For ‘Understanding non-human consciousness: methods, prospects, and challenges’, in Henry Shevlin (ed.) Special Issue, *Philosophical Topics*.

Consciousness, Concepts and Natural Kinds

Tim Bayne and Nicholas Shea

Abstract

We have various everyday measures for identifying the presence of consciousness, such as the capacity for verbal report and the intentional control of behaviour. However, there are many contexts in which these measures are difficult (if not impossible) to apply, and even when they can be applied one might have doubts as to their validity in determining the presence/absence of consciousness. Everyday measures for identifying consciousness are particularly problematic when it comes to ‘challenging cases’—human infants, people with brain damage, non-human animals, and AI systems. There is a pressing need to identify measures of consciousness that can be applied to challenging cases. This paper explores one of the most promising strategies for identifying and validating such measures—the natural kind strategy. The paper is in two broad parts. Part I introduces the natural kind strategy, and contrasts it with other influential approaches in the field. Part II considers a number of objections to the approach, arguing that none succeeds.

Part I

(1) The Validation Challenge

Although few theorists would take issue with the claim that adult neurotypical human beings have a standing capacity for consciousness, there is little else that can be said about the distribution of consciousness that is not highly controversial. Some theorists hold that a capacity for consciousness requires language and/or a theory of mind, and does not extend much beyond the so-called higher mammals. Others would include reptiles, birds, fish, and perhaps even insects within the circle of sentience. Another group of theorists argue that consciousness is ubiquitous, and that even the basic building blocks of the physical world might harbour a kind of conscious awareness. As AI becomes increasingly sophisticated, debates about the distribution of consciousness in the biological realm will be joined by debates about the distribution of consciousness in the world of computing machines. Some argue that we are on the verge of creating conscious machines—or even that current AI systems might be conscious in some way. Others are sceptical about the prospects of machine consciousness, arguing that conscious AI is at best a remote possibility. Indeed, uncertainty about the distribution of consciousness also encompasses the members of our own species. There is debate about: when the capacity for consciousness first appears in infancy; the conditions under which it is retained in the context of severe brain damage; and when—or even whether—it is lost in sleep and anaesthesia.

There is, then, a clear need to develop measures ('markers', 'indicators', 'indexes', 'signatures') of consciousness that might be applied not just to neurotypical adult human beings in the awake state but to a much broader range of individuals, such as human infants, humans who have suffered from severe brain damage, nonhuman animals, and artificially intelligent agents. Ideally, we would have measures that would enable us to tell not only *whether* an entity is conscious but also *how* it is conscious—that is, what global state of consciousness it is in (e.g., ordinary waking awareness or dreaming awareness) and what its conscious contents are (i.e.., what perceptual, affective and cognitive experiences it has). We focus here on the quest to identify measures of consciousness as such (that is,

'generic consciousness'), but much of what we say applies also to the identification of global states of consciousness and the contents of consciousness.

A number of measures of consciousness have been proposed in the literature. For example, it has been suggested that the capacity to detect an unusual conjunction of perceptual features (Faivre et al. 2014; Mudrik et al. 2014) or to detect violations of second-order regularities (Faugeras et al. 2012; King et al. 2013) can be used as measures of consciousness. Other proposed measures of consciousness are neural—such as the 'ignition' that appears to follow the presentation of a perceptual stimulus (Dehaene & Naccache 2001) and the perturbational complexity index ('PCI'), a measure of the neural complexity that results from a TMS pulse (Casali et al. 2013, Casarotto et al. 2016). These and other putative tests of consciousness are already being used as tentative measures of consciousness in challenging cases, having been invoked in debates about the distribution of consciousness in infants (Goksan et al. 2015; Kouider et al. 2013), sedated individuals (Huang et al. 2018), and severely brain-damaged patients (Gosseries et al. 2014; Sitt et al. 2014).

The central challenge, of course, concerns the *validation* of any putative measure of consciousness ('sentience', 'awareness'—we use these terms interchangeably). It is one thing for a neural or cognitive property to be practicable as a measure of consciousness in one or more challenging populations, but it is another thing to know (or be warranted in thinking) that the property in question can play that role. Are we able to justify treating a particular neural or cognitive response as an indicator of the presence (or absence) of consciousness in 'challenging cases'—that is, in cases in which our pretheoretical measures of consciousness fail to deliver a robust verdict? Call this the *validation challenge*.

In the contemporary landscape, positive responses to the validation challenge tend to take one of three forms. The first appeals to what we will call the *replacement strategy*. (It could also be called the *deflationary strategy*.) Here, one meets the challenge of validating measures of consciousness by replacing CONSCIOUSNESS[1] with another concept (or concepts) that is/are more amenable to

[1] We use small caps to refer to concepts.

measurement. A recent example of this approach is provided by Dehaene, Lau & Kouider (2017), who claim that "consciousness" is a pre-theoretical term that "conflates" three quite distinct phenomena: vigilance/wakefulness; global availability to the organism (what they call 'C1'); and metacognition (what they call 'C2'). Setting vigilance/wakefulness to one side, Dehaene and his colleagues go on to equate the question of whether a target entity is conscious with the question of whether it exemplifies global availability or metacognition, a question which they take to be relatively straightforward to address.

Is the replacement strategy viable? It is certainly a familiar lesson from the history of science that terms can turn out to be equivocal and in need of 'splitting' (Carey 1991; Kitcher 1988; Thagard 1992). Prior to Galileo average velocity was not clearly distinguished from instantaneous velocity (Kuhn 1964), and prior to Joseph Black heat and temperature were not clearly distinguished (Wiser and Carey 1983). The suggestion that 'consciousness' too might stand in need of splitting is an important one and should be taken seriously.

That being said, we have seen no good reason to adopt the replacement strategy at this point. The primary problem is that we have been given no good argument that the successor properties (e.g., *global availability* and *metacognition*) capture components of the concept we started with—namely CONSCIOUSNESS. We need good reason to think that by measuring global availability and metacognition we aren't simply changing the subject, but are instead accounting for the phenomenon or phenomena that we have in mind when asking what the distribution of consciousness is. It is not enough that C1 and C2 are more amendable to scientific study. Dehaene and his colleagues seem to be relying on the idea C1 and C2 correlate to some extent with pre-theoretic measures of consciousness. We object to the correlation strategy in a moment, but in any event the replacement strategy requires more. As in the case of velocity or heat/temperature, we need an argument that the original concept conflates two properties that together account for the target phenomena without residue. Whatever argument is inserted here, the replacement strategy will be backed by some approach to validating the

replacement properties. We are, then, thrown back on approaches of the kind that we consider below.[2]

A second response to the validation challenge is the *correlational strategy*.[3] The idea here is that it is reasonable to treat a putative measure as valid if it correlates strongly with our pre-theoretical markers of consciousness. In this approach, the pre-theoretical markers of consciousness function as a gold standard against which candidate markers are to be assessed. The correlational strategy instructs us to begin with our ordinary, pre-theoretic markers of consciousness (such as verbal report, storage in episodic memory and the control of voluntary behaviour), and looks for correlations between these pre-theoretical markers and the proposed measure of consciousness.

The correlational strategy is also problematic. For one thing, there are many contexts in which the pre-theoretical measures of consciousness cannot be applied (at least, not in any straightforward manner). For example, neither neonates nor vegetative state patients are capable of report, and it is unclear what capacities they have for episodic memory or the intentional control of behaviour. One could of course take these considerations as decisive evidence that neonates and vegetative state patients cannot be conscious, but that inference would be premature given the uncertainty surrounding the relationship between consciousness and cognition. A second problem with the correlational approach is that we want our 'scientific' measures of consciousness to be able to correct (and not merely supplement) our pre-theoretical makers of consciousness, for our

[2] A further objection to the replacement strategy is that we currently lack a sufficiently precise formulation of global availability to make it clear what it would be to measure global availability in an organism; similarly, we lack a sufficiently precise understanding of what form of metacognition we should be measuring.

[3] This corresponds to the theory-neutral approach, in Birch's useful taxonomy (Birch, unpublished). Our theory-based approach below corresponds to Birch's theory-heavy approach, and our preferred approach, the natural kind strategy, would count as theory-light in Birch's taxonomy.

pre-theoretical markers of consciousness are surely not beyond debate. However, the structure of the correlational approach is such that it could never provide us with reasons for taking a novel measure to trump our pre-theoretical measures.

A third response to the validation challenge is the theory-based strategy. Here, one justifies a putative measure of consciousness on the grounds that it is supported by a certain theory of consciousness. For example, the P3b ERP response is taken to be a measure of consciousness on the grounds that it fits with the global workspace theory of consciousness (Dehaene et al. 2011). The theory-based strategy is also hinted at in some presentations of the perturbational complexity index (PCI), one of the most influential proposed measures of consciousness to date. As well as justifying the PCI on the basis of a correlation between PCI-based measures and pre-theoretical assumptions about the distribution of consciousness, theorists have also motivated PCI on the grounds that it "follows" from Giulio Tononi's *Integrated Information Theory* (IIT) theory of consciousness (Casali et al. 2013; Casarotto et al. 2016).

There are certain respects in which the theory-based strategy is superior to the correlational strategy; for one thing, the theory-based strategy would allow us to treat a novel measure as trumping our pre-theoretical markers. However, the theory-based approach has troubles of its own. One problem is that no extant theory of consciousness enjoys widespread support, and any putative measure of consciousness that is justified on the basis of theoretical fit will be as controversial as the theory on which it is based. Further, the strength of evidence for a theory of consciousness will depend on a range of measures of consciousness used to test to the theory. The validation of these measures cannot then be wholly derivative from the theory.

In short, none of the dominant strategies for validating proposed measures of consciousness is satisfactory. In response to this sorry state of affairs, one might be tempted to embrace doubts about whether the validation challenge *can* be met. Chalmers has defended a view along these lines, arguing that "the primary criterion for consciousness will always remain the functional property that we started with: global availability, or verbal report, or whatever." (Chalmers 1996: 243). Chalmers's arguments have rightly received considerable attention. Here,

we will focus instead on motivating a positive account of how novel measures of consciousness can be validated: the natural kind strategy. If the strategy we are suggesting here succeeds it will do much to answer Chalmers's scepticism.

(2) The Natural Kind Strategy

Consider the scientific study of hepatitis (Seeff 2009). At one time hepatitis was diagnosed only on the basis of set of 'observational' signs and symptoms, jaundice (turning yellow), fever, the patient's history, and clinical examination of the abdomen. The scientific investigation of hepatitis examined the patterns of association and dissociation between these various marks, looking for ways in which they clustered together. An early distinction was found between one variety of hepatitis that spread in outbreaks suggestive of a faecal-oral route and a second variety with no obvious chain of transmission and a longer incubation period. The latter was later discovered to be related to blood transfusions. Having identified such patterns, investigators looked for the underlying mechanisms that would explain them, discovering that they were explained by viruses of various types (initially hepatitis A and B, later C). The tests that we now have target those viruses. Although we still use many of the pre-theoretical symptoms of hepatitis as rough-and-ready markers of the different conditions, tests for the presence of the relevant virus are used to both supplement and (where necessary) correct those pre-theoretical tests.

The strategy used to identify and validate novel measures of hepatitis is what we call the *natural kind* (NK) strategy, for it treats the target phenomenon as a natural kind (Shea 2012; Shea & Bayne 2010; Block 2007 advances the related idea of relying on inference to the best explanation). Applied to consciousness, the idea is that we should begin with a variety of dispositions and capacities ('marks') that are putatively associated with consciousness. The marks will include the various pre-theoretic indicators of consciousness, such as verbal report and volitional behavioural control, but they are by no means limited to such features (helpfully, the approach does not rely on being able to distinguish between pre- and post-theoretic marks). Other marks might concern the structure of consciousness, such as the spatial and temporal dimensions of perceptual experience (Phillips 2014a; Phillips 2014b), or the distinctive forms of unity that seem to characterize

consciousness (Bayne 2010; Dainton 2000). Still others will concern the functional profile of consciousness, such as the (apparent) fact that learning a novel predictive relationship between two stimuli requires (or is greatly facilitated by) awareness of both stimuli (Faivre et al. 2014; Travers et al. 2018; cf. Scott et al. 2018).

The NK strategy instructs us to determine the degree to which the marks of consciousness pattern together (cp. Godfrey-Smith 2020, p. 2). Do they tend to come and go in groups, across changes of conscious content and global state? Can some subsets be used to predict the presence or absence of others? If so, we have good reason to think consciousness (and/or its determinables) involves a natural kind—an underlying property that is responsible for the clustering of the various features that have been identified.

The next step is to investigate the reasons behind this clustering. It might be a matter of a shared computational property, something about information being processed in a characteristic way. That would be substrate-neutral. Alternatively, the clustering could be due to the way in which neurons and neural assemblies process information, such as the dynamics and resonance properties that follow from their electrochemical profile. A third possibility is that the clustering is due to features of neural structure and function, such as patterns that result from developmental pathways (cp. the patterning of deficits caused by strokes due to the anatomy of blood supply in the brain) or features due to common selective pressures with somewhat different aetiologies (cp. evolution of the camera eye).

The NK strategy does not presuppose that the reason the marks of consciousness cluster together (to the extent that they do) must be explained by appeal to properties like the 'eternal kinds' of physics and chemistry. A natural kind can be a 'homeostatic property cluster' (Boyd 1991), where a natural process has the effect that some features tend to cause the presence of others, without there being a single unifying property that accounts for the presence of all the features in the cluster. The clustering of features can also arise due to shared causal history ('historical kinds', Millikan 1998). For example, individuals of the same species share features, not because there is some synchronic essential property they share, but because they all descend, through a conservative copying process, from

a common ancestor with those features (Griffiths 1999). Here we adopt a broad understanding of what it takes to be a natural kind: it is any collection of natural properties that cluster together better than chance for a univocal reason. 'Natural kind' is sometimes used for the cluster of features and sometimes for the property that generates the cluster of features (e.g. being gold).[4] When we need to disambiguate, we use 'natural kind property' for the latter.

The remainder of this paper explores a number of objections to the NK strategy—objections which in effect argue that it is illegitimate to treat consciousness as a natural kind. But before considering those objections, we need to clarify what exactly the NK strategy does—and does not—commit one to.

First, it is important to set to one side a metaphysical issue which is orthogonal to our discussion. The NK strategy can seem to be inextricably tied to physicalism (or materialism), but it is in fact also compatible with some species of dualism. Physicalist versions of the NK strategy treat consciousness and its determinates as physical properties. By contrast, non-physicalist versions of the strategy hold that consciousness and its determinates are non-physical properties that march in step with physical properties as a matter of at least nomological necessity. Our focus here is on the NK strategy itself, and we take no stand on the metaphysics of consciousness. For simplicity we will refer to consciousness *being* a physical natural kind, but that is merely a *façon de parler*, and the position that we will develop is compatible with any view on which consciousness is tightly connected with (e.g. is necessarily co-instantiated with) a physical kind.

Second, the NK strategy is viable only if it remains an open empirical possibility that there is a natural kind in this area (i.e. closely connected to CONSCIOUSNESS or "consciousness"). A number of theorists have argued that there is no natural kind corresponding to "consciousness" (e.g., Allport 1988; Churchland 1988; Irvine 2013). If those claims are right, then the NK strategy is a non-starter. But in our

[4] It is also sometimes used for the collection of individuals that are members of the kind, e.g. 'the members of the species form a natural kind'.

view whether there is a natural kind in the vicinity of consciousness remains an open question.

Third, the NK strategy does not presuppose that there will always be a determinate answer to the question of whether an entity is conscious, for natural kinds can admit of borderline cases: things which are neither determinately an instance of the kind nor determinately not an instance of the kind. This point applies not only to natural kinds as understood on the homeostatic property cluster view, but extends also to natural kinds as historical kinds. The fact that an entity might lie at the border of a kind is not a matter of how many of the 'marks' associated with the kind are present, for a determinately 100% instance of the kind could have most or all of its marks masked. Instead, it's a matter of whether the operative principle that is responsible for the marks is at work. That could be a matter of degree (e.g. if the operative principle is a process). The upshot of this is that even if consciousness is a natural kind, it doesn't follow that there will always be a determinate answer as to whether a particular entity is conscious or not. Taking that possibility seriously is a mildly counter-intuitive consequence of adopting the NK strategy, but it is not unusual to find, especially in the realm of biology, that a property that we take to be all-or-nothing in fact comes in degrees or admits of borderline cases (cp. life).

Part II

(3) The Phenomenal Concept Objection

The NK strategy has been outstandingly successful in other areas of science—indeed, it is arguably the standard method for validating novel measures of a natural phenomenon (e.g., hepatitis, chemical elements, temperature). Thus, it is puzzling why the strategy has received so little attention in the consciousness literature. This neglect might be understandable if the approach had been shown to be fatally flawed, but that is not the case. We know of very few discussions of consciousness that even consider the NK strategy, let alone claim to have shown it to be untenable. Scientific neglect of the NK strategy might stem (at least in part) from ignorance of the approach, but we suspect that philosophical neglect derives from doubts about its legitimacy. The remainder of this paper examines four challenges to the legitimacy of the NK strategy. We argue that none succeeds,

although they do throw light on why the NK strategy has been overlooked in studying consciousness.

We begin with the *phenomenal concept objection*. Whatever other ways we have of conceptualising consciousness—in terms of a semi-technical notion of global access, say, or through a collection of folk-psychological platitudes about functional roles—it is widely assumed that we each possess the capacity to establish an intimate cognitive relation to our own conscious states. This relation is mediated by what have come to be known as *phenomenal concepts* (e.g. Loar 1990/1997; Tye 2000; Perry 2001; Papineau 2002). Although theorists disagree about precisely how phenomenal concepts should be understood (see Balog 2009 and Sundström 2011 for reviews), at the heart of any commitment to phenomenal concepts is the idea that they can be directly and non-inferentially applied as result of being in the type of experiential state to which they refer. (If they are genuine concepts—that is, components of thought satisfying some kind of generality constraint—then they must also be applicable indirectly and/or inferentially, as when we apply them to other people or when we deploy them to judge that we are not in certain kinds of conscious states.) Most theorists also hold that a phenomenal concept is acquired by instantiating the experiential property to which it refers, or at least one very similar to it.[5]

Although the existence of phenomenal concepts is controversial (Macdonald 2004; Crane 2005; Ball 2009; Tye 2009), on our reading of the literature most contemporary philosophers of mind not only recognize the existence of phenomenal concepts, they also view them as central to the way that consciousness is ordinarily conceptualised. Thus, it is a matter of some importance whether a commitment to phenomenal concepts (PCs) is at odds with the NK approach.

[5] When it comes to determinates of consciousness (such as the experience of a particular shade of colour), one may be able to acquire the corresponding phenomenal concept by instantiating near neighbours of the state (cp. the missing shade of blue). We will suppress this qualification in our subsequent discussion.

How might a PC-based argument against the NK strategy go? One version of the argument runs as follows:

(1) CONSCIOUSNESS is a phenomenal concept.

(2) If CONSCIOUSNESS is a phenomenal concept, then it is not a natural kind concept.

(3) CONSCIOUSNESS is not a natural kind concept. (From (1) and (2))

(4) The NK strategy could be successful only if CONSCIOUSNESS were a natural kind concept.

Therefore,

(5) The NK strategy cannot be successful. (From (3) and (4).)

Central to this argument is the notion of a 'natural kind concept'. How should this notion be understood?

One conception of a natural kind concept identifies it with a concept that refers to a natural kind property. On this understanding, CONSCIOUSNESS would qualify as a natural kind concept if and only if consciousness itself is a natural kind. But this reading of what it is to be a natural kind concept cannot be the intended one here, for the version of the objection that it generates would amount only to the charge that consciousness isn't a natural kind. Consciousness may turn out not to be a natural kind property (as we have already noted), but the PC-based argument promised to deliver a different kind of objection to the NK strategy. If the PC-based argument is to make good on that promise it must employ a different conception of what it is to be a natural kind concept.

The following comments from Peter Carruthers point us in the direction of such an account:

> [One] thing that everyone has agreed on, at least since Kripke (1980), is that terms referring to conscious mental states aren't used as natural-kind terms. In contrast, it is generally agreed that our concepts for substances like water are natural-kind ones. Even before we knew anything about chemistry, we

> used the concept water to refer to the underlying nature or essence of the recognizable stuff that fills our lakes and rivers (H_2O); and it turned out that it was that very same stuff that presents as ice in some circumstances (frozen water) and as mist in others (evaporated water). But our concepts of the felt qualities of our conscious experiences aren't like that. We don't use them with the intention of referring to whatever natural kind underlies those experiences, whatever that might turn out to be, and however that kind might be presented in other creatures. On the contrary, we mean to refer just to those felt qualities themselves. (Carruthers 2018: 53)

The key idea here is the claim that CONSCIOUSNESS isn't a natural kind concept because it isn't used with the intention to refer to the natural kind that underlies experience, but is (instead) used to refer to "just those felt qualities themselves." The supposed contrast, of course, is with concepts such as WATER, which are supposedly used with the intention to refer to a natural kind property. On this reading, a natural kind concept is a concept where the user means or intends in some way to refer to a natural kind. We can broaden that somewhat to include not just intentions, but whatever it is in the thinker's conceptions and use of the concept that operates to fix its reference. So we will take a natural kind concept to be any concept whose semantics is somehow constrained such that it must refer to a natural kind if it refers at all.

Let us return now to the PC-argument, understanding what it is to be a natural kind concept in these terms.

There are three ways in which one might respond to it. First, one might reject premise (1), arguing that while the concepts associated with fine-grained conscious states (such as experiences of red, pain or sounds) are phenomenal concepts, CONSCIOUSNESS itself isn't a phenomenal concept. We, however, are not inclined to endorse this line of reply. We are certainly open to the idea that certain concepts associated with consciousness are not phenomenal concepts. For example, the concepts associated with global states (or 'levels') of consciousness—such as alert wakefulness, dreaming, and the psychedelic state—may not qualify as phenomenal concepts (Bayne et al. 2016). But in our view CONSCIOUSNESS itself does qualify as a phenomenal concept. More carefully, even if

there are concepts of consciousness that aren't phenomenal concepts (such as ACCESS CONSCIOUSNESS), the property that lies at the heart of the quest for measures of consciousness is phenomenal consciousness, and the concept corresponding to phenomenal consciousness is a phenomenal concept.

A second response to the PC argument targets premise (2). On the face of it, that premise appears to be secure. After all, phenomenal concepts are contrasted with natural kind concepts. (Consider again the quotation from Carruthers reproduced above.) Phenomenal concepts are self-applied directly in virtue of being in a state of the type referred to, they have distinctive acquisition conditions, and they are inferentially isolated from functional and physical concepts, whereas natural kind concept appear to have none of those features. How then—one might wonder—could CONSCIOUSNESS be both a phenomenal concept *and* a natural kind concept?

Insofar as CONSCIOUSNESS has the features just listed, it isn't a *canonical* natural kind concept. However, we would argue that CONSCIOUSNESS could nonetheless qualify as both a phenomenal concept and a natural kind concept, for one could think of the property that one is able to directly self-ascribe as a natural kind property (much like any other biological property). In the passage quoted above Carruthers contrasts intending to refer to a natural kind that underlies experience with intending to refer to "felt experiential qualities", but we would argue that this contrast is a false one, for one could intend to refer to *both* of these features. Indeed, if one regards consciousness as a natural kind then one *will* intend to refer to both felt experiential qualities and a natural kind. In short, being a phenomenal concept is inconsistent with being a *canonical* natural kind concept, but there is no reason to think that it is inconsistent with being a natural kind concept in an extended sense of the term.

A third response to the PC argument rejects premise (4), arguing that the success of the NK strategy does not require CONSCIOUSNESS to be a natural kind concept, but requires only that the property to which CONSCIOUSNESS refers is in fact a natural kind. This response will be attractive to those who doubt that referential intentions play a central role in fixing the content of a concept. Concept users may not intend to refer to a natural kind—they may even positively intend to refer to the felt quality of experience (although we doubt the need for such referential

intentions, and their existence)—but that does not rule out that the property to which CONSCIOUSNESS refers is in fact a natural kind property.

Is there another way in which phenomenal concepts could be disqualified from referring to a natural kind? Their distinguishing feature is that the user can correctly apply the concept to instances of the property, ones which she herself instantiates, in virtue of having the experience. Maybe the argument is simply that the property to which concept-users intend to apply CONSCIOUSNESS in their own case is the property of having a phenomenally conscious experience ('the felt quality' of experience), and that *that* is not a natural kind property. But then we are back to the claim that they intend to refer to something which is not in fact a natural kind. As we have seen, that would be to beg the question against the NK strategy. In short, we don't see how (1) to (5) could form the basis for a good argument against the NK strategy.

(4) The Objection from Revelation

Thus far we have focused on the thought that the NK strategy is doomed to failure because CONSCIOUSNESS is a phenomenal concept rather than a NK concept. We turn now to a slightly different version of 'the' phenomenal concept objection—a version that makes heavy use of the notion of revelation. The basic idea is that the phenomenal concept user knows something about the nature of consciousness from their own case—and that what is known is at odds with the NK strategy. Whereas the PC objection is focused on the concept of consciousness, the focus of the revelation objection is on the nature of consciousness itself.

An oft-expressed thought in the consciousness literature is that self-application of a phenomenal concept while being in the conscious state it picks out reveals something substantive about the nature of its referent. Different authors express this idea in slightly different ways. According to Levine, a successful account of phenomenal concepts should explain 'the especially immediate and intimate cognitive relation between phenomenal concepts and their objects' – as afforded by phenomenal concepts (2006: 162). Chalmers describes a phenomenal concept as a concept that picks out its referent 'in terms of its intrinsic nature' (2003: 225). Nida-Rümelin says that our grasp of phenomenal properties goes via a grasp of phenomenal concepts, and that 'to grasp a property is to understand what having

that property essentially consists in' (2007: 307). Horgan and Tienson claim that when one thinks of a phenomenal property via a phenomenal concept one thinks of it 'directly, as it is in itself' (2001: 311). Finally, Goff writes: 'phenomenal concepts reveal the essence of the states they denote' (2017: 107).

One line of resistance would be to deny that phenomenal concepts reveal anything substantive about the nature of their referents. However, for the sake of engaging with the argument we will grant that there may be something to the idea that phenomenal concepts reveal something about their referents in a way in which (say) canonical natural kind concepts such as WATER do not. But the crucial issue is whether what is revealed in the first-person application of phenomenal concepts is at odds with consciousness being a natural kind property. We think not.[6]

To see why, consider Goff's explication of the idea that phenomenal concepts reveal the nature of their referents:

> Surely, you know exactly what your pain is—what it is for someone to feel pained in precisely that way—just by attending to pain and thinking about it in terms of how it feels. There is nothing in any way hidden from you about the reality of how you're feeling; nor is it possible that you're not really feeling that way. And that's because the feeling is "right there" for you, in such a way that its reality cannot be doubted. (2017: 108)

There are two readings of the idea that 'there is nothing in any way hidden from you about the reality of how you're feeling'. One reading involves what we call lightweight revelation; the other involves heavyweight revelation. We grant the former but reject the latter.

Lightweight revelation focuses on the experience itself. If you're in pain then there is something it is like for you to have that experience. We can talk about 'knowing'

[6] Revelation is also advanced, alongside conceivability arguments and the explanatory gap, as grounding an argument against physicalism (Goff 2011; Goff 2015; Horgon & Tienson 2001; Trogdon 2017). It is not our purpose here to defend physicalism in general, so we will not engage with that argument here.

what it's like in a way which depends only a conscious subject having an experience, and does not depend on the subject having or using any concepts of experience. In this sense, the experience is indeed 'fully revealed' to its subject. Nothing about what it's like to have the experience is hidden from the subject. If the subject is having that experience, then she really is feeling that way – she 'knows what it's like'. But there is nothing more to this 'knowing' than the having of the relevant experience, and thus lightweight revelation is unexceptional. Confusing having an experience with thinking about a current experience could make these platitudes seem more substantial than they are.

Heavyweight revelation, by contrast, is neither unexceptional nor platitudinous, for it involves the deployment of concepts, and the idea that self-application of a phenomenal concept reveals substantial truths about its referent. Even if consciousness is a natural kind, it is implausible that the content *consciousness is a natural kind* would be revealed as a result of self-applying the phenomenal concept. Nor is it plausible that *consciousness is not a natural kind* would be revealed by the application of a phenomenal concept. Instead, the claim must be that self-application of a phenomenal concept reveals substantial truths about its referent that are at odds with the possibility that consciousness is a natural kind property. But what substantive truths about consciousness could be revealed by a grasp of phenomenal concepts?

The literature contains few answers to this question. Indeed, to the best of our knowledge the only example of a substantive truth about consciousness that revelationists have offered is due to Goff, who provides, 'If x is pain then x is bad'. Accepting (if only for the sake of argument) that this is indeed a substantive truth about the nature of pain that is mediated by self-application of PAIN, it's not at all clear that similar examples can be provided for knowledge of other kinds of phenomenal states, or indeed for knowledge of consciousness itself. Furthermore, the truth of 'if x is pain then x is bad' is not at odds with the possibility that pain (or the determinable, phenomenal consciousness) is a natural kind. As an example of what is positively revealed, it does not suggest that what is revealed is incompatible with consciousness being a natural kind, and hence with the presuppositions of the natural kind strategy.

Finally, it is worth remembering that the NK strategy as such is not tied to one side of the metaphysical debate between physicalism and dualism. As we noted above, the strategy is equally valid if dualism is true, provided consciousness is associated with a natural kind property as a matter of nomological or metaphysical necessity. Thus, even if—contrary to what we have argued—the first-person deployment of phenomenal concepts did somehow reveal that consciousness is non-physical, that wouldn't itself undermine the NK strategy. The NK strategy could still be pursued, provided that the various marks of consciousness cluster together and that the non-physical property of being conscious marches in step with the underlying cause of that clustering, whatever it may be.

(5) The Starting Point Objection

We turn now to one of the few objections to have been specifically raised against the NK approach: the *starting point objection* (Phillips 2018). The worry is that the NK strategy cannot get started because there isn't enough pre-theoretical agreement on the markers/indicators of consciousness. As Phillips puts it:

> It is undoubtedly true that some measures such as explicit verbal report of awareness do provide fairly uncontroversial positive evidence of consciousness. However, such superficial consensus masks the fact that even very early on in our inquiry we face profound and longstanding controversies concerning how to measure consciousness. Furthermore, it is not unreasonable to think that our initial choice of evidence will make a dramatic difference to our initial sample—a difference dramatic enough to change the number of clusters eventually found by our causal modelling.

The starting point objection would represent a serious threat to the NK strategy if rival camps started with just one or two criteria for the ascription of consciousness, and if there were no overlap between their pre-theoretical starting points. But on our reading of the literature that is not the case. Although there certainly is disagreement about putative markers of consciousness, there is also a great deal of agreement between theorists. The vast majority of theorists hold that

consciousness is implicated in verbal report, intentional behaviour, storage in episodic memory, and much else.

Second, the NK strategy is relatively insensitive to one's starting point, since an important part of the method is to enlarge the initial marks into a much wider collection of putative indicators of consciousness. We can think of a natural kind as a sort of attractor, and thus one would expect that research groups that begin with different sets of marks would each enlarge their tests in different ways and eventually converge on the same underlying kind (should it exist).

Of course, there is no guarantee that research groups with different starting points (in terms of their pre-theoretical commitments) *will* converge on the same natural kind. Perhaps different groups will be drawn towards distinct natural kinds (in the way that one research group might be towards the virus associated with hepatitis A and another might be drawn towards the virus associated with hepatitis B). Or perhaps one research group will be drawn towards a natural kind property whereas research groups with distinct starting points won't. But the fact that these possibilities cannot be ruled out *a priori* cuts no ice against the NK strategy, for they also leave open the possibility that the NK strategy *will* succeed, with different research groups converging on the same kind. We see no way of deciding whether convergence is likely short of actually pursuing the NK strategy.

In our view, the only scenario that would undermine the NK strategy from the get-go is one in which pre-theoretical disagreement about the marks of consciousness is so radical that there is little reason to think that the marks used by different research groups cluster together—or that if they cluster, it isn't in virtue of a univocal natural reason. But we are not in that position. There is substantial overlap between the marks that different researchers take to be associated with consciousness, as we noted above. Furthermore, the extent to which researchers from different laboratories and traditions succeed in communicating about consciousness—going to the same scientific conferences, publishing in the same journals—suggests that their starting points are not so radically different.

In short, examining our starting point—the current state of play in consciousness science—should not discourage us from pursuing the NK strategy.

(6) The Anthropocentric Objection

We turn now to a final—and, from the perspective of this special issue, most pressing—objection: the anthropocentric objection. The worry here is that the NK strategy must start by interrogating the putative markers of consciousness in humans (more specifically, in adult humans), but it is then unclear how the results of its investigation can be legitimately extended to non-humans. If we begin with markers of consciousness in human beings, then we run the risk of assuming that features of consciousness that are merely contingently connected with consciousness (i.e., in us, but not in other creatures) are necessarily connected with consciousness. Put another way: how do we distinguish markers of consciousness as such from markers of human consciousness?

This question has particular bite here, for we earlier touted the NK strategy as having virtues that other approaches to validating measures of consciousness lack. If it turns out that the NK strategy is either impotent (that is, gives no verdict) or unreliable (that is, gives the wrong verdict) when applied to non-human animals and AI systems then we would have made little progress in meeting the validation challenge. For example, we might find an excellent diagnostic measure for consciousness in mammals that targets aspects of mammalian neural structure: perhaps it's that thalamico-cortical loops resonate in a characteristic way. That would be entirely useless for measuring consciousness in octopuses, whose brain and nervous system is structured differently. Measures based on neuron types, transmitter types and gene expression could similarly turn out to be unhelpfully parochial. They might tell us nothing about how to measure consciousness in birds, reptiles, crustaceans and insects, let alone AI systems.

So how do we distinguish markers of consciousness as such from markers of human consciousness? The answer will depend on successfully pursuing the NK strategy and then going on to understand how consciousness produces the observable marks of consciousness. Compare nitrogen, say. Being a gas is characteristic of all of our ordinary interactions with nitrogen; also of many years of scientific interaction. But we learnt what it is that produces the characteristic properties of nitrogen and we can see that being a gas is a contingent property: it is a result of how the operative principle interacts with ordinary temperatures and

pressures. Go outside that and nitrogen can be a liquid. Similarly with consciousness. Suppose we find that there is some fundamental information-processing property that accounts for why most of the marks of consciousness tend to cluster together. We could then see why the presence of a certain type of neuron—von Economo neurons, as it might be—is crucial to the realisation of that information-processing property in the mammalian brain. That is consistent with the same information-processing property being realised a different way in (say) the brains or birds or cephalopods.

The NK strategy doesn't guarantee that we'll be able to achieve the sort of understanding that would allow us to work out which marks are clade-specific. So the worry is a real one: we might end up with a good way of diagnosing consciousness in mammals but with little clue of what to say about consciousness in crustaceans. On the other hand, the NK strategy, together with the normal scientific abductive method, does contain tools that, in favourable circumstances, would allow us to go beyond our pre-theoretical measures.

Whether we could also say something useful about consciousness in AI systems depends on the nature of the natural kind uncovered by the NK strategy. If an information-processing property unites the cluster, then it would be relatively straightforward to assess whether an artificially intelligent computer processes information in that characteristic way. If the explanatory property is instead proprietary to living things, or perhaps even to organic life, then difficult questions would arise about the possibility of non-organic forms of consciousness (parallel to questions about whether non-organic life is possible). The NK strategy would not on its own answer those questions, but it would have taken us a long way forward in our understanding of the phenomenon we're interested in.

Finally, we might ask whether the NK strategy is not just anthropocentric, but adult-centric. Will it lead to measures that are only applicable to adult humans and not, for example, to infants? The literature on infant pain points out that not all features of adult pain apply to infants (Goksan et al. 2015; Hu & Iannetti 2016; Tracey 2011). The NK strategy is well equipped to tackle this problem. Since infants are humans, there is little prospect that consciousness is generated in a different way—with a different substrate, say—in infants. (That would require

that one way of being conscious arises in infants, and is then switched off in development and replaced with a different way of being conscious.) Discovering that consciousness is a particular natural kind property would provide us with the resources to say which features of adult pain are constitutive of pain and which are merely contingent. If it succeeds, we will have discovered the underlying univocal reason why the diverse marks of conscious pain cluster together. The question about pain in infants is answered by asking whether that property is present in infants or not.

(7) Conclusion

One of the main challenges facing the science of consciousness is that of validating putative measures of consciousness. Existing research programmes appeal either to correlations with current, pre-theoretical measures of consciousness (such as verbal report), or to contested theories of consciousness. Neither approach is satisfactory. We argue that there is a more plausible alternative to both approaches: the NK strategy. Our aim in this paper has been to outline the underlying motivation for pursuing the NK strategy, and to address the most pressing objections to it.

Of course, even if we have been successful in these aims it by no means follows that the NK strategy will itself succeed. Not only might there be additional objections to the NK strategy, it might also fail for the simple reason that consciousness isn't a natural kind. But these are topics for another occasion. Here we have attempted only to remove some of the chief impediments that lie in the path of the NK strategy.[7]

[7] We thank Jonathan Birch, Peter Carruthers, Cecily Whiteley and two referees for their very helpful comments on previous drafts of this paper. This research was supported by a Future Fellowship from the Australian Research Council (FT 150100266) (TB), the Brain, Mind & Consciousness program of the Canadian Institute for Advanced Research (CIFAR) (TB); and the European Research Council under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 681422 (MetCogCon) (NS).

References

Allport, A. 1988. What concept of consciousness? In A. Marcel and E. Bisiach (Eds.) *Consciousness in Contemporary Science*. Oxford: OUP.

Ball, D. 2009. There are no phenomenal concepts. *Mind*, 118: 935-62.

Balog, K. 2009. Phenomenal concepts. In B. McLaughlin, A. Beckermann & S. Walter (Eds.) *Oxford Handbook in the Philosophy of Mind*. Oxford University Press, pp. 292-312.

Bayne, T. 2010. *The Unity of Consciousness*. Oxford: Oxford University Press.

Bayne, T., Hohwy, J. & Owen, A.M. 2016. Are there levels of consciousness?, *Trends in Cognitive Sciences*, 20/6: 405-13.

Birch, J. Unpublished. Invertebrate Consciousness: Three Approaches.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences,* 30(5), 481-498.

Boyd, R. 1991. Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies*, 61(1-2): 127–148.

Carey, S. 1991. Knowledge acquisition: Enrichment or conceptual change? In S. Carey and R. Gelman (Eds.), *The Epigenesis of Mind: Essays on Biology and Cognition.* Erlbaum, Hillsdale.

Carruthers, P. 2018. Comparative psychology without consciousness. *Consciousness and Cognition*, 63; 47-60.

Casali, A. G., Gosseries, O. & Rosanova, M. et al. 2013. A theoretically based index of consciousness independent of sensory processing and behavior, *Science and Translational Medicine*, 5: 198ra105. DOI: 10.1126/scitranslmed.3006294.

Casarotto, S., Comanducci, A. & Mario, R. et al. 2016. Stratification of unresponsive patients by an independently validated index of brain complexity. *Annals of Neurology*, 80/5: 718-29.

Chalmers, D. 1996. *The Conscious Mind*. New York: OUP.

Chalmers, D. 2003. The content and epistemology of phenomenal belief. In Q. Smith and A. Jokic (Eds.) *Consciousness: New Essays*. Cambridge: Cambridge University Press.

Churchland, P.S. 1988. Reduction and the neurobiological basis of consciousness. In A.J. Marcel and E. Bisiach (Eds.) *Consciousness in Contemporary Science*. Oxford: Oxford University Press, pp. 273-304.

Crane, T. 2005. Papineau on phenomenal concepts. *Philosophy and Phenomenological Research*, 71/1: 155-62.

Dainton, B. 2000. *Stream of Consciousness.* London: Routledge.

Dehaene,S., Lau, H. & Kouider, S. 2017. What is consciousness, and could machines have it? *Science*, 358(6362): 486-92.

Dehaene, S. & Naccache, L. 2001. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79: 1-37.

Dehaene, S., Changeux, J.-P. & Naccache, L. 2011. The Global Neuronal Workspace Model of Conscious Access: From Neuronal Architectures to Clinical Applications. In S. Dehaene and Y. Christen (Eds.) *Characterizing Consciousness: From Cognition to the Clinic? Research and Perspectives in Neurosciences*. Springer-Verlag Berlin, pp. 55-84.

Faivre, N., Mudrik, L., Schwartz, N. and Koch, C. 2014. Multisensory Integration in complete unawareness: Evidence from audiovisual congruency priming, *Psychological Science*, 25/11: 2006-2016.

Faugeras, F. Rohaut, B. & Weiss, N. et al. 2012. Even related potentials elicited by violations of auditory regularities in patients with impaired consciousness. *Neuropsychologia*, 50: 403-18.

Godfrey-Smith, P. 2020. Varieties of subjectivity. *Philosophy of Science*.

Goff, P. 2011. A posteriori physicalists get our phenomenal concepts wrong. *Australasian Journal of Philosophy*, 89/2: 191-209.

Goff, P. 2015. Real acquaintance and physicalism. In P. Coates and S. Coleman (Eds.) *Phenomenal Qualities*. Oxford: Oxford University Press.

Goff, P. 2017. *Consciousness and Fundamental Reality*. Oxford: Oxford University Press.

Goksan, S. et al. 2015. fMRI reveals neural activity overlap between adult and infant pain, eLife 4, e06356.

Gosseries, O., Di, H., Laureys, S., Boly, M. 2014. Measuring consciousness in severely damaged brains. *Annual Review of Neuroscience*, 37: 457-78.

Griffiths, P. E. 1999. Squaring the circle: natural kinds with historical essences. In. R. A. Wilson (ed) *Species: New Interdisciplinary Essays*. Cambridge, M.A, MIT Press, pp. 209-228.

Horgan, T. and Tienson, J. 2001. Deconstructing new wave materialism. In C. Gillett and B. Loewer (Eds.) *Physicalism and its Discontents*. Cambridge: Cambridge University Press.

Hu, L. & Iannetti, G. 2016. Painful issues in pain prediction. *Trends in Neurosciences*, 39/4: 212-20.

Huang, Z. et al. 2018. Brain imaging reveals covert consciousness during behavioral unresponsiveness induced by propofol, *Scientific Reports*, 8, 13195.

Irvine, E. 2013. *Consciousness as a Scientific Concept: A Philosophy of Science Perspective*. Dordrecht: Springer.

King, J.R., Faugeras, F., Gramfort, A. 2013. Single-trial decoding of auditory novelty responses facilitates the detection of residual consciousness, *Neuroimage*, 83: 726-38.

Kitcher, P. 1988. The child as parent of the scientist. *Mind and Language*, 3: 217-28.

Kouider, S. et al. 2013. A neural marker of perceptual consciousness in infants. *Science*, 340: 376-80.

Kripke, S. 1980. *Naming and Necessity*. Oxford: Blackwell.

Kuhn, T. S. 1964. A function for thought experiments. In *The Essential Tension: Selected Studies in Scientific Tradition and Change*. University of Chicago Press. pp. 240-265.

Levine, J. 2006. Phenomenal concepts and the materialist constraint. In T. Alter & S. Walter (Eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford: OUP, pp. 145-166.

Loar, B. 1990. Phenomenal states. *Philosophical Perspectives*, 4: 81-108.

Loar, B. 1997. Phenomenal states [second version]. In N. Block, O. Flanagan, & G. Güzeldere (Eds.) *The Nature of Consciousness*. London/Cambridge MA: MIT Press, pp. 597-616.

Macdonald, C. 2004. Mary meets Molyneux: The explanatory gap and the individuation of phenomenal concepts, *Noûs*, 38/3: 503-524.

Millikan, R. G. 1998. A common structure for concepts of individuals, stuffs, and real kinds: More mama, more milk, and more mouse. *Behavioral and Brain Sciences*, 21(1), 55-65.

Mudrik, L., Faivre, N. and Koch, C. 2014. Information integration without awareness, *Trends in Cognitive Sciences*, 18(9): 488-96.

Nida-Rümelin, M. 2007. Grasping phenomenal properties. In T. Alter & S. Walter (Eds.) *Phenomenal Concepts and Phenomenal Knowledge*. New York: Oxford University Press.

Papineau, D. 2002. *Thinking About Consciousness.* Oxford: OUP.

Perry, J. 2001. *Knowledge, Possibility, and Consciousness*. Cambridge, MA: MIT Press.

Phillips, I. 2014-a. Experience of and in time, *Philosophy Compass*, 9/2: 131-44.

Phillips, I. 2014-b. The temporal structure of experience. In D. Lloyd and V. Arstila (Eds.) *Subjective Time: The Philosophy, Psychology, and Neuroscience of Temporality*. Cambridge, MA: MIT Press.

Phillips, I. 2018. The methodological puzzle of phenomenal consciousness, Phil. Trans. R. Soc. B 373 (1755), 20170347.

Scott, R.B., Samaha, J., Chrisley, R. & Dienes, Z. 2018. Prevailing theories of consciousness are challenged by novel cross-modal associations acquired between subliminal stimuli, *Cognition*, 175: 169-85.

Seeff, L. B. 2009. The history of the "natural history" of hepatitis C (1968–2009). *Liver International*, 29(s1), 89-99.

Shea, N. 2012. Methodological encounters with the phenomenal kind, *Philosophy and Phenomenological Research*, 84(2): 307-44.

Shea, N. & Bayne, T. 2010. The vegetative state and the science of consciousness, *British Journal for the Philosophy of Science*, 61: 459-84.

Sitt, J. et al. 2014. Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. *Brain*, 137, 2258-2270.

Sundström, P. 2011. Phenomenal concepts. *Philosophy Compass*, 6/4: 267-81.

Thagard, P. 1992. *Conceptual Revolutions*. Princeton, NJ: Princeton University Press.

Tononi, G. 2008. Consciousness as integrated information: a provisional manifesto. *The Biological Bulletin, 215*(3), 216-242.

Tracey, I. 2011. Can neuroimaging studies identify pain endophenotypes in humans? *Nature Reviews Neurology*, 7: 173-81.

Travers, E., Frith, C. & Shea, N. 2018. Learning rapidly about the relevance of visual cues requires conscious awareness, *Quarterly Journal of Experimental Philosophy*, 71/8: 1698-1713.

Trogdon, K. 2017. Revelation and physicalism. *Synthese*, 194/7: 2345–2366.

Tye, M. 2000. *Consciousness, Color, and Content*. Cambridge, MA: MIT Press.

Tye, M. 2009. *Consciousness Revisited: Materialism without Phenomenal Concepts*. Cambridge, MA: MIT Press.

Wiser, M. & Carey, S. 1983. When heat and temperature were one. In D. Gentner and A. Stevens (Eds.), Mental Models. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 267-97.