Consciousness, Pseudo-consciousness, and the Moral Significance of Consciousness[1]

Geoffrey Lee

Forthcoming in Lee and Pautz eds. *The Importance of Being Conscious* OUP

Philosophers have tended to picture consciousness (as in "phenomenal consciousness": sentience or subjective experience) as a kind of "inner light" whose presence marks a deep divide in nature, and which has great significance for the beings whose inner life is illuminated by it. However, I believe that the "inner light" metaphor has mislead us, and that once we understand consciousness's true metaphysical nature, in particular once we accept that some form of broadly physicalist view must be correct (so e.g. it is a complex high-level property of the human brain), it may be reasonable to revise our attitude to its significance. A better picture is that consciousness is but one of many kinds of "inner light", and that other kinds can be just as significant for the creatures that have them. This paper will (partially) describe the case for this view, with particular attention to the moral and practical significance of consciousness, and the question of how a revisionary metaphysical view can motivate a morally revisionary view.

A conscious experience is a type of mental event that plays a certain role in our cognitive economy. Some aspects of this role are well-known to us as part of our shared folk-psychological self-understanding. For example, we use perceptual experiences as a guide to forming beliefs about, and acting on, the environment around us. Other aspects may only become clear with empirical research. For example, certain kinds of memory or perceptual learning may depend on conscious experience, whereas others may be possible through subliminal perception also (Allen (2017)). This role may be partly grounded in intrinsic features of the experience such as its representational and mereological structure, or features such as the representational content of the experience – I will use "role" in a very broad sense here, to include these causally/explanatorily relevant features also.

If a type of internal state plays a role that at least is roughly similar to the role played by our conscious states, we can say it is "pseudo-conscious", or "consciousness-like". And we can consider the class of beings that have consciousness-like states, the *pseudo-conscious beings*. Let's assume that pseudo-conscious beings who are not conscious (*merely* pseudo-conscious beings) are possible (I do not say this is obvious, and will say more about why we might believe it below). An example might be an intelligent AI with a cognitive architecture broadly similar to a human brain – e.g. it has perceptual systems feeding into cognitive system responsible for planning and executive motor control. Suppose, despite this, they are a mere zombie machine. My question is : what

[1] For helpful comments and discussion I'm very grateful to Olivia Bailey, Brian Cutter, Shamik Dasgupta, Uriah Kriegel, Alex Jackson, David Papineau and Adam Pautz, as well participants in a seminar at UC Berkeley where material from this paper was presented, a number of metaphysical mountaineers (you know who you are), and audience members on a number of occasions when I presented this material.

practical attitude should we have to them? What significance do their pseudo-conscious states have?

The standard view is that if they lack the "inner light", then their internal life is severely impoverished – in fact it might be a stretch to even use the term "internal life" to describe what is going on with them. They are mere physical machines, with no genuine subjectivity or interiority. They lack the special epistemic connection to the external environment and to their mental life that we conscious beings enjoy. Moreover, from a practical point of view, they lack a central ingredient that makes life worth living. And since they have no genuine feelings, it may be morally acceptable to inflict damage on their bodies and brain-analogs in ways that cause analogs of pain and emotional distress – that is, internal states that play similar roles.

Now, arguably this thinking rests on the intuition that there is a profound metaphysical difference between genuine conscious states and these merely pseudo-conscious states (a version of what I call the "big difference intuition" – the intuition that there is a big difference between conscious states and mere functional simulacra). It's because their internal state is nothing like ours – the "lights are off" – that treating them so differently seems justifiable. However, I believe that this metaphysical intuition is a mistake, at least if a physicalist metaphysics of consciousness is correct. If consciousness is merely a high-level physical or functional property of our brains, then there is no deep metaphysical difference between what we enjoy, and the pseudo-conscious internal states of, say, some intelligent aliens, or silicon-based AIs. There is mere an arbitrary difference in how aspects of the consciousness-role are physically and functionally implemented. I think this can reasonably shake our faith in the unique significance of consciousness.

To get a feel for why physicalism, properly understood, might put pressure on the "conscious beings are special" view (following recent precedent I'll call this "sentientism" – to be defined more carefully below), consider an analogy. Many people have a moral ideology that puts profound significance on the property of being a human being. Only humans have any intrinsic importance – everything else only matters in so far as it affects humans. Those who hold such an ideology might mistakenly picture humans as fundamentally different from other beings, and take their normative scheme to be partly grounded in this deep divide. If they discover that the distinction between humans and other creatures is much more messy and superficial than they thought, they might reasonably be prompted to question their moral thinking. For example, suppose hypothetically that Neanderthals had not been made extinct, but continued to live in a part of the world isolated from Homo Sapiens, and then suddenly contact was made again. Although the human-worshippers could intransigently hold onto their normative attitudes, classify the Neanderthals as non-human, and treat them as mere means, a reflective member of the human-worshippers might reasonably be shaken by the profound similarity between the two species. Their sense of a deep divide having been pulled away, they might think about what surrogate to their old idea of "deep humanity" really exists in the world, and how something close enough to the old

morality can be reconstructed around it. That might well lead them to include Neanderthals as intrinsically important, despite their being a different species[2].

Or consider the following kind of case (described vividly by Dave Baker in his story "The Hunter Captain"[3]). We encounter a species of merely pseudo-conscious aliens, who conclude that we are not "conscious beings" (their language), on the grounds that we do not have the neuroanatomy that their science of "consciousness" has revealed is crucial to the "consciousness" that they enjoy. Since we are merely "insentient" machines that lack the inner light, they conclude that it is unproblematic to torture and enslave us. On my view, neither species is more or less *metaphysically* correct in describing what they enjoy as an "inner light". Once we acknowledge this, I say that there is motivation for adopting a moral framework that treats both species as morally important, even if our word "consciousness" does not in fact apply to them, and even if we started as sentientists: i.e. those that believe consciousness is required for moral status (i.e. mattering for one's own sake morally) or for living a valuable life.

Of course, sentientists may not yet be moved. They may insist that consciousness is a deep metaphysical divide in the way that being a human is not. This might extend to insisting that there is a deep metaphysical asymmetry between us and our alien oppressors, who are simply subject to an unfortunate *illusion* in believing it is *they* who are special. Or they may concede the metaphysical point, but think that our investment in the significance of consciousness is so foundational that there is no reasonable alternative than to hold onto it. I reject both of these ideas; in this discussion I presuppose the metaphysical point, and focus on the normative issue[4].

One way to frame the normative revisionism that I advocate is as "Deflationary Pluralism". On this view, there are pseudo-conscious states (i.e. states playing a similar role to conscious states) that have the same significance as conscious states despite lacking phenomenology. I call these "quasi-conscious states". For example, perhaps consciousness is essentially a kind of biological phenomenon, dependent on neural activity. Some AIs might lack this property, but have an artificially grounded property that plays a similar role. The deflationary pluralist says that it can be just as normatively significant as consciousness.

Since there are different kinds of significance we attach to consciousness, there are different versions of the deflationary pluralist view. Here I focus on the moral/practical significance of consciousness[5]. In brief, what the deflationary pluralist insists on is:

(1) Mere pseudo-pains and pseudo-pleasures can be just as good and bad (prudentially or morally) as pains and pleasures despite lacking phenomenology.
(2) A merely pseudo-conscious life can be just as worth living as a conscious life.

---

[2] Kagan (2016) also presses the analogy between speciesism and sentientism.
[3] Baker (2016). Baker actually leaves it open in the story whether his aliens ("the Nampranth") are conscious beings; so there is a possible interpretation of the story where both species mean the same thing by "conscious" and the Nampranth are mistaken in concluding we lack what they have.
[4] See Lee (2018) and Lee (manuscript) for more discussion of the metaphysical issue.
[5] In Lee (2013) I discuss the epistemic significance of consciousness.

(3) Merely pseudo-conscious beings can have the same moral status as sentient humans.

(1) and (2) are intended to make deflationary pluralism a prudential as well as moral view. For example, on this view, one could calmly accept a "zombification" procedure that replaced one's neural hardware with a functionally similar but consciousness-incompatible artificial hardware[6] (note the parallel to Parfit's (1984) well known buddhist-inspired deflationary view of personal identity).

On "moral status": most simply, the idea here (as mentioned) is that pseudo-conscious beings can matter as much sentient humans for the purpose of moral deliberation. One way to elaborate this is in terms of the idea that moral deliberation takes into account the *interests*, *welfare* or *well-being* (I assume in what follows that these are equivalent ideas) of different agents, and that to grant agents equal moral status is to say that their interests matter equally (Lee, A. (2022)). Put in these terms, (1) and (2) can be read as saying that non-conscious beings can have interests in the same way that conscious beings can, and (3) says that these interests can matter as much as sentient human interests (so (3) presupposes (1) and (2)).

I then take sentientism[7] to be the denial of these claims[8]. It has a weak or strong reading depending on whether mere pseudo-consciousness has *no* value, or *less* value (or similarly : pseudo-conscious beings have *no* moral status vs *lesser* moral status, or theirs lives are *not* worth living at all vs *less* worth living). I assume the strong reading, but I don't think that affects anything in what follows.

Note that sentientism is not committed to welfare hedonism – the view that all that matters for welfare is pleasure and pain. The sentientist could think that other factors (e.g. life projects) make our lives go well or badly – but we have to be conscious in the first place for these things to matter (more on this below).

Importantly, although my view can be framed as deflationary pluralism, I would rest content with other construals. I'm also a *conceptual pluralist* about phenomenal consciousness, in the sense that I think that there are a number of important concepts in

---

[6] For an opposing perspective on zombification see Siewert (1998, 2021)

[7] If we count welfare hedonist as a form of sentientism, then discussion of sentientism has always been a theme in debates about the good life, including in non-western philosophy (e.g. in discussions of materialism in classical indian philosophy (Adamson and Ganieri (2020) ch.32). Recent advocates of sentientism include Lee A. (2022), Lin (2021), Shepard (2018), Siewert (1998, 2021), Singer (2009).

[8] Lee, A. (2022) argues that sentientism should be understood as the view that non-conscious beings don't even have interests in the first place (see Singer (2009), pp 7-8) – whereas a speciesist view would say that although another species has interests, those interests are less important than human interests. I don't know if much turns on this for present purposes. Pseudo-conscious beings can care about how their life goes in a functional sense of "care", and can have functional analogs of pleasures and pains and other welfare goods. My argument below that the metaphysical symmetry between these "interests" and our interests motivates a moral symmetry would ipso facto imply that they are "real interests" (whatever that means!) in the first place.

the vicinity of "phenomenal consciousness" that need to be distinguished. I think most people working on consciousness have an *empirical realist* construal in mind, and it's that construal that I think pairs naturally with deflationary pluralism. On the empirical realist view, consciousness is a natural phenomenon which we ostend in ourselves, and which in fact plays the "consciousness role" in us (a role that need not be a priori). So on this view, there is an intelligible scientific/philosophical project of figuring out what the underlying *nature* of consciousness is – i.e. the feature in us that plays the consciousness role (hence we have "theories of consciousness"). But we can also consider :

> **The Superficial Conception** : To be conscious just is to be pseudo-conscious (i.e. to have a property that plays the consciousness-role).
>
> **The Imaginative Conception** : To be conscious is to be the legitimate target of empathetic imagination.
>
> **The Normativist Conception** : To be conscious is to have consciousness-like property that is normatively significant in the way that human consciousness is.

I tend to think that these concepts are more revisionary than the empirical realist concept, which is therefore a plausible construal of ordinary experience-talk. But that's less my overall concern than to press the view that these concepts are both conceivably, and in actuality, *not extensionally equivalent to each other* (I only partially argue for this here).

This said, I will proceed on the assumption that superficialism is false, and that therefore merely pseudo-conscious beings are possible. It's worth clarifying why I think this is plausible. The pseudo-conscious states, as I think of them, do not all have in common any interesting or deep functional profile (or any other common thread), because they needn't resemble our conscious states *in the same way* (talk of *the* consciousness-role is therefore misleading[9]). As mentioned, conscious states are going to be characterizable both in terms of their folk-psychological properties and the kinds of features that empirical science determines them to have. I'm thinking there is a plurality of different kinds of "consciousness-like" roles that can be constructed from different aspects of the rich characterization of human consciousness we might ultimately hope for. Pseudo-consciousness is therefore a matter of family resemblance. So for example, a baby seahorse and an AI[10] could count as pseudo-conscious in virtue of completely different resemblances to us, and in fact have nothing in common physically or functionally beyond very abstract structural features that are not sufficient for consciousness (compare how we might loosely pick out all phenomena that are "waterish" or "lifeish" or "humanoid").

This "role pluralism" is one reason why "consciousness is superficial consciousness" is not an attractive view: we think of conscious states as a metaphysical unity, not as a loose confederation (I do not say this could not be revised though!). The other is simply

[9] Recall also that I understand "role" in a very broad way to include explanatorily relevant structural features of experience, such as its mereological or representational structure.

[10] Mclaughlin (2003) p.184 offers a similar example.

that 'the consciousness role' as I understand it, is defined in terms of folk-psychological and empirical beliefs about the structure and role of consciousness, and famously, we do not think of these features as definitional (as on e.g. an analytical functionalist view), but rather we think of consciousness as a phenomenon that we pick out in a quasi-demonstrative way in our own mind, and we may independently *believe* has these various features (features which give us leverage in figuring out empirically what it really is). This suggests that "alien variants" of consciousness - properties distinct from consciousness which play the consciousness-role in other beings – are both coherent and metaphysically possible. They are what we find if we decenter the project of figuring out the identity of consciousness on other beings, treating them as if they are themselves conscious[11].

If you reject this picture (e.g. because you are a superficialist), then I can fall back on this formulation: it would be a mistake to think that the scientific/metaphysical project of figuring out what consciousness is, *as it is usually understood*, automatically also uncovers what matters morally. So for example if we discover that octopi do not have whatever plays the consciousness-role in us, we should not jump to conclude that what plays the pain-role for them does not *matter* in the way that pain does. (Below I also argue for similar view for the targets of imaginative empathy: e.g. to be able to legitimately "imagine what it's like" for an octopus, it need neither have "consciousness" (=what plays the consciousness role in us), nor need it have a feature that matters morally).

It is also worth emphasizing that such alien realizers of the consciousness-role are probably not the only group of properties that realize this role. As discussed in the introduction to this volume, there are likely to be many slightly different physical/functional properties that actually belong to human beings, all of which are "consciousness-like", and are thus reasonable candidates for what we mean by "consciousness". If we suppose that exactly one of these is "consciousness", then the arguments for deflationary pluralism apply just as much here too. The reason why I focus on the case of alien realizers in octopi, AIs, extra-terrestrials etc. is that I think these provide the most interesting cases of pseudo-consciousness, precisely because they involve properties that we humans *do not in fact have*, and therefore are

---

[11] Two things to note here. One is that rejecting superficialism is not the same thing as rejecting functionalism. An empirical realist might believe that it is a functionally defined property (e.g. the broadcasting of information to a global workspace (Baars 2005)) that unites all the conscious states. The second is that the deflationary pluralist is not just concerned with a cloud of consciousness-like states that humans actually have, and which are reasonable candidates for what we mean by "consciousness". When we decenter on other creatures or systems, we may find consciousness-like states that we definitely do *not* have and which are determinately *not* conscious. Still, on a deflationary pluralist view they might be as significant as consciousness. This is one way in which a deflationary pluralist view need not be analogous to Papineau's view (2002) view that "consciousness" is surprisingly *indeterminate*. That said, the deflationary spirit of Papineau's materialist view is very much in keeping with my own thinking (see in particular Papineau (2007) section 5.4, and his contribution in this volume).

determinately not consciousness. Nonetheless, they can have the same significance (or so say I).

My discussion in what follows has two parts. First, I want to explain why "digging below bedrock" and aiming to unseat a foundational normative attitude to consciousness through metaphysical means is not an unreasonable project. Second, I will address several important objections to the deflationary pluralist perspective. My concessions to these objections will highlight the priority for me of conceptual pluralism over deflationary pluralism.

Digging below bedrock

The Deflationary Pluralist says that there is a family of consciousness-like properties that are equally significant as consciousness; I take this to be *revisionary*, in that it rejects the necessity of consciousness in order to have states that matter in the way that conscious states do (here I assume that sentientism is the default view). This is different from a revisionism that claims consciousness *lacks* significance – it is the elevation of it onto a supreme pedestal that is rejected. The view is also not a kind of eliminativism or illusionism about consciousness ("consciousness doesn't exist!") – it is a deflationary form of realism. At least for consciousness, pluralism is the forgotten sibling of eliminativism (aka illusionism) and more standard realism. An alternative to denying that consciousness exists is to say "there are many consciousnesses!", rather than treating it as a single elite property. This said, I do think that thinking through what our practical attitude would be, were we to discover that consciousness doesn't exist, is a useful heuristic for thinking about our actual situation correctly. I call this the "eliminative heuristic". Even though many have claimed that a world without consciousness would be a world without value, I nonetheless believe that the most natural (and reasonable) reaction to discovering eliminativism is *actually true* would be to look for a *surrogate* for consciousness to be what we care about. This is rather like how the human supremacist, on discovering that "deep humanity" as they understood it, doesn't really exist, has to find a surrogate – and that might be a surrogate that includes non-human beings. Similarly, although I think that consciousness exists, I don't think that "deep consciousness", as depicted by the inner-light metaphor and the "big difference" intuition, exists, and the fallout is rather similar to the illusionist fallout.

Again, my concern here is not to develop the metaphysical part of this argument. I believe that on a wide range of construals of a "physicalist view", consciousness will be just one among many high-level physical properties, that will not stand out metaphysically (it has no special "glow")[12]. A zombie martian scientist, in comparing us

[12] I think this is particularly obvious on a reductive or "identification" physicalist view, on which consciousness is a complex physical or functional pattern (see the introduction to this volume). However, even on a strongly "non-reductive" or "grounding physicalist" position (see introduction), the point might go through. It all depends on how sparse the properties are that resist a reductionist treatment. Conceivably, human consciousness might resist reduction, but an alien species might lack this special property and only have functional structure that is amenable to a reductionist treatment. That would make for a metaphysical asymmetry. Still, I think this kind of "quasi-dualist" view is unmotivated. The intuition that there is a deep divide

with Baker's oppressor aliens, will not see some special metaphysical asymmetry between us; they will just see two different physical systems with different properties playing the consciousness-role. I say this is diagnostic of there not *existing* any such difference between us. This is (roughly) what I mean when I say that consciousness "lacks strong natural significance".

It is also worth briefly noting here that there is a plausible debunking argument against a metaphysical "big difference" intuition. Even granting that the intuition is correct, we can ask: do we have the intuition *because* of the big difference? I won't argue the point here, but it's very plausible that the answer is "no". It's not as if God set up a giant chasm between consciousness beings and the rest and then made sure that we appreciate her marvelous creation. Rather, our sense of a deep divide between "light on" and "lights off" plausibly has to do with such factors as the all-or-nothing application of imaginative empathy, factors that are not sensitive to a deep metaphysical divide, even assuming that it exists.

Now, in previous work, I framed what is at stake here in terms of the following argument:

The Natural-Normative Grounding Argument:

(1) If consciousness has strong normative significance, then it has strong natural significance
(2) Consciousness does not have strong natural significance
(3) Therefore, consciousness does not have strong normative significance

("Strong normative significance" here indicates the kind of *unique* normative significance that consciousness would *not* have, if merely pseudo-conscious states can have a similar normative significance.)

I now have some misgivings about this framing (as I will explain), but I still think it's a helpful framework for thinking about the issue.

Those who wish to reject the conclusion may give two different justifications for this (independently of whatever arguments they give against the premises). They may either attempt to give a deeper explanation of why consciousness is strongly normatively significant, or claim that it is a foundational normative commitment that cannot be dislodged by any argument of this kind.

On the first horn, one could look in detail at different options for explaining the normative significance of consciousness – for example, in Lee (2013) I look in detail at options for explaining the epistemic significance of consciousness. Here I will rest

---

does not (according to me) have epistemic weight (e.g the aliens have it too). And anti-reductionism is probably best motivated by a desire to treat the ontology of high-level sciences in a metaphysically heavyweight way. But in that case, both aliens and humans will have irreducible properties in great abundance (i.e. those appealed to by explanatory theories of how they function). Consciousness might be among them, but a functional analog of consciousness might be among those enjoyed by the aliens.

content with a general point about why such deeper explanations tend to be unpromising. What we are looking for is some normative principle (or other fact) that is not itself a basic commitment about the significance of consciousness, but which can nonetheless explain the significance of consciousness. The problem is that it's implausible that one could find such a neutral deeper ground that didn't also ground pseudo-consciousness having a similar significance. Consider for example, the suggestion that consciousness is morally or epistemically significant because we have a special, direct, way of *accessing* it. For example, a pain or pleasure wouldn't matter if you weren't aware of it! Such a suggestion faces a dilemma – either the relevant kind of access already involves consciousness, in which case we haven't explained the significance of consciousness in terms that don't presuppose its significance, or it is also available to the pseudo-conscious being as well. Similar points apply to concepts like "being a subject" or "having a point of view". They might well play a foundational role in our normative thinking, but that can't really explain why consciousness matters, because we can always ask why e.g. *being a subject* in a conscious-involving way, rather than in a purely functional way, is what is important.

So the sentientist should probably take their view to be normative bedrock. Does this end the debate? I say no! For one thing, basic normative commitments can still have presuppositions which might fail, and therefore call them into question. An obvious example is the *existence* of consciousness. As mentioned, if we discover it in fact doesn't exist, we might reasonably question whether it is what matters, even if that is a basic commitment. My positive case for deflationary pluralism hinges on the idea there is a similar presupposition involving the deep metaphysical significance of consciousness. **We think there is a big difference, metaphysically, between having a real pain (lights on!) and having a zombie ersatz pain (lights off!), and that is *why* treating these cases very differently normatively is perfectly reasonable.**

But if this metaphysical presupposition fails, it's not that we now triumphantly reason our way to a new moral framework. This is one way in which the "deductive argument" framework above could mislead. It's more that we face a kind of normative crisis, where *all* our options are somewhat revisionary, and we will have to do our best to find new footing (hence the parallel with illusionism). This (I claim) opens space for taking seriously the more inclusive view, but it also leaves plenty of room for an intransigent sentientist. They can say : sure, consciousness isn't the kind of thing we thought it was, and our treating pseudo-conscious beings differently cannot be justified by saying that what they have is nothing like what we have. Still, our position is the best, because least revisionary.

My opponents may also point to the fact that the more inclusive view cannot be justified on the grounds that there is a more inclusive property that *does* have the special metaphysical glow we have discovered consciousness to lack. So whatever property we attach strong normative significance to won't satisfy the analog of premise (1). So why, once all the chips are on the table, is it a problem if consciousness doesn't satisfy it? A closely related complaint about casting a wider net is that there is no natural stopping point. Suppose we agree that consciousness is just the arbitrary human realizer of the consciousness role, and are motivated to find a different criterion for moral status. Surely whatever criterion we choose will have its own arbitrariness? For example, we could use pseudo-consciousness, or an unrelated criterion like being a certain kind of

intentional system. But then couldn't it always be complained that there are systems outside the net who are not too different from those in the net? Or more generally, can't we always wonder what's so special about the physical configurations we treat as morally significant beings? After all, everything is ultimately just atoms swirling in the void. A third related complaint is that we should be wary of adopting a principle that in general requires moral joints to line up nicely with natural joints. Is the proponent of the argument presupposing some such principle?

In response, there is certainly no presupposition here that in general natural and moral joints need to line up. For example, as I'll discuss below, there can be moral and political motivations for the concepts we use that are quite different from the motivations of explanatory projects in the natural sciences. My picture is that in certain *specific* cases - consciousness, personal identity, perhaps others – we tend to think in terms of a heavyweight metaphysics the phenomena in fact lack, and we put emotional weight on this. Once the true metaphysical picture comes to light, this therefore creates space for reconsideration of our moral/practical thinking. Furthermore, I agree that wherever we land will tend to look somewhat arbitrary from a cosmic perspective. I certainly am not expecting a transcendent Kantian justification for our moral system, and agree that whatever we carve out as mattering is ultimately just atoms swirling in the void.

This might make the shift to a more inclusive view seem completely unmotivated. But that ignores the fact that there are other strands in our thinking about moral status we can reasonably put weight on. In particular, I assume pseudo-conscious beings are capable of being agents with projects and other emotional investments in the world. I claim that once consciousness is revealed as lacking the metaphysical depth we thought it had, and these other beings at least have something *similar*, tilting the scales towards these other criteria should seem attractive (that's my reaction, anyway!).

We can put the point here in terms of a puzzle (raised forcefully in Bradford (forthcoming)) about views of well-being that reject pure hedonism (only experiences matter for well-being). If states of affairs external to your experience matter for how well your life is going (for example, having your desires satisfied, or pursuing and completing valuable projects), and a zombie could also have a form of well-being involving such goods (because, e.g. they can have desires and pursue projects), why can't they have a life worth living?

Now, in fairness to the sentientist view, there *is* a way for those who reject pure hedonism to make sense of "no zombie well-being". By way of analogy, consider the attitude of a car enthusiast, who really loves owning a car, and really loves the particular car they own. For them, although having a car is their central concern, it's important that their car has excellent features: sweet rims, an abrasive sound system, leopard print seat covers, flames painted on the side, etc. These things make the car worth owning – they would not want to drive around in a totally plain car, especially if they see out on the road other people driving the kind of car they would like to drive. Now although these accessories play this central role, that doesn't mean that they would be a source of value absent the car. They wouldn't necessarily care to own a boat with an abrasive sound system and flames painted on the side, for example. Similarly, I think we see the value of life-projects and other external-to-consciousness goods as like

accessories to the conscious life. They are needed for life to go well, but the thing at the center that is made to go well by them is a *conscious life*.

That said, the fact that some pseudo-conscious zombies could have a form of life remarkably similar to ours, where they value similar kinds of projects and so on, does point to a peculiar lack of unity in the non-hedonistic sentientist view. For the deflationary pluralist, we are a bit like car enthusiasts who insist on making the internal combustion engine an essential part of the definition of a car, so that an electric car isn't even really a kind of car, and so is not worth owning. This, even though the electric car fans can have a very similar cultural practice, valuing the same kinds of accessories etc., and enjoying their electric cars in entirely analogous ways. Importantly, what enables this similarity is that the way the accessories bring value doesn't depend on the kind of engine in the vehicle – there is no engine-accessory synergy. Similarly, the "life-accessories" we value so much do not seem to essentially require consciousness as opposed to pseudo-consciousness (although conscious experience might (locally for us) be a means that in fact is important in achieving our external goals – e.g. being able to see is useful in writing a book). So in the end I do think that rejecting a purely hedonistic concept of well-being is a motivation against the sentientist view.

There are also pragmatic arguments here. One important way to think about morality (probably not the only legitimate way) is that it's about the norms we agree on to live together as a "moral community"[13]. Part of the game on this conception is to find a mutually beneficial framework we can all agree on. But who gets a seat at the table in the first place? In situations where groups are living near or around each other and are prone to conflict, there could be an argument from peace and stability for casting a wider net. If two antagonistic groups mutually agree to treat each other as "ends in themselves", and think of themselves as part of a single moral community, this might be mutually beneficial as a means not just to keep the peace, but also enjoy the benefits of cooperation and collaboration (as some proponents of liberal politics have stressed). Conceivably, we could face a situation involving, say, pseudo-conscious intelligent aliens who we come into conflict with, that fits this mold (perhaps something like this is happening in Baker's story). (It must be acknowledged though, that there are also many cases involving an asymmetrical power dynamic where, if anything, it is more convenient for the powerful group to treat others as moral outsiders (presumably that's our intention with future AI!) – so this consideration may only apply in a limited range of contexts).

A closely related point is that functional zombies are certainly capable of being moral *agents*, in the sense that they have the cognitive capacity to engage in moral thinking and live in morally-regulated communities. There's real tension between acknowledging this and denying that they are moral *patients*, i.e. they have moral status. If we lived in community with a zombie moral agent, we (presumably) would want to be able to hold them responsible, forgive them, debate with them about what is right and wrong etc., and expect similar treatment from them. But we could not expect such reciprocity if at the same time we tell them we think their life has no importance!!

[13] With idealization built in, so they aren't automatically just our *actual* norms.

Finally, I want to stress that my aim is quite modest. I'm not really trying to show you an unavoidable route to the moral truth about consciousness, but rather warm you to the more inclusive view as a view to be taken seriously. My attitude to moral philosophy here is pluralistic[14]. There is a space of different moral frameworks one could articulate that are grounded in our moral psychology. We can explore this space of frameworks and assess their attractiveness in terms of what they do or not preserve in our untutored moral thinking, what the consequences would be if we lived by them, whether they have deep unifying principles, and so on. In this way I reject any vision of ethics as a "quest for the moral truth". Since my thinking about sentientism is informed by this ethical/meta-ethical sensibility, it might not resonate with those who have the more conquistadorean outlook. For example, if a sentientist takes it is as a direct insight into the objective moral truth that a zombie's life cannot matter at all, then on discovering illusionism is true, they may be committed to believing that all human lives do not matter[15]. As mentioned though, I think there's room for a pragmatic shrug here, and an attempt to find a reasonable *surrogate* for what we once cared about. And there's surely not one correct way to do this – if we settle on a new moral framework, this is to some extent more like a leap of faith than embracing a scientific theory. That is, I'm comfortable with the (analog of) epistemic norms for moral and practical commitments being different (i.e. looser in some ways), than those governing non-normative theoretical enquiry.

To sum up, there are at least three reasons to take the inclusive view seriously, even if the sentientist view is normative bedrock:

(1) Merely pseudo-conscious beings are not as different from us (metaphysically) as it first appeared (i.e. when we were in the grip of the 'inner light' picture).
(2) Merely pseudo-conscious beings can have other features that we think of as central to moral status, such as being able to pursue a life with projects they care about.
(3) If we lived among merely pseudo-conscious moral agents, both groups might have a pragmatic interest in being "seen as conscious" by the others, and forming a moral community that fostered stability and cooperation.

Again, I don't think these considerations force anyone to give up the sentientist view; but I hope they at least make my position seem like an intelligible place to land, and not just a perverse provocation!

In the rest of this discussion I turn to considering three different responses/objections to my position.

<u>Response 1 : Flipping the script?</u>

What if the sentientist attempts to tollens my ponens in favor of strong natural significance (despite whatever argument has already been made against it) (see Cutter

[14] An intellectually towering embodiment of (something like) this mindset can be found in Isaiah Berlin's work.
[15] This is admittedly a bit simplistic, because the moral realist sentientist could also have a high prior that human life matters.

(2017))? For example, they may believe that dualism about consciousness would save the day. Here I will (mostly) rest content with simply asserting my position: I'm strongly opposed to this kind of reasoning, which strikes me as blatant wishful thinking! Individuals with the same evidence and the same background beliefs should not be licensed to hold different views about the metaphysics of the world, in virtue of having different value commitments. That is, our practical normative commitments should be a *conservative extension* of our theoretical commitments (to put it in Fieldian terminology (Field 1980))[16].

This doesn't mean that our reasoning about the physical world can never involve normative premises; it's just that this reasoning should be a kind of convenient proxy for purely theoretical reasoning[17]. So to take a well-known kind of example (Dorr (2002)):

(1) If a political/social system is deeply unjust, it will eventually become unstable and collapse
(2) The system of country X is deeply unjust
(3) Therefore, the system of country X will eventually become unstable and collapse

Arguably this reasoning is only legitimate if it is a proxy for a purely non-normative argument concerning justice-attitudes or other non-normative correlates of injustice (see Enoch (2003) and Lenman (2003))'s responses to Dorr (2002)). So for example, the idea here could be that if the physical conditions for injustice obtain in a country (whatever those are according to the reasoner's vision of justice), then inhabitants of that country will find the situation intolerable and demand change. One reason this is plausible is that the justification is presumably available even to someone who rejects the reasoner's vision of justice (perhaps they think that the only just society is an authoritarian one), but who has the same evidence. They can substitute something like the purely descriptive "deeply unjust according to the standards of western progressives" and they are good to go (see also footnote 9).

Clearly, the argument that dualism must be true because only then can the significance of consciousness be saved, is not a proxy for purely descriptive reasoning. For example, it cannot be appropriated in a modified form by someone who is not a sentientist. My challenge to those who reason in this non-conservative way is to develop a plausible epistemology on which their normative commitments and conditional commitments

---

[16] This probably does need refining to allow for certain legitimate forms of pragmatic encroachment. For example, as William James pointed out, and as signal detection theorists love to emphasize, setting the threshold for believing involves a pragmatic trade-off between false positives and false negatives.

[17] If we have conservativeness, then whatever non-moral commitments justified introducing the moral commitments will independently justify whatever non-moral conclusions we get from moral premises (Lenman (2003)). I can imagine a kind of reliabilist-inspired relaxation of conservativeness according to which, the propositions that serve as factual premises in introducing moral claims need not actually be *believed* by the subject, but instead merely be facts reliably correlated with moral intuitions (e.g. their causes), or perhaps propositions mentally represented without being believed that cause moral beliefs etc.

linking the normative and non-normative realms are *reliable insights*, and not mere prejudices.

I will also note here that the situation is clearly far more delicate if it is *epistemic* normativity that is at issue, because patently, different epistemic schemes *do* license different descriptive beliefs. Still, "the world must be this way because otherwise our epistemic norms will need to be revised" also is (it seems to me) a form of wishful thinking – and so I am strongly inclined to think that "flipping the script" on an epistemic version of the current argument would be equally unattractive (we see something like this is Pautz (2017)). Clearly this deserves much more discussion, however.

Response 2 : The Imaginative Empathy Objection

A sentientist might also argue as follows. If pseudo beings lack consciousness, this means that *we can't imaginatively empathize with them* – we can't take up their subjective perspective. But isn't that a necessary condition for caring about another being? How then can we say that they have moral status?

To elaborate, consider the following way of expressing the consciousness/moral status link:

> **Status Norm** : You ought not to care about S unless S has conscious experience

One way to unpack this idea (and I do not say that it has to be unpacked this way, or that the following norms are more fundamental in our thinking) is in terms of the following two norms, which jointly entail the status norm:

> **Status-Imagination Norm** : You ought not care morally about S unless it is appropriate to imaginatively relate to them
>
> **Consciousness-Imagination Norm** : You ought not imaginatively relate to a being unless they have conscious experience

The important point here is that since these norms entail the status norm, then if you reject the status norm (as I will have to), you have to reject at least one of these norms.

A couple of clarifications. First, I use the term "imaginatively relate to" to mean something like "imaginatively empathize with", or "take up the point of view of in imagination". I deliberately intend to exclude a moral or emotional connotation. In particular, I can "imaginatively relate" to you without necessarily *caring* about you. Second, the kind of "appropriate imaginative relating" that I have in mind here is the minimal kind of appropriateness we get from the target simply being a conscious being. Granted that e.g. a bumblebee is conscious, it makes sense to try to picture their experiences in imagination, in the way that it wouldn't for an entirely unconscious rock. Whether these imaginings are *accurate*, or whether we are even *capable* of accurately imagining the experiences of the target being, is a further matter that is not relevant to whether they are "appropriate" in the relevant sense.

Which norm should an anti-sentientist like me reject? Now, you might think that the Status-Imagination norm would be the most plausible target for me here. Perhaps I should argue as follows:

(1) Having the capacity for imaginative relationships to each other is inessential for a group of agents to have a cogent moral practice.
(2) If having imaginative relationships are inessential to a moral practice, then the status-imagination norm is false
(3) Therefore, the status-imagination norms is false.

In defense of (1), we can ask : is imaginative relating really all that central to *our* moral practice? If it isn't, why couldn't we have a cogent moral practice where we live alongside zombie moral agents who we do not imaginatively relate to (and so it wouldn't matter if we *couldn't* legitimately do this)? There are several arguments in favor of this. First, presumably I could (at least in theory) completely lack imagination but still care about other beings (including pseudo-conscious beings), in the sense that I could make morally appropriate decisions concerning them. That is: imagination isn't computationally necessary for moral decision making (even if we sometimes or often in fact use it in our deliberations). Second, imaginative relating might even be *morally unhelpful* in some ways: e.g. biasing us towards individuals who we familiar with, or causing negative affect that tempers our moral motivation (Bloom (2017), Prinz (2011)). Third, some of the most important moral emotions (compassion, indignation) arguably don't require imaginative empathy, and so it's not clear how big a role it plays in moral psychology (Bloom (2017), Prinz (2011)). Finally, it is also noteworthy that some individuals like radical ecologists already treat entities like ecosystems as mattering morally without thinking of them as imaginatively relatable; why not think of pseudo-conscious beings in a similar light?

Now, although there is a possible case here for rejecting the status-imagination norm, these considerations do give me some pause. The thing is this: a lack of imaginative empathy might be harmless or even helpful for a detached decision maker ruling on others' behalf. Thus, the kind of moral norms that apply to such a ruler perhaps need not mention imaginative empathy, and thus could direct decisions that were impartial with respect to conscious and merely pseudo-conscious beings. But if we think of morality as concerned with how we treat each other as we live *together*, then it will encompass situations involving collaboration and coexistence, such as interactions among individuals who live in the same house and make collective decisions. I think it's quite implausible that the capacity to take up the point of view of others in imagination isn't a crucial part of this kind of *collective* moral life, at least *given the kind of psychology we have as humans*. If we wanted to live not just *alongside* pseudo-conscious beings (perhaps ruling over them and making detached decisions on their behalf), but *with them* as part of the same community of autonomous moral *agents*, then our relationship would be much impoverished if we did not, or could not, imaginatively relate to them (e.g. if we tried to be friends with them or collaborate with them). So there is probably an important sense in which imaginative relating *is* quite central to (human) moral life.

Having said this, this richer kind of empathetic moral relationship is probably not needed to care about pseudo-conscious beings in a more minimal way that constitutes

treating them as *mattering* in the relevant sense. And so rejecting the status-imagination norm might still be defensible. Perhaps that's right. Still, I don't think I need to settle for this, because I think we *can* appropriately imaginatively relate to these beings! That is, it is (perhaps surprisingly) defensible to reject the consciousness-imagination norm, and this is my preferred response to the objection.

To be clear then : I want to say that we can legitimately "imagine what it's like" for non-conscious beings (although this terminology doesn't fit this activity very well!). The first pass of my defense of this takes the form of a parable:

> **The Parable of the Ice-houses** : An isolated community lives in little 1-person ice-houses, and no-one is allowed inside another person's ice-house. However, each person can use their ice-house as a reasonably accurate model of the ice-houses of others, and they have engaged fruitfully in this practice for thousands of years. They also worship the ice-houses as sacred (one of them, a philosopher named Naleg Noswarst, even declares that the ice-house/non-ice-house distinction is the most important in the whole of reality![18]) One day, they discover a different community that lives in a similar ultra-private 1-person houses, except that these houses are made from blocks of clay, not blocks of ice, albeit with similar functional properties nonetheless. An intrepid philosopher from the ice-house community, Leefrey Ogef, suggests that it would not be unreasonable to use his ice-house as a model for a clay-house. After all, in many ways, they have a similar design. He would also be totally cool with moving into one of them; because he can model one of these clay-houses with his own house, he's pretty confident that life would be just as good. Other philosophers respond that these are not ice-houses, and therefore it is absurd to use an ice-house as a model for them. When we model, we are modeling the *ice-house*-state, the state of the sacred ice-house! Moreover, if you can model clay-houses with ice-houses, where does this stop? You could model almost anything with an ice-house. After all, ice-houses resemble many things in many different ways. Are all these many things as important as ice-houses? Would Ogef live pretty much anywhere? Is nothing sacred?

At the heart of my response is a certain view of how we can understand the role of sensory imagination. According to *Simulationism*, imaginative projection involves using one's internal state as a *model* for another's. A model, in the sense that it is the *similarity* between the model and target that determines the success of the modelling exercise (which is therefore a matter of degree). Moreover, and importantly, what kind of similarity is relevant depends on the use that the model is being put to (it's functional role), which can be changed in a flexible and context-dependent way. Thus, even though we may have a firmly established practice of restricting the targets of imaginative projection to conscious beings, there is no reason why the practice couldn't be extended if the use of the modelling state was changed somewhat. This makes sense on a deflationary pluralist metaphysics, because there is no profound dissimilarity between us and a merely pseudo-conscious being. For example, if we had two beings

---

[18] "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)

Barely and Almost, such that one was barely similarly enough to us to count as conscious and the other was barely too dissimilar, any attempt to model their internal states with our internal state would be *comparably successful*, because they are *comparably similar*. There is no deep ground for the appropriateness of imaginative empathy (no "inner light") that would decisively determine that trying to imaginatively relate to Almost would be like trying to relate to a rock.

On this view, imaginatively relating to a pseudo-conscious being (especially if we consider a fairly *reliable* and fairly *accurate* act of imaginative simulation) would be comparable to "perceiving" a virtual environment. Some theorists might be reluctant to count this as "really" perception (especially if being "perceptually manifest" had moral or normative importance), but one can understand a philosophical point of view that treats the difference as only superficial (see Chalmers (2022)).

I do not have space to properly develop and defend the Simulationist point of view here, but let me defend it against two important objections.

First, like the ice-people, you might object that my conscious state resembles many things in many different ways; therefore, I could use it to model many different things in different ways. So, surely there is more to imagining another's mental state than *mere* modeling or simulation. My response to this is simply to bite the bullet. Note that your experiences already have a multiplicity of modeling uses – in particular, they are *also* models of the manifest environment (which many also think is "illuminated"!). What an experience is used to model will make different functional roles appropriate. And if we are modeling a *mental state* through simulation (even if not a conscious one), that is very different use from a perceptual use. But in theory, yes, experience can be used to model anything that it is useful to model. In this sense, *everything is illuminated*![19]

A second important objection is that it is simply part of the meaning of "phenomenal consciousness" that all and only phenomenally conscious beings are such that we can appropriately imaginatively relate to them. So my view is ruled out a priori. My response to this is to concede that this is indeed a central strand in our thinking about consciousness, and if we want, we can doggedly hold onto it as a conceptual truth. This results in what I call the "imaginative conception" of consciousness. I think there's nothing wrong with making such a stipulation, but then it must be conceded that that this notion is going to be revisionary in other ways. In particular, it's arguably a matter

[19] But isn't it is simply part of the *content* of states of imaginative relating that it is a conscious state that is being represented in the other being? Thus, isn't imagining the inner-life of a non-conscious being bound to involve misrepresentation? I say no! I think that whether this is part of the content of imaginative states depends on *how they being used*. I reject the idea that some such content is simply baked into the phenomenology of a state of imagining, completely independently of how it is used (indeed I would say this about all contents of experience). Thus we could deliberately and unproblematically use states of imagining to simulate the inner states of non-conscious beings. Or we could imagine situations where it is the *function* of such states to simulate non-conscious beings – for example, suppose that we evolved living together with another species that does not have conscious states, but large part of the function of imagination was to get a grip on the inner life of this other species.

of degree how successfully you can imaginatively relate to another being, it is an anthropocentric matter (similarity to *me* is what matters), and the group of entities that resemble us in the relevant ways need not have any deep similarity to each other. Perhaps most importantly, the result will be that on the imaginative conception, we must reject empirical realism and thus the idea that we can determine which "theory of consciousness" is correct by investigating which natural property (or properties) plays the consciousness-role in us. Instead, learning what plays the consciousness role at best informs us what the base property is that is being used to simulate other beings; successful simulation need not require that they actually *have* this property, however.

This is therefore a case where my "conceptual pluralism" kicks in as more philosophically fundamental to my position than deflationary pluralism (which pairs with empirical realism). In the final section of the paper I consider another view, Normativism, where my response has a similar structure.

Response 3 : The Normativist View

As we have seen, on my view, not being conscious is not an obstacle to having a consciousness-like state which matters in the same way. A natural reaction to this is to wonder why it isn't correct to simply classify (perhaps as purely a stipulative matter) all the beings that matter in the way that conscious beings matter, as *thereby* bona-fide cases of conscious beings. Now, as above, I think we can legitimately get this result through *conceptual stipulation*, giving us a "normativist" concept of consciousness. But as with the "imaginative" concept, I will argue that the normativist concept is revisionary, and there probably not our ordinary concept of consciousness (or the one typically in play in philosophical debates). However, a version of it could, in some contexts, be useful to adopt in a revisionary spirit.

The Normativist puts the normative significance of consciousness center stage in defining what consciousness is. On their account, to be a conscious being just *is* to have a consciousness-like state that has the special significance of consciousness. That is, Phenomenal Consciousness = Quasi-consciousness. This rules out quasi-conscious beings that are not conscious just by conceptual fiat.

One helpful analogy here might be with belief-desire psychology. Some theorists claim that propositional attitudes concepts like "belief" are normative concepts in the sense that one is engaging in a special kind of rationalizing explanation in ascribing them, and so is making normative claims about the rational relations between a subject's states in ascribing them – they are a rational being complying with rational norms. Having a belief just is having a state playing a certain rational role, so the idea of quasi-beliefs that play the same role but aren't really beliefs makes no sense (see Davidson (1973), Dennett (1971)).

Normativism pairs naturally with strong normative realism, the view that there are fundamental mind-independent facts about what is morally (or epistemically) important. We can then picture the Normativist agreeing with the deflationary pluralist that there is no deep non-normative natural joint between conscious beings and the rest, but holding rather that the objective selection of a particular, arbitrary seeming,

physical/functional property as objectively strongly *normatively* significant is what the deep distinction between consciousness and other similar looking properties consists in.

On a "strong normativist" position, when we ascribe consciousness to a being, we are making a claim that is true only if certain normative conditions hold – it is true only if they are in a state that has a certain special practical (or epistemic) significance. A weak normativist does not put normative conditions in the truth conditions for such statements, but rather involves them in a different way, which I'll get to momentarily.

There are several problems with the strong normativist view. For one, it seems to reverse the normal order of explanation: we don't think that a creature is conscious because they are morally important, but rather they are morally important because they are conscious. So Strong Normativism is a strange inversion of our ordinary thinking. By way of analogy, caring about creatures because they are conscious is like caring about them because they are intelligent, which also doesn't strike us as, by its very nature, a morally normative property (rather it's a *functional* property).

A second problem is that that the view ties the metaphysics of consciousness to the metaphysics of normativity. For example, if I was a normative anti-realist of a certain stripe – an error theorist or expressivist perhaps – then this view would commit me to a corresponding error theoretic or expressivist view of statements about consciousness. But it's quite implausible that normative anti-realism commits one to denying that consciousness exists, or saying that statements about consciousness are not wholly factual statements. Related to this, the intuitive picture is that we know what consciousness is through simply *confronting it in introspection*, not through commitment to a certain normative way of thinking. So if the strong normativist is saying that we explicitly conceive of consciousness a priori as strongly normative, this seems to be mistaken (perhaps they could retreat to an a posteriori version of the view, but that has other problems).

A retreat from strong normativism is a weak normativist position, according to which consciousness is a completely natural property, so there are no normative commitments involved in its application condition; nonetheless, normative commitments come into play more indirectly in determining which natural property it is – they play a role in our reference fixing intentions (or at least our intentions in selecting a concept to be associated with the word "consciousness", if that is a distinction with a difference). I think this view is still implausible for reasons similar to the strong normativist view, although it crucially avoids some issues, such as tying consciousness-realism to normative-realism. However, it is very much worth thinking about, because I believe we may want to adopt it as a *revisionary* concept for certain moral or political purposes.

Here a comparison with other ethically important subject-categories, such the concept of a person, of a family, or categories of gender or race, is helpful. Consider, "family" for example. In an explanatory context in biology, it might be important to be only talking about biological families as "families". But if we are considering who gets to be a "family" for moral or political purposes, it would be obnoxious to insist that only biological families count (and that would be true, even if the word "family" had been typically used in ordinary contexts with that restriction in mind). We may well favor a quite heterogenous variety of social arrangements as "families" for the purpose of say,

family law, or for how we talk about families in everyday life (for example, we include adoptive families as families, and we count adoptive parents as parents).

A similar point holds for gender[20], where I see patterns of thinking that mirror the consciousness case. When it comes to womanhood (or manhood), there is a tendency to think as if there is a deep essence of womanhood, and we need to have a heated debate about who has it. But isn't there simply a plurality of womanhood-ish properties that emphasize in different ways different diagnostic features, and which we should want to talk about will depend on the motivations in a given context (see Saul (2012), Diaz-Leon (2016) for related views)? In scientific-explanatory contexts, we may be interested in a biological property, a cultural or social property (e.g. Witt (2011)), a psychological property (e.g. Bettcher (2013)), or a hybrid of these categories (and presumably there are many relevant properties in each of these categories). But importantly, there are also many contexts where political or ethical motivations are highly relevant (i.e. everyday contexts or in legal/political/moral debates) (Haslanger (2012)). For example, consider the statement "trans women are women", the negation of which is sometimes asserted by people opposed to recognizing the gender identity of trans people. The way I understand the point of such a statement, it's irrelevant whether certain causal-explanatorily useful ways of understanding gender or sex in biology or elsewhere exclude trans women[21], or even whether "woman" as uttered from the mouths of ordinary folks in the past has been trans-inclusive (in much the way that these factors are also (I assume) irrelevant for who should be described as a "family", what counts as "marriage" etc etc)[22]. What those of us who say this care about is expressing a commitment to freedom and dignity for trans people, including the view that we *ought*

[20] For the comparison between family and gender, see Grace-Chappell (2018).

[21] As an aside, I would note that the extent to which this is the case is often overstated, for example in the unfortunate use of terminology like "biological man" / "biological woman". Many trans women/men have transitioned biologically through hormones and surgeries, and so *do* count as women / men on many reasonable biological criteria.

[22] See Byrne (2023) for a prominent example of the kind of view I'm responding to here.

to categorize trans women as women (and trans men as men)[23] (compare "adoptive families are families[24]!", "same-gender couples can be married!" etc).

Similarly, I could envisage situations where it would be morally or politically well-motivated to understand "conscious" in a weakly normative way. For example, suppose that octopi do not have "consciousness" in a scientific realist sense (because that turns out to refer to a quite specific feature of mammalian brains), but nonetheless we enlightened philosophers know they have an analogous neural state that matters in the same way. Using "consciousness" in a (perhaps revisionary) inclusive sense could be morally motivated if we want people to treat octopi appropriately. I mention this case because it is actually realistic, but note that in the more hypothetical Baker-style scenario where two intelligent species both care very much that the *other* group see them as conscious beings (despite their having very different brains), the motivation for using the normativist concept would be a lot stronger.

I expect some will react with a sense that this kind of conceptual engineering involves a disagreeable kind of false-consciousness. "They aren't *really* conscious, but you are insisting on pretending that they are because of your weird morality. I refuse to submit to this Orwellian mind-control!". However, I see this reaction as of a piece with the misleading 'inner light' or "big difference" thinking that tends to rule our thinking about consciousness. Once we have a more pluralistic view at the metaphysical level, it is natural to see it as far more discretionary what we pick out with the words "conscious being". Setting aside moral considerations, octopi could well be "conscious" in a number of reasonable senses even if they don't have what we have – e.g. they may

---

[23] If one takes this position, that doesn't mean that the truth condition of all uses of "woman" in the trans-inclusive sense are literally assertions about how we ought to be using "woman". In many contexts they are probably best interpreted as purely factual assertions about a group of people who happen to include trans women. The point is that politically normative considerations can go into determining which concept of "woman" is best (i.e. morally / politically best for everyday use, not best for scientific causal-explanatory purposes), and under certain circumstances, I can be expressing or implicitly asserting my view about which concept these political considerations point towards, in saying something like "trans women are women" in, e.g. a political debate.

A similar point can be made about "belief" and "desire". I used to think that a serious problem for Normativism about these concepts is that even if Normativism is true there will surely be purely functional (or physical) necessary and sufficient conditions for, e.g. a person believing that grass is green, so why not take believing that grass is green to be such a purely functional property? I now think this offends the letter but not spirit of belief-desire normativism, because one can hold the position that what selects these functional properties as belief properties is their role in our practice of rationalizing explanation; they count as beliefs and desires with certain contents because of the way their causal role approximates a rational ideal. Again, we can think of this as a kind of meta-semantic normativism, where now the relevant norms are rational norms rather than political norms. Conceivably, the (perhaps purely functional or physical) properties that this practice picks out might be different from any that would picked out by a more purely causal-explanatory scheme (so there is some truth in McDowell's (1985) view).

[24] See Grace-Chappell (2018).

have a state that plays the consciousness-role, and they may be legitimately the targets of our imaginative empathy. And so once moral considerations kick in, the sense that we are willfully misrepresenting them if we call them "conscious" is actually much less compelling that it first appears[25].

So : although I doubt that our ordinary concept is the weak normativist concept, it could potentially have a legitimate role similar to other morally/politically motivated concepts that also have "scientific naturalist" analogs.

Summing Up

At a crude first pass, you could sloganize my view as "zombies can matter too!". But hopefully I've said enough here to make it clear that what really matters to me is the underlying conceptual pluralism. There are various important strands in our thinking about phenomenal consciousness, and the mainstream view is that we can find a single magnificent property in the world that does all the work for us. Part of my project is to pull the strands apart to explicate different variants on the concept of consciousness. I think when we do this, we can see that it's plausible that they do not all pick out the same phenomena in the world. Here I've tried to make the case that the scientific naturalist project of figuring out what consciousness really is, should not be construed as automatically also telling us which beings matter, or who we can imaginatively relate to.

Citations

Adamson, P., & Ganeri, J. (2020). *Classical Indian Philosophy: A history of philosophy without any gaps, Volume 5*. Oxford University Press.

Allen, C. (2017). "Associative Learning". In the Routledge Handbook of the Philosophy of Animal Minds, Andrews and Beck (eds.). Routledge.

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in brain research*, *150*, 45-53.

Baker, D. (2016). "The Hunter Captain." Featured in Escape Pod, episode 526.

[25] Simion (2018) argues that politically motivated conceptual engineering projects could be defective if they ignore epistemic constraints on concepts, to the effect that they should carve nature at the joints in order to be representationally adequate ("a concept should be ameliorated only insofar as this does not translate into epistemic loss" p.914). I sense something of the "false consciousness" complaint in her work; however her position ignores the fact that we can have a plurality of concepts that have different motivations – there is no need to choose between political and epistemic motivations. Also, if such a view were intended to motivate more exclusive uses of "consciousness", "woman" etc. (to be clear, Simion does not address these cases), it probably requires rejecting the kind of metaphysical pluralism I advocated here.

Bettcher, Talia Mae (2013). "TransWomen and the Meaning of 'Woman'". In A. Soble, N. Power & R. Halwani (eds.), Philosophy of Sex: Contemporary Readings, Sixth Edition. Rowan & Littlefield 233–250.

Bloom, P. (2017). Empathy and its discontents. *Trends in cognitive sciences*, *21*(1), 24-31.

Bradford, G. (2023). Consciousness and welfare subjectivity. *Noûs*, *57*(4), 905-921.

Byrne, A. (2023). *Trouble with gender: Sex facts, gender fictions*. John Wiley & Sons.

Chalmers, D. J. (2022). *Reality+: Virtual worlds and the problems of philosophy*. Penguin UK.

Cutter, B. (2017). the Metaphysical Implications of the Moral Significance of Consciousness. *Nous-Supplement: Philosophical Perspectives*, *31*(1), 103–130

Davidson, D. (1986) A coherence theory of truth and knowledge. In Ernest LePore (ed.), *Truth and Interpretation. Perspectives on the Philosophy of Donald Davidson*. Basil Blackwell.

Davidson (1973) Radical Interpretation. *Dialectica* 27 (1):314-328.
Dennett, D. (1971) Intentional Systems. Journal of Philosophy 68 (February):87-106

Diaz-Leon E (2016) Woman as a politically significant term: a solution to the puzzle. Hypatia 31(2):245–258

Dorr, C., (2002) "Non-Cognitivism and Wishful Thinking", Nous 36(1), pp. 97–103.

Enoch, D. (2003). "How Noncognitivists Can Avoid Wishful Thinking". The Southern Journal of Philosophy (41) 527-545.

Field, H.,(1980) Science Without Numbers: A Defense of Nominalism. Princeton: Princeton University Press.

Grace-Chapell, S. (2018). Trans Women/Men and Adoptive Parents : An Analogy. APA Blog Post 7/20/2018. https://blog.apaonline.org/2018/07/20/trans-women-men-and-adoptive-parents-an-analogy/

Haslanger, S. (2012). 'Gender and Race: (What) Are they? (What) Do we want them to be?' in Resisting Reality. Oxford OUP, p. 221–247.

Kagan, S. (2016). What's Wrong with Speciesism? (Society for Applied Philosophy Annual Lecture 2015). *Journal of Applied Philosophy*, *33*(1), 1-21.

Kriegel, U. (2019). The value of consciousness. *Analysis*, *79*(3), 503-520.

Lee, A. Y. (2022). Speciesism and sentientism. *Journal of Consciousness Studies*, *29*(3-4), 205-228.

Lee, G. (2013). Materialism and the Epistemic Significance of Consciousness. In Kriegel (ed.) Debates in Philosophy of Mind.

Lee, G. (2018). Alien Subjectivity and the Importance of Consciousness. In Pautz and Stoljar (ed.) Blockheads. MIT Press.

Lee (manuscript). The Search for the Inner Light: Finding Consciousness in a Physical World.

Lenman, J. (2003). Noncognitivism and Wishfulness. Ethical Theory and Moral Practice 6: 265–274

Lin, E. (2021). The experience requirement on well-being. *Philosophical Studies*, *178*(3), 867-886.

McDowell, J. (1985). Functionalism and Anomalous Monism. In Brian P. McLaughlin & Ernest LePore (eds.), *Actions and Events: Perspectives on the Philosophy of Donald Davidson*. Blackwell

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* Oxford: Oxford University Press

Papineau, D. (2007) Phenomenal and perceptual concepts, in Alter, T. & Walter, S. (eds.) Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism, Oxford: Oxford University Press.

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

Pautz, A. (2017). The Significance Argument for the Irreducibility of Consciousness. *Philosophical Perspectives*, *31*, 349-407.

Prinz, J. (2011). Against empathy. *The Southern Journal of Philosophy*, *49*, 214-233.

Saul, J. (2012) Politically significant terms and philosophy of language: methodological issues. In *Out from the Shadows: Analytical Feminist Contributions to Traditional Philosophy*, ed. Sharon Crasnow and Anita Superson. Oxford University Press.

Shepherd, J. (2018) Consciousness and Moral Status, New York: Routledge.

Siewert, C. (1998). *The significance of consciousness*. Princeton University Press.

Siewert, C. (2021). 'Consciousness: Value, Concern, Respect.' *Oxford Studies in Philosophy of Mind, 1.*

Simion, M. (2018). The "should" in Conceptual Engineering. Inquiry (61) 914-928.

Singer, P. (2009) Animal Liberation, updated edition. New York: Harper.

Strawson, G. (1994). Mental Reality. MIT Press.

Van der Deijl, W. (2021). The sentience argument for experientialism about welfare. *Philosophical Studies*, *178*(1), 187-208.

Witt, Charlotte (2011). The Metaphysics of Gender. OUP USA.