Draft paper, 31 March 2021

Ethics and the Mind-Body Problem

Jonathan Birch

Centre for Philosophy of Natural and Social Science,
London School of Economics and Political Science,
Houghton Street, London, WC2A 2AE, UK.
j.birch2@lse.ac.uk
http://personal.lse.ac.uk/birchj1

Abstract

Phenomenal consciousness has an important role in ethics: it is plausible that it is at least a necessary condition for a distinctive kind of moral status. There is a mismatch between this ethical role and an a posteriori (or "type-B") materialist solution to the mind-body problem. I argue that, if type-B materialism is correct, then the reference of the concept of phenomenal consciousness is indeterminate between properties that are coextensive in the case of (fully conscious) humans but have radically different extensions in non-human animals. The result is that the moral status of many non-mammalian animals is indeterminate. Some ways of managing this disturbing indeterminacy are evaluated.

Phenomenal consciousness matters in ethics. It is, in particular, relevant to assessments of moral status. To give two examples: (i) our moral obligations towards a patient in a permanent vegetative state plausibly differ from our obligations towards a patient who is minimally conscious, because there is, at least some of the time, something it's like to be the latter but not the former (Kahane & Savulescu 2009); (ii) if we could show that crabs and lobsters have a capacity for phenomenal consciousness—that there is something it's like to be a crab—this could form the basis of an argument that they have morally significant interests in avoiding suffering (Birch 2017). The nature of the relationship between phenomenal consciousness and moral status merits further discussion (see Section 1), but it is widely agreed that there is such a relationship.

Given this, there are possible solutions to the mind-body problem that would create serious trouble for ethics. This is most obviously true of eliminative materialism about consciousness, also known (in recent literature) as strong illusionism.[1] This view holds that phenomenal consciousness does not exist. If strong illusionism is correct, and if phenomenal consciousness plays an important role in ethics, ethics may need significant revision. Kammerer (2020) has called this the "normative challenge" for strong illusionism. But perhaps ethicists should not be unduly worried about this, since the challenge can be avoided by denying strong illusionism.

My focus here will be on a significantly more popular approach to the mind-body problem: a posteriori materialism, or "type-B" materialism in Chalmers' (2010) terminology. The type-B materialist holds that, although there is an epistemic gap between physical and phenomenal truths that is exploited by anti-materialist arguments, this epistemic gap can be explained without positing any ontological gap between physical and phenomenal facts. In a particularly influential version of the view, the epistemic gap is to be explained by appeal to phenomenal concepts. These are concepts we use for thinking about our experiences, and they are posited to be radically different from the concepts we use to think about the physical world, with the result that there are no a priori connections between phenomenal concepts and non-phenomenal concepts (Loar 1990; Carruthers 2000; Papineau 2002). This has come to be known as the "phenomenal concept strategy" (Stoljar 2005; Balog 2012).

Since the type-B materialist does not deny the existence of phenomenal consciousness, it is not so obvious that that the view holds revisionary consequences for ethics. But I will argue that it does. I will argue that, if type-B materialism is true, then, given a plausible account of the relationship between phenomenal consciousness and moral status, facts of great ethical significance turn out to be indeterminate.

The next section argues for the claim that a capacity for phenomenal consciousness is necessary condition for a step change in moral status. The subsequent section argues for the claim that, if type-B materialism is true, then the reference of the concept PHENOMENAL

[1] I am using the term "strong illusionism" here in the sense of Chalmers (2018). Frankish (2016) uses the term in a slightly different sense.

CONSCIOUSNESS is indeterminate between physical properties that are coextensive in the case of (fully conscious) humans but have very different extensions in non-human animals. These premises combine to yield the disturbing conclusion that, in many cases, it is indeterminate whether or not an animal has the kind of moral status that is associated with phenomenal consciousness. I then consider various ways to manage this indeterminacy—and find all the options troubling.

1. Phenomenal consciousness, valence and moral status: in search of common ground

A being with *moral status* has at least some interests that matter morally *in their own right*, and not just because they matter to some other being. I want to leave open the possibility that non-conscious beings can have moral status. Some maintain, for example, that ecosystems, or foetuses not yet capable of forming conscious states, or unconscious patients who will never regain consciousness, have moral status. On some ethical theories, such as hedonic utilitarianism (Singer 1995) and Regan's (2004) animal rights theory, the capacity to have conscious experiences is a necessary condition for moral status. However, I want to avoid assuming any particular ethical theory.

In order to remain neutral on these claims, I will say merely that a *step change* in moral status is associated with a capacity for conscious experience. My hope is that this can be a point of broad ethical consensus. The term "step change" is intended to be neutral between a step from zero moral status to moral status, and a step from a pre-existing, basic form of moral status to a new form. I will use the rather awkward term **"phenomenality-linked moral status" (PLMS)** to label the stepped-up form of moral status associated with phenomenal consciousness.[2]

What can we say about the nature of the step change without losing the desired neutrality? I think it can be a point of wide agreement that a conscious being's interests ought to factor into moral deliberation in a distinctive way, such that these interests receive greater weight than they would have received if they were the interests of a non-conscious being, if we allow that the non-conscious have interests at all. When we find evidence that a patient thought to be permanently unconscious is in fact having conscious experiences at least some of the time, their interests (e.g. in adequate nutrition and hydration) appropriately receive greater moral weight. Likewise, when we find evidence that an animal thought to be wholly unconscious has conscious experiences, their interests (e.g. in humane treatment and good welfare) appropriately receive greater moral weight.

Is it phenomenal consciousness *as such* that explains this step change, or a special type of phenomenally conscious state? Various authors (e.g. DeGrazia 1996; Singer 2011; Korsgaard 2018; Shepherd 2018) have proposed that it is not conscious experience as such but *valenced* conscious experience that explains the step change. Valenced conscious experiences are experiences that feel bad or feel good. Negatively valenced experiences include pain,

[2] See also Shevlin (2020a) on "psychological moral patiency". The quest for neutrality among substantively different ethical theories seems to lead inevitably to awkward terminology.

pleasure, distress, anxiety, boredom, tiredness, hunger and thirst. These experiences typically motivate actions to alleviate them. Positively valenced experiences include pleasure, joy, warmth, comfort, satiety and excitement. These experiences typically motivate actions to sustain them. Any capacity for valenced conscious experience will ground interests to escape, or to sustain, certain types of experience, and these interests matter morally.

One might wonder: could valence alone be enough to explain the step change, independently of phenomenal consciousness? This depends, first of all, on whether there can be valence without phenomenal consciousness. On one account of valence (that of Carruthers 2018), valence is the nonconceptual representation of value, and it is plausible that a state may represent value nonconceptually without being phenomenally conscious. Indeed, it seems necessary to posit non-conscious representations of value in order to explain the possibility of subliminal motivation (Pessiglione et al. 2007) and subliminal instrumental conditioning (Pessiglione et al. 2008, though cf. Skora et al. 2021). For example, Pessiglione et al. (2007) presented evidence that subjects find larger sums of money more strongly motivating than smaller sums, even if the amounts of money are presented subliminally. This suggests that the larger sums are assigned higher value than the smaller sums by unconscious motivational processes.

Granting that there can be valenced states that are not phenomenally conscious, it makes sense to ask whether a capacity to form valenced mental states might already suffice for the step change in moral status we took to be associated with phenomenality. But I contend that a capacity for representing value would *not* suffice if the representations were always non-conscious. A reinforcement learning algorithm represents value, but it is usually taken to do so non-consciously, and in a way that intuitively confers no moral status on it. Tomasik (2014) has argued that reinforcement learning algorithms do possess moral status (inspiring an organization called "People for the Ethical Treatment of Reinforcement Learners") but Tomasik rests his case on the idea that algorithms may have a capacity for phenomenal consciousness, not on the idea that non-conscious representation of value is already sufficient.

One might also wonder: could phenomenal consciousness alone be enough to explain the step change, independently of valence? At least in principle, there can be phenomenal consciousness without valence: experiences that feel neither bad nor good. It is not clear that humans can have such experiences: our overall conscious state arguably always contains an element of mood. But we can conceive easily enough of a being that has a subjective point of view on the world in which non-valenced states feature (it consciously experiences shapes, colours, sounds, odours, etc.) but in which everything is evaluatively neutral. Would such a being have the distinctive kind of moral status associated with conscious experience? If it would, it raises the possibility that valence, not phenomenal consciousness, is a redundant condition for PLMS.

Chalmers (quoted in Wiblin et al. 2019) has offered the example of a Vulcan. The original Vulcans in Star Trek are not wholly without valenced experiences, but we can conceive of a

"philosophical Vulcan" in which all valenced experience is dialled down to nothing, while leaving conscious perceptual experience, conscious thought, imagination, and episodic memory in place. Carruthers (2005) also considers such a being, which he names "Phenumb". Intuitively, a philosophical Vulcan has some moral status: it would be wrong (for Chalmers, "monstrous") to destroy such a being for no reason at all. In a similar vein, Kriegel (2019, p. 515) suggests that "we have duties towards not only human beings but all conscious beings, including non-human conscious animals: these animals ought to be treated as ends, quite independently of the hedonic quality of their lives".

In opposition, Lee (2019) offers the example of a being that experiences a maximally simple non-valenced experience, such as an experience of slight brightness. The example is reminiscent of Ginsburg and Jablonka's (2007, p. 220) example of a being who experiences only "white noise"—Ginsburg and Jablonka speculate that the first conscious experiences in the earliest nerve nets were something like this. Is the presence of conscious experiences of slight brightness, or white noise, enough to make the difference between the presence or absence of PLMS? Plausibly, it is not.

Can we reconcile the conflicting intuitions elicited by these cases? Recall that moral status is crucially tied up with the possession of *interests*. What the philosophical Vulcan shows us, I suggest, is that interests can be grounded independently of valence. An autonomous rational being capable of reflectively endorsing goals and projects has interests, whether or not it has experiences of frustration, joy (and so on) associated with the success or failure of those projects. The step change in moral status associated with phenomenal consciousness is the change that comes when events that promote or thwart a being's interests are registered in experience. The significance of valence is that it provides the necessary grounding for interests in beings who lack rational agency.

The overall picture, then, is one on which a capacity for phenomenal consciousness as such is a *necessary condition* for PLMS. Facts sufficient to ground interests, either in the form of valence or the autonomous, rational endorsement of goals, must also be in place, and the promotion or thwarting of those interests must register in experience. A consequence of this picture is that the moral status of non-human animals depends a great deal on which ones are capable of forming phenomenally conscious states. This is because, for many animals, valence is undoubtedly in place: we have good evidence of their ability to form representations of value and disvalue that guide flexible decision-making.

To illustrate, consider Crook's (2021) recent study of responses to injury (injection of acetic acid) in Bock's pygmy octopus (*Octopus bocki*). Injured octopuses showed directed grooming at the site of the acetic acid injection that was abolished by a local anaesthetic, lidocaine. More than this, they came to prefer chambers in which they had been placed after receiving lidocaine, and disprefer chambers in which they had received an injury. This type of evidence is widely regarded in animal welfare science as evidence of pain, an exemplar of a valenced experience.

There is evidence of this general type (reviewed in Sneddon et al. 2014) for mammals, birds, reptiles, fish, cephalopod molluscs, and decapod crustaceans. But this evidence invites the objection that to show valenced *experience*, it is not enough to show mere representation of value and disvalue—you also have to show that these representations are consciously experienced (Dawkins 2021; Paul et al. 2020).

A step change in moral status for many animals thus hinges on the question of whether or not their representations of value and disvalue are consciously experienced. This is a hard question to answer conclusively, presenting serious methodological challenges that are not the topic of this article, but we can at least hope that the science of consciousness will one day be able to overcome these challenges, and that in the meantime it will produce relevant, albeit inconclusive evidence (for discussion of the methodological challenges, see Birch 2020). This hope, however, rests on there being a determinate fact of the matter to be found.

2. Mild and radical indeterminacy

On any type-B materialist view that identifies phenomenal consciousness with a non-fundamental, higher-level natural kind, some degree of sorites-style vagueness seems likely to arise at the edges, since this is a common and perhaps ubiquitous feature of such kinds. To illustrate, suppose we think phenomenally conscious states are patterns of thalamocortical activity supported by pyramidal neurons in layer 5 of the neocortex (a hypothesis set out by Aru et al. 2019). Various neurobiological kinds are in play in this hypothesis: thalamus, neocortex, pyramidal neuron, layer 5. If we could see the evolution of these traits unfolding over time, we would expect to see borderline cases of all of these kinds. For example, when the laminated structure of the mammalian neocortex was in the process of evolving, it seems likely that there would have been borderline cases between laminated and non-laminated cortices.

This vagueness clashes with the intuition that phenomenal consciousness must always be determinately on or off, with no borderline cases, an intuition Anthony (2006) has called the "intuition of sharpness" (see also Simon 2017). I take it that type-B materialists who regard conscious experience as an evolved, non-fundamental natural kind must reject the intuition of sharpness and assert that borderline cases of conscious experience are possible (see also Lee 2017; Godfrey-Smith 2020). That is simply part of the price of the view: a sense in which it is counterintuitive, to be added to the counterintuitiveness (for some) of rejecting the possibility of zombies.

Several authors have noted that sorites-style vagueness, combined with a close connection between phenomenal consciousness and moral status, threatens to lead to borderline cases of moral status (Dunaway 2016; Cutter 2017; Godfrey-Smith 2020). However, there is no particular reason to expect sorites-style vagueness to lead to a widespread meltdown of our ethical deliberations regarding animals. Firstly, there may not be any extant species occupying the borderline region for any of the relevant kinds (for example, all extant mammals determinately have a neocortex). Secondly, even if there are some extant borderline cases, they are likely to represent a very small fraction of species. This is implicit

in the idea that neurobiological and cognitive kinds are natural kinds that in some sense "carve nature at the joints", with most cases lying between the joints. If we found a putative cognitive/neurobiological kind such that a vast majority of animals were borderline cases with respect to that kind, this would lead scientists to revise the kind. To be sure, sorites-style vagueness regarding moral status has interesting implications for meta-ethics, as Cutter (2017) has noted, since meta-ethical positions incompatible with vagueness about moral obligation will also be incompatible with plausible forms of materialism. Yet in so far as it need not threaten our practical deliberations about what to do outside of a small number of cases, sorites-style vagueness is a mild form of indeterminacy.

Several prominent defenders of type-B materialism (Papineau 1993, 2002, 2003, 2020; Carruthers 2019; Balog 2020), have noted that it raises the spectre of a much more pervasive and troubling form of indeterminacy. The threat here is that the reference of the concept PHENOMENAL CONSCIOUSNESS is indeterminate between properties that are coextensive in the paradigm case of a human who can report their experiences, but that differ radically in their extensions outside of this paradigm case.

Why would that be? For most concepts, the story of how the reference of the concept is fixed will normally give at least some role to *conceptions*: the stock of beliefs a subject associates with the concept. We need not be internalists about mental content to allow *some* role for conceptions in fixing reference. For example, my correct belief that platypuses are egg-laying mammals is likely to form part of the explanation for why my concept PLATYPUS successfully refers to platypuses.

Yet for the type-B materialist, most if not all of the conceptions we associate with phenomenal consciousness are misconceptions. The type-B materialist parts ways here with the analytic functionalist (e.g. Lewis 1983), who takes us to conceive of conscious experiences as states that play a certain type of functional role. The type-B materialist agrees with the dualist that there are *no a priori links between phenomenal consciousness and any functional concept*. The concept of phenomenal consciousness is not even partly the concept of a functional role or its realizer. Instead, we tend to think that phenomenal consciousness is non-functional, irreducible, intrinsic, qualitative, primitive, ineffable, physically inexplicable, unknowable from the outside, that its essence is fully revealed to us first-personally, and so on. But the type-B materialist parts ways with the dualist by holding that these intuitive judgements are false. Rejecting these conceptions allows the rejection of dualism, but it means the type-B materialist must hold that PHENOMENAL CONSCIOUSNESS refers to a physical property *in spite of* the conceptions associated with the concept, not because of them.[3]

I say "most, if not all" to allow that some fairly minimal conceptions may survive the type-B

[3] Alternatively, the type-B materialist may argue that no conceptions at all are associated with PHENOMENAL CONSCIOUSNESS, because it is a bare, indexical concept, its content being roughly "this sort of thing" (see Carruthers 2019). This is still a view on which conceptions play no role in fixing reference.

materialist bonfire. In particular, a type-B materialist will endorse the conception that phenomenal consciousness is a property that the referents of our phenomenal concepts have in common, and the conception that phenomenal consciousness has effects in the physical world (without being definable in terms of these effects). The problem is that the surviving conceptions are far too minimal to triangulate a single physical property.

The result is that, when giving a positive account of how the concept's reference is fixed, the type-B materialist must work with a very limited set of resources. When explaining the reference of a concept like ARTHRITIS, we can (as Burge 1979 argued) appeal to deference to experts to show how successful reference can be compatible with serious misconceptions. But it would be implausible to appeal to deference to experts to explain the reference of PHENOMENAL CONSCIOUSNESS, since it is generally supposed to be a concept that we can grasp intuitively by reflecting on our own conscious experiences. There are no textbook definitions for us to consult; the textbooks refer us back to our own experiences.

With conceptions and deference off the table, there is not much the type-B materialist can say except that our applications of PHENOMENAL CONSCIOUSNESS successfully *track* a physical property, despite all our substantive misconceptions about the nature of that property. They track the property in the sense that all and only the states to which we are disposed to apply the concept first-personally in fact possess that property. The materialist can then argue that successful tracking is enough for successful reference, even in the presence of substantial misconceptions. But now indeterminacy looms, because our applications of PHENOMENAL CONSCIOUSNESS are likely to successfully track more than one physical property. These properties are coextensive in our own (first-person) case, and are generally coextensive in fully conscious humans (as opposed to humans in a minimally conscious state), but have very different extensions outside these cases.

I will focus here on a version of the problem from Papineau (2002), which I take to be the most troubling version. For Papineau, PHENOMENAL CONSCIOUSNESS successfully tracks properties at least two different levels of organization: a high-level functional property and its neuronal realizer (Papineau 2002, p. 214). Our first-person applications of PHENOMENAL CONSCIOUSNESS to our own states (as when I judge, for example, that I am consciously experiencing a perception of a blue sky, but am not consciously experiencing the digestion of my breakfast) will track a high-level functional property. The science and philosophy of consciousness gives us several important candidates for this functional property. In broad terms, it may be entry to a global workspace (Dehaene and Changeux 2011; Dehaene 2014) or something causally upstream of entry to a global workspace, such as entry to fragile short-term memory (Block 2007, 2011) or something causally downstream of entry to a global workspace, such as becoming the object of a higher-order thought (Rosenthal 2005; LeDoux 2019). Let us assume, perhaps optimistically, that the science of consciousness will eventually reach consensus about what this high-level functional property is, and let us call it property *F*.

Crucially, this high-level functional property, whatever it is, will be coextensive in humans

with a particular neuronal mechanism (or set of mechanisms) that realizes it. For example, entry to the global workspace may be coextensive with activation of a neuronal mechanism involving pyramidal neurons in the prefrontal cortex and the "global ignition" of many cortical regions (Dehaene and Changeux 2011; Dehaene 2014; Mashour et al. 2020). Again, let us assume optimistically that the science of consciousness will converge on a single such neuronal mechanism, and call it property *N*. I want to assume as little as possible about the nature of *N*. In particular, I leave open the possibility that *N* will be shared by all mammals and will not be specific to primates or to humans. What seems very unlikely at this stage, however, is that *N* will be shared by a wide range of non-mammalian animals. This is because we already have clear evidence that conscious experience is intimately related to mechanisms in the neocortex, a brain region that has evolved since the divergence of the mammals from other lineages (Dehaene 2014; Koch et al. 2016; Frith 2019; Aru et al. 2020). If *F* has evolved in other, non-mammalian lineages, then it must have a *non-cortical* neuronal implementation that differs substantially from its cortical neuronal implementation in mammals.

Papineau's point is this: there will be no way to resolve the question of whether PHENOMENAL CONSCIOUSNESS refers to *F* or to *N*. Moreover, this is not, he suggests, merely the result of epistemic limitations. This is a point of contrast with Block (2002), who assumes that there must be some fact of the matter about whether *F* or *N* is phenomenal consciousness, but argues that we cannot know this fact. Papineau contends that the reference of the concept PHENOMENAL CONSCIOUSNESS is indeterminate between *F* and *N*. We are disposed to apply the concept, in our own case, to states that instantiate both properties. There is nothing in the concept, or in its associated conceptions, or in our use of it, that could fix just one of these properties as the unique referent. They are equally eligible candidates for reference. And yet the distribution of *F* and *N* in the natural world may well be very different: *N* is likely to be specific to mammals for the reasons noted above, whereas *F* may turn out to be possessed by a very wide range of animals (birds, reptiles, fish, cephalopods, arthropods) which have evolved a different neuronal implementation of the same functional property. This point is highlighted in the context of global workspace theory (the source of one important candidate for *F*) by Dehaene, who writes "I would not be surprised if we discovered that all mammals, and probably many species of birds and fish, show evidence of a convergent evolution to the same sort of conscious workspace" (Dehaene 2014, p. 246). *F* but not *N* may also be possessed by non-living entities, such as future AI systems and robots, as emphasized by Dehaene et al. (2017).[4]

Carruthers, another prominent type-B materialist, arrives at a similar conclusion via a different route.[5] Carruthers (2019, pp. 155-160) draws an analogy with a person who

[4] Papineau's argument has received surprisingly little discussion. Taylor (2013) and Balog (2020) are exceptions. Taylor rebuts a distinct argument from Papineau (2002), the so-called "methodological meltdown" argument, but does not rebut the argument for referential indeterminacy between *F* and *N*.

[5] There are differences between Papineau and Carruthers that I lack the space to discuss here. For Carruthers, the main threat is not one of indeterminacy between *F* and *N*, but indeterminacy between functional properties specified at different grains of analysis. For example, the coarse-grained functional property of possessing a

sometimes remarks, of a neighbourhood, that it is "that sort of neighbourhood". Suppose there is more than one property that these judgements track, and that these properties happen to be coextensive in the part of the world where the person lives. Perhaps they track both low socioeconomic deprivation and low levels of gun ownership. Now we ask: which neighbourhoods would be "that sort of neighbourhood" in a different country where these properties are no longer coextensive? To settle such a question in practice, we could present the person with those neighbourhoods (or pictures of them) and see what they judge. Whether or not we actually do this, the person will still have dispositions to judge counterfactual neighbourhoods as "that sort" or "not that sort", and these dispositions may be enough to triangulate a single property as the referent of "that sort of neighbourhood".

For Carruthers, the concept PHENOMENAL CONSCIOUSNESS works like the phrase "that sort of neighbourhood". We acquire the concept by picking out various particular experiences first-personally, and then forming a concept of "that sort of thing". Our applications of the concept track different properties that are coextensive in our own case. We then ask: to which non-human mental states does this concept apply? As with "that sort of neighbourhood", the way to settle such a question would be to present that subject with the non-human mental states in question, first-personally, and see what they judge. We can't do that, but one might hope that—as in the neighbourhood case—we could use a subject's dispositions-to-judge regarding non-human mental states to triangulate a single physical property as the referent. But these counterfactuals, Carruthers argues, are non-evaluable: there is simply no fact of the matter about whether I would, or would not, first-personally judge a particular non-human mental state to be phenomenally conscious, given the chance.[6] Given this, Carruthers argues, we should accept that there is no fact of the matter about whether a non-human mental state is phenomenally conscious or not. We have run out of reference-fixing resources.

One possible way to resist the threat of radical indeterminacy is suggested by Shea (2012). Shea notes that there is one more reference-fixing resource available to the type-B materialist: the role played by PHENOMENAL CONSCIOUSNESS in inductive inferences.[7] Return here to "that sort of neighbourhood". If we find that the person uses this phrase in inductive inferences (e.g. "it's that sort of neighbourhood, so litter on the street will soon be picked up.") we can ask what property explains the success of these inferences. This property, argues Shea, is a more eligible candidate for reference than a coextensive property that does less work, or no work at all, in explaining the inductive utility of the concept. For example, if low socioeconomic deprivation explains why "that sort of neighbourhood" is inductively fruitful, but low gun ownership does not, low socioeconomic deprivation is a

---

global workspace *of some kind* (F_1) is coextensive in humans with possessing a global workspace *with the specific cognitive architecture of the human global workspace* (F_2). Papineau (1993, p. 124) discussed a similar issue very briefly in earlier work. Shevlin (2020b) discusses a related idea under the heading of "the specificity problem". See also footnote 9.

[6] Papineau (1993, p. 126, footnote 23) makes a similar point, briefly.

[7] Shea (2012, p. 335): "irrespective of whether we conceive of [phenomenal consciousness] as being the occupant of a functional role, our concept refers to whatever property underpins the successful inductions in which it is deployed." Shea credits this idea to Millikan (2000).

more eligible referent—it attracts the reference of the concept more strongly.

Can we use inductive considerations to discriminate between *F* and *N*? It is plausible that we sometimes use PHENOMENAL CONSCIOUSNESS in successful inductions. For example, there are inductive links between conscious experience and memory: if I am having a phenomenally conscious experience of perceiving a stimulus *S* now, I am likely to retain an episodic memory of perceiving *S* later; but if I perceive *S* unconsciously, it is very unlikely that I will form an episodic memory of perceiving *S*. There are also inductive links between conscious experience and imagination. If I have had phenomenally conscious experiences in a perceptual modality *M* (e.g. colour vision), I am likely to be able to imagine having experiences in *M*; whereas if I have had never such experiences, I will struggle to imagine what they would be like. Perhaps *F* will play a much greater role than *N*, or vice versa, in explaining why these inductions work.

Yet I am doubtful that the relation of realization allows enough space between *F* and *N* for one to be substantially more relevant than the other to the explanation of our inductive successes. If we can explain the inductive links between conscious experience, memory and imagination in our own case by appealing to *F* and its integration with the cognitive architecture of memory and imagination, then we can also explain them by appealing to *N* and its integration with the human neural implementation of memory and imagination. The explanation can proceed equally well at either level, cognitive or neural, and the two explanations will complement each other. Whatever successful induction we choose, there will be a cognitive-level explanation for its success that appeals to *F* and its connections to other cognitive properties, and a neural-level explanation that appeals to *N* and its connections to the human neural implementations of other cognitive properties.

To find successful inductions for which *F* and *N* differ in their explanatory relevance, we would have to admit (as relevant for reference-fixing purposes) successful inductions concerning systems without *N*, such as inductions about how conscious experience relates to memory and imagination in insects or robots. But we cannot take it for granted in this context that any such inductions are actually successful—to regard them as successful would beg the question by assuming that PHENOMENAL CONSCIOUSNESS refers to *F* rather than to *N*.[8] In sum, appealing to inductive utility seems to help with "that sort of neighbourhood", where the candidates for reference are distinct in a way that gives them very different types of explanatory significance, but it seems not to help with cases where the two properties vying for reference are related by realization.[9]

[8] This is related to a point made by Michel (2019). Michel argues: to test the claim that pain is multiply realizable, we need to settle the question of whether it is present in any animals with different neural states that play the same functional role as pain. But to settle this question, we first need to know whether pain is multiply realizable. The threat here is one of epistemic circularity. But there is also a threat of semantic circularity for a semantic theory such as Shea's that ties reference to inductive success.

[9] Shea's response is more useful, I suggest, for defusing Carruthers' concern about indeterminacy between F_1 and F_2 (see footnote 5). We might well find that one of these cognitive properties is more relevant than the other to the cognitive-level explanation of our inductive successes. There will plausibly be a cognitive property that

Given the above, it is plausible that, if type-B materialism is true, then the reference of PHENOMENAL CONSCIOUSNESS is indeterminate between *F* and *N*. It is worth considering briefly why the same problem does not resurface for different solutions to the mind-body problem. It does not arise for forms of dualism (interactionist or epiphenomenalist) or Russellian monism (including forms of panpsychism and panprotopsychism) because these views accept our intuitive conceptions about PHENOMENAL CONSCIOUSNESS, including conceptions that concern its non-functional, intrinsic, distinctively first-personal nature. These views posit a special type of property that answers to those conceptions. The hard questions for these views are why we should believe that such properties exist, and how to reconcile their existence with a scientific worldview—not whether (if they do exist) the concept PHENOMENAL CONSCIOUSNESS succeeds in picking them out.

What about an analytic functionalist, such as Lewis (1983)? The analytic functionalist might deny that PHENOMENAL CONSCIOUSNESS refers to anything, favouring instead more specific concepts such as PAIN, for which the corresponding a priori functional role is more easily specified (they would then be a kind of strong illusionist). Alternatively, they might try to construct an a priori functional specification of PHENOMENAL CONSCIOUSNESS. Let us set aside for the moment the problem posed to this idea by the conceivability of zombies. If they were to take this route, they could further argue that PHENOMENAL CONSCIOUSNESS is ambiguous between two functionally defined concepts, one of which determinately refers to a role-property (what Lewis might have called the "attribute of having phenomenal consciousness"), and the other of which determinately refers to *N*, the neural state that realizes the phenomenal consciousness role in humans (cf. Lewis 1983, note 6). The corresponding problem for ethics would be one of choosing how to resolve this ambiguity in ethical contexts. That would be an interesting problem in its own right, but it is distinct from the problem that confronts the type-B materialist, and I will not discuss it further here.

3. Managing radical moral indeterminacy

By combining the considerations from the last two sections, and assuming type-B materialism is true, we can run the following argument:

> (**Premise 1**) For non-rational animals that possess valenced states (including many non-mammalian animals), the question of whether or not the animal has phenomenality-linked moral status (PLMS) hinges on whether at least some of the valenced states are consciously experienced.
> (**Premise 2**) The reference of PHENOMENAL CONSCIOUSNESS is indeterminate between a neuronal realizer property (*N*) that is not shared by non-mammalian animals and a high-level functional property (*F*) that is shared by many non-mammalian animals.

includes just enough architectural specificity to explain what needs explaining (e.g. the links between consciousness, memory and imagination) but no more specificity than is necessary.

(**Disturbing conclusion**): For many non-mammalian animals, it is indeterminate whether or not the animal has PLMS.

Can we accept the disturbing conclusion, or would doing so lead to a catastrophic meltdown of our rational deliberations regarding the treatment of non-human animals?

It is helpful here to distinguish between objective obligations and what have been called "subjective oughts" or "decision oughts" (Williams 2017). If an animal's moral status is indeterminate, our objective moral obligations will be indeterminate in relation to that animal. But it seems we can still ask: in practical deliberation, ought I give its interest the special weight owed to the interests of a conscious being, or not? I must choose between these options; the structure of deliberation forces a choice upon me. So ought I treat the animal *as if* it had PLMS or *as if* it did not have PLMS? The ought in question is a decision ought.

If there is no fact of the matter in relation to decision oughts, then we do indeed face a meltdown of practical deliberation. But objective indeterminacy may not have to spill over into indeterminacy at the level of decision oughts. There are various principles we could endorse that would prevent indeterminacy from derailing practical deliberation, allowing us to move from objectively indeterminate moral status to determinate decision oughts. I will call these principles "blocking principles".

One possible blocking principle draws inspiration directly from Williams (2017). Williams proposes that "a choice to *X* is decision permissible iff it is not determinately objectively impermissible to *X*" (2017, p. 670). If neither of two options is determinately objectively impermissible, then we may (subjectively, in our practical deliberations) treat both options as if they were permissible. In moving from indeterminacy to decision, we err on the side of permissiveness. A natural way to apply this idea to the present problem is the following:

> ***Blocking principle 1:*** If an animal has objectively indeterminate PLMS, then all choices with respect to that animal that are not objectively determinately impermissible are decision-permissible, as long as they are diachronically consistent with the same agent's other choices.

The motivation for Blocking principle 1 is the same as the motivation for Williams' principle. Faced with indeterminate obligations, rational decision must avoid *neutral sanction*: sanction from the point of view of someone who takes no stand on indeterminate matters. Following Blocking principle 1 allows an agent to avoid neutral sanction.

Williams includes a diachronic consistency constraint: you ought to avoid not just neutral sanction for doing something objectively impermissible, but also neutral sanction for being objectively inconsistent. Accordingly, once you have made a judgement call about an indeterminate fact, you ought to decide consistently with that judgement call in the future, as long as your views about the other issues at stake do not change. So, if you initially choose to

associate *F* with PLMS (e.g. when faced with the case of a bird), then all your subsequent decisions should be consistent with that (e.g. when faced with the case of a cognitively sophisticated invertebrate or robot).

Williams notes that such a principle is likely to generate "queasiness" (2017, p. 671), since you may find yourself rationally compelled (by your initial arbitrary judgement call) to make long sequences of decisions that an alter-ego who made a different judgement call would find subjectively impermissible. In the present case, "queasiness" seems too weak a word: the principle leads to profound unease. Blocking principle 1 might have been acceptable if the only indeterminacy we faced were the mild, sorites-style form, so that our initial arbitrary judgement call would only very occasionally constrain our future choices. But when the indeterminacy infects our dealings with a very wide range of non-mammalian beings (including birds, reptiles, fish, invertebrates, future robots and AI), a huge amount seems to hang on the initial judgement call. If we choose to take *N* as the ground of PLMS, we will give the interests of mammals greater moral weight in any choice scenario where the interests of mammals and non-mammals clash. If we choose *F*, we will not give greater moral weight to the interests of mammals—and indeed will regard this as a form of baseless taxonomic chauvinism.

The unease can be compounded by the following thought experiment: imagine an avian species one day evolves a human-level capacity for introspective and ethical thought. The avian creature constructs its own phenomenal concepts, which (assuming type-B materialism) refer to physical properties of its own brain. Suppose it constructs a concept of phenomenal consciousness* that refers to the sort of thing its phenomenal concepts pick out. This concept comes to carry great ethical significance for the avian creature. Unfortunately, its reference is indeterminate between *F* and *N**, the neuronal realizer of *F* in the avian brain. It endorses Blocking principle 1, permissibly chooses to take *N** as the ground of PLMS, and regards mammals, including humans, as lacking PLMS. The prospect of our avian counterpart reasoning in this way about us creates unease, and the source of the unease is Blocking principle 1.

A different approach begins with the way we would approach cases of *uncertain* moral status, assuming a sharp boundary between sentient and non-sentient life—and then aims to treat indeterminacy on the model of epistemic uncertainty. When consciousness is uncertain but determinately present or absent, there is a strong case for applying a precautionary principle and erring on the side of attributing PLMS in any case where we find both credible evidence of valenced states and widespread practices that cause extreme negative valence (Birch 2017). The same general thought, carried over to the case of indeterminacy, leads to the suggestion that, if an animal has objectively indeterminate PLMS, then it is decision-obligatory to treat it in all respects *as if it had PLMS*.

But this leads to the objection: is it plausible that, when you face a decision problem in which the interests of a determinately conscious animal come into conflict with the interests of an indeterminately conscious animal, determinacy has no moral significance at all? Suppose, for

example, you are a government or philanthropic foundation trying to decide whether to prioritise improving the welfare of mammals or birds. In the case of epistemic uncertainty, we can in principle handle trade-offs using probabilities, even though it is very challenging to estimate the probabilities. By contrast, in a case where we know that consciousness is indeterminate, it is very unclear what we should do, even in principle. It is certainly far from clear that we should attach no weight to determinacy.

Three main variants of "treating indeterminate cases as if they had PLMS" arise, representing different ways of handling trade-offs. First:

> ***Blocking principle 2:*** If an animal has objectively indeterminate PLMS, then it is decision-obligatory to treat it in all respects as if it had PLMS, drawing no distinction (*per se*) between those animals that determinately possess PLMS and those animals whose PLMS is indeterminate.

This version faces the criticism that it ignores the ethical relevance of indeterminacy. Faced with a choice between doing something determinately objectively obligatory and something indeterminately obligatory, it is plausible that our determinate obligations take priority (cf. Williams 2017, p. 655). For example, we plausibly have an obligation to intervene when a being with PLMS is tortured for no reason in front of us. Assume this, and suppose we are forced to choose between saving a determinately conscious being from torture and saving an indeterminately conscious being. On Blocking principle 2, we ought not take the determinacy into consideration as a morally relevant factor, and this seems wrong.

The intuitive pull of granting some ethical relevance to determinacy can be captured by either of the following:

> ***Blocking principle 3:*** If an animal has objectively indeterminate PLMS, then it is decision-obligatory to treat it in all respects as if it had PLMS, subject to the qualification that lexical priority should be given to the interests of animals that determinately possess PLMS.

> ***Blocking principle 4:*** If an animal has objectively indeterminate PLMS, then it is decision-obligatory to treat it as if it had PLMS, while giving greater weight (a "determinacy multiplier") but not lexical priority to the interests of animals that determinately possess PLMS.

However, both principles bring us back to the problem of genealogical unease raised by our example of the introspective avian creature. They involve giving either lexical priority or greater weight to beings that share our own neuronal mechanisms, simply because they happen to share our own neuronal mechanisms. The introspective avian would be entitled to do likewise, deprioritizing or giving reduced weight to the interests of mammals. That prospect should give us pause before devaluing non-mammals in our own ethical deliberations. Of course, Blocking principle 4 faces an additional problem: that of finding a

non-arbitrary determinacy multiplier.

These four blocking principles suggest that we are stuck between a whirlpool and a rock: the whirlpool is entirely denying the ethical significance of determinacy, and the rock is a kind of taxonomic chauvinism that should leave us profoundly uneasy, since it would allow a moral agent with a different neuronal realization of *F* to reason its way to chauvinism about us. To the extent that type-B materialism appears to leave us with a choice between abandoning a strongly plausible link between phenomenal consciousness and moral status, succumbing to a meltdown of practical deliberation regarding animals, or endorsing a blocking principle with profoundly troubling consequences, this is something we must add to the price of type-B materialism.

Acknowledgements
I thank Tim Bayne, Liam Kofi Bright, Campbell Brown, Susanne Burri, Giacomo Giannini, Anna Mahtani, Matthias Michel, David Papineau, Lewis Ross, Nick Shea and Robert Williams for their comments and advice. This research is part of a project that has received funding from the European Research Council (ERC) under the European Union's (EU) Horizon 2020 research and innovation programme, Grant No. 851145.

References

Adamo, Shelley A. (2016) Consciousness explained or consciousness redefined? *Proceedings of the National Academy of Sciences of the United States of America* 113:E3812.

Antony, M. V. (2006). Vagueness and the metaphysics of consciousness. *Philosophical Studies* 128:515-538. https://doi.org/10.1007/s11098-004-7488-8

Aru, J., M. Suzuki, R. Rutiku, M. E. Larkum, and T. Bachmann. (2019) Coupling the state and contents of con- sciousness. *Frontiers in Systems Neuroscience* 13:43. https://doi.org/10.3389/fnsys.2019.00043

Balog, Katalin (2012) In defense of the phenomenal concept strategy. *Philosophy and Phenomenological Research* 84:1-23.

Balog, Katalin (2020) Hard, harder, hardest. In Arthur Sullivan (ed.), *Sensations, Thoughts, and Language: Essays in Honor of Brian Loar*. New York, USA: Routledge. pp. 265-289.

Barron, Andrew B. and Klein, Colin (2016). What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences of the United States of America* 113:4900–4908.

Birch, Jonathan (2017) Animal sentience and the precautionary principle. *Animal Sentience* 2(16):1. https://animalstudiesrepository.org/animsent/vol2/iss16/1/

Birch, Jonathan (2020) The search for invertebrate consciousness. *Noûs* https://doi.org/10.1111/nous.12351

Block, Ned (2002) The harder problem of consciousness. *Journal of Philosophy* 99:391-425

Block, Ned (2007) Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences* 30:481-499.

Block, Ned (2011) Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences* 12:567-575.

Burge, Tyler (1979) Individualism and the mental. *Midwest Studies in Philosophy* 4:73-122.
Carruthers, Peter (2000) *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge: Cambridge University Press.
Carruthers, Peter (2018) Valence and value. *Philosophy and Phenomenological Research* 97:658-80.
Carruthers, Peter (2019) *Human and Animal Minds: The Consciousness Questions Laid to Rest*. Oxford: Oxford University Press.
Chalmers, David J. (2010) *The Character of Consciousness*. Oxford: Oxford University Press.
Chalmers, David J. (2018) The meta-problem of consciousness. *Journal of Consciousness Studies,* 25(9-10):6-61.
Crook, Robyn J. (2021) Behavioral and neurophysiological evidence suggests affective pain experience in octopus. *iScience* 24:102229. https://doi.org/10.1016/j.isci.2021.102229
Cutter, Brian (2017) The metaphysical implications of the moral significance of consciousness. *Philosophical Perspectives* 31:103-130. https://doi.oirg/10.1111/phpe.12092
Dawkins, Marian Stamp (2021) *The Science of Animal Welfare: Understanding What Animals Want*. Oxford: Oxford University Press.
DeGrazia, David (1996) *Taking Animals Seriously: Mental Life and Moral Status*. Cambridge: Cambridge University Press.
Dehaene, Stanislas (2014) *Consciousness and the Brain: Deciphering How the Brain Encodes Our Thoughts*. New York: Viking Press.
Dehaene, Stanislas and Changeux, Jean-Pierre (2011) Experimental and theoretical approaches to conscious processing. *Neuron* 70:200-227.
Dunaway, Billy (2016) Ethical vagueness and practical reasoning. *Philosophical Quarterly* 67:38-60. https://doi.org/10.1093/pq/pqw038
Frankish, Keith (2016) Illusionism as a theory of consciousness. *Journal of Consciousness Studies* 23:11-39.
Frith, Chris D. (2019) The neural basis of consciousness. *Psychological Medicine*.
Ginsburg, Simona and Jablonka, Eva (2007) The transition to experiencing: I. Limited learning and limited experiencing. *Biological Theory* 2:218-230.
Godfrey-Smith, Peter (2016) *Other Minds: The Octopus, the Sea, and the Deep Origins of Consciousness*. New York: Farrar, Strauss and Giroux.
Godfrey-Smith, Peter (2020) Metazoa: Animal Minds and the Birth of Consciousness. New York: Farrar, Strauss and Giroux.
Kahane, Guy and Savulescu, Julian (2009) Brain damage and the moral significance of consciousness. *Journal of Medicine and Philosophy* 34:6-26.
Kammerer, Francois (2020) The normative challenge for illusionist views of consciousness. *Ergo* 6: 891-924.
Koch, Christof, Massimini, Marcello, Boly, Melanie, and Tononi, Giulio (2016) Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience* 17: 307–322.
Korsgaard, Christine M. (2018) *Fellow Creatures: Our Obligations to the Other Animals*. Oxford: Oxford University Press.

Kriegel, Uriah (2019) The value of consciousness. *Analysis Reviews* 79:503-520.
LeDoux, Joseph (2019) *The Deep History of Ourselves: The Four-Billion-Year Story of How We Got Conscious Brains*. New York: Viking Press.
Lee, Andrew Y. (2019) Is consciousness intrinsically valuable? *Philosophical Studies* 176:655-671.
Lee, Geoffrey. (2019) Alien subjectivity and the importance of consciousness. In Pautz, Adam and Stoljar, Daniel (Eds.), Blockheads! Essays on Ned Block's Philosophy of Mind and Consciousness. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9196.003.0014
Lewis, David K. (1983) An argument for the identity theory. In his *Philosophical Papers, Volume 1* (pp. 99-107). New York: Oxford University Press.
Loar, Brian (1990) Phenomenal states. *Philosophical Perspectives* 4:81-108.
Mackie, John L. (1974) *The Cement of the Universe: A Study of Causation*. Oxford: Clarendon Press.
Magee, Barry and Elwood, Robert W. (2016) Trade-offs between predator avoidance and electric shock avoidance in hermit crabs demonstrate a non-reflexive response to noxious stimuli consistent with prediction of pain. *Behavioural Processes* 130:31-35.
Mashour, G. A., Roelfsema, P., Changeux, J. P., and Dehaene, S. (2020) Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105 (5):776–98. https://doi.org/10.1016/j.neuron.2020.01.026
Mather, Jennifer (2019) What is in an octopus's mind? *Animal Sentience* 4(26):1. https://animalstudiesrepository.org/animsent/vol4/iss26/1/
Michel, Matthias (2019) Fish and microchips: On fish pain and multiple realization. *Philosophical Studies* 176:2411-2428. https://doi.org/10.1007/s11098-018-1133-4
Millikan, Ruth G. (2000) *On Clear and Confused Ideas. Cambridge: Cambridge University Press*.
Papineau, David (1993) *Philosophical Naturalism*. Oxford: Blackwell.
Papineau, David (2002) *Thinking about Consciousness*. Oxford: Oxford University Press.
Papineau, David (2003) Could there be a science of consciousness? *Philosophical Issues* 13:205-220.
Papineau, David (2020) The problem of consciousness. In Kriegel, Uriah (Ed.), *The Oxford Handbook of the Philosophy of Consciousness* (pp. 14-38). New York: Oxford University Press.
Paul, Elizabeth, Sher, Shlomi, Tamietto, Marco, Winkielman, Piotr, and Mendl, Michael T. (2020) Towards a comparative science of emotion: Affect and consciousness in humans and animals. *Neuroscience and Biobehavioral Reviews* 18:749-770.
Pessiglione, M., Schmidt, L., Draganski, B., Kalisch, R., Lau, H., Dolan, R. J., and Frith, C. D. (2007) How the brain translates money into force: A neuroimaging study of subliminal motivation. *Science* 316:904-906.
Pessiglione, M., Petrovic, P. Daunizeau, J., Palminteri, S., Dolan, R. J. and Frith, C. D. (2008) Subliminal instrumental conditioning demonstrated in the human brain. *Neuron* 59:561-567.
Regan, Tom (2004). *The Case for Animal Rights*. 2nd edition. Berkeley: University of California Press.

Rosenthal, David M. (2005) *Consciousness and Mind*. Oxford: Clarendon Press.
Shea, Nicholas (2012) Methodological encounters with the phenomenal kind. *Philosophy and Phenomenological Research* 84:307-344.
Shepherd, Joshua (2018) *Consciousness and Moral Status*. London: Routledge.
Shevlin, Henry (2020a). Which animals matter? Comparing psychological approaches to psychological moral status in non-human systems. *Philosophical Topics*.
Shevlin, Henry (2020b). Non-human consciousness and the specificity problem. *Mind and Language*.
Shoemaker, Sydney (2007) *Property Realization*. Oxford: Oxford University Press.
Simon, Jonathan A. (2017). Vagueness and zombies: Why 'phenomenally conscious' has no borderline cases. *Philosophical Studies*, *174*, 2105–2123. https://doi.org/10.1007/s11098-016-0790-4
Singer, Peter (1995) *Animal Liberation*. 2nd revised edition. London: Pimlico.
Singer, Peter (2011) *Practical Ethics*. 3rd edition. Cambridge: Cambridge University Press.
Skora, L. I., Yeomans, M. R., Crombag, H. S. and Scott, R. B. (2021) Evidence that instrumental conditioning requires conscious awareness in humans. Cognition 208: 104546.
Sneddon, Lynne U., Elwood, Robert W., Adamo, Shelley A., & Leach, Matthew C. (2014). Defining and assessing animal pain. *Animal Behaviour* 97:201-212.
Stoljar, Daniel (2005) Physicalism and phenomenal concepts. *Mind and Language* 20:469-494.
Taylor, John Henry (2013) Is consciousness science fundamentally flawed? *Journal of Consciousness Studies* 20:203-21.
Tomasik, Brian (2014) Do artificial reinforcement-learning agents matter morally? *arXiv*, https://arxiv.org/abs/1410.8233, accessed 27 August 2020.
Wiblin, Rob, Koehler, Arden and Harris, Keiran (2019) David Chalmers on the nature and ethics of consciousness. The 80,000 Hours Podcast, 16 December 2019. Retrieved from: https://80000hours.org/podcast/episodes/david-chalmers-nature-ethics-consciousness/
Williams, J. Robert G. (2017) Indeterminate oughts. *Ethics* 127:645-673.