UNIVERSITY OF BIRMINGHAM
University of Birmingham Research at Birmingham

Consciousness as a natural kind and the methodological puzzle of consciousness

Taylor, Henry

DOI:
10.1111/mila.12413

License:
Other (please specify with Rights Statement)

Document Version
Peer reviewed version

Citation for published version (Harvard):
Taylor, H 2022, 'Consciousness as a natural kind and the methodological puzzle of consciousness', *Mind & Language*. https://doi.org/10.1111/mila.12413

Link to publication on Research at Birmingham portal

Publisher Rights Statement:
This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions. This article may not be enhanced, enriched or otherwise transformed into a derivative work, without express permission from Wiley or by statutory rights under applicable legislation. Copyright notices must not be removed, obscured or modified. The article must be linked to Wiley's version of record on Wiley Online Library and any embedding, framing or otherwise making available the article or pages thereof by third parties from platforms, services and websites other than Wiley Online Library must be prohibited.

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

•Users may freely distribute the URL that is used to identify this publication.
•Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
•User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
•Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.

Download date: 26. Apr. 2024

Consciousness as a Natural Kind and the Methodological Puzzle of Consciousness

Henry Taylor

Abstract

A new research programme conceives of consciousness as a natural kind. One proposed virtue of this approach is that it can help resolve the methodological puzzle of consciousness, which involves distinguishing consciousness from cognitive access. The present paper raises a novel problem for this approach. The problem is rooted in the fact that we may have misclassified episodes of consciousness as not episodes of consciousness. I argue that conceiving of consciousness as a natural kind cannot distinguish consciousness from cognitive access.

1: The methodological puzzle

Standardly, we ascertain that a subject is conscious of a stimulus through the subject's report. To report a stimulus, the subject has to cognitively access information about it. Therefore, if there are conscious experiences that the subject does not cognitively access, we cannot establish their existence using reportability measures. That is, standard measures of consciousness cannot lead us to establish the existence of the following scenario (which I call the 'no access scenario' or NAS):

> **(NAS):** Subject S has a phenomenally conscious experience at time t, but S did not cognitively access the experience at t.

The problem is not that situations like (NAS) exist. They may not. The problem is that *if* they did exist, standard methodology couldn't establish them. That is the methodological puzzle of consciousness (Block 2007).[1]

This puzzle lies at the heart of the disagreement between alternative scientific theories of consciousness. Global workspace theory says that unconscious sensory information is only boosted to consciousness when attentional amplification allows it to be cognitively accessed, and thereby brought into the global workspace. This system makes information directly available to a wide variety of consumer systems such as voluntary action control and language. The workspace is associated with prefrontal and

[1] Michel (2019) traces the problem throughout the history of consciousness science.

anterior cingulate cortices (Dehaene *et.al.* 2006, Sergent *et.al.* 2005, Cohen and Dennett 2011). Global workspace theory claims that information only becomes conscious when it is cognitively accessed, so denies that (NAS) ever occurs. Conversely, recurrent processing theorists claim that (sensory) consciousness arises whenever there is recurrent processing of sensory information. This includes areas of visual cortex such as V1 and V4, many of which are not cognitively accessed by the global workspace system (Lamme 2004, 2006). Thus, recurrent processing theorists claim that (NAS) occurs. In these ways, two major scientific camps concerning consciousness offer different answers to whether (NAS) occurs. The methodological puzzle poses a challenge to our ability to know whether this is the case, and is thus an obstacle to resolution of the debate between global workspace and recurrent processing theorists. For this reason, it is amongst the most pressing challenges to consciousness science.

Many paradigms have been proposed to help with the puzzle, including partial report (Block 2014), no-report, (Cohen *et.al.* 2020), no-cognition (Brascamp *et.al.* 2015, Block 2019), and abductive inferences from psychological theories to phenomenology (Denison *et.al.* 2020). This paper's primary focus is a less explored approach, which is to study consciousness as a natural kind (Shea 2012, Bayne 2018, Bayne and Shea forthcoming, Shea and Bayne 2010). The natural kinds approach is ambitious and wide-ranging, but has so far received little critical attention (one exception is Phillips (2018)). A thorough appraisal is due.

The natural kinds framework is intended as a general methodology for consciousness science. Given that one of the main claimed virtues of the natural kinds approach is its ability to resolve the methodological puzzle, its failure in this regard will correspondingly reduce our credence in the methodology as a whole. So, the consequences of this paper ripple beyond the methodological puzzle.

By 'consciousness' I mean 'phenomenal consciousness' throughout. I won't attempt to define 'consciousness', or 'cognitive access' but assume they are clear enough for discussion. There is a difference between information *accessed* by the workspace, and information *accessible* to the workspace. Different theories place different conditions on what kind of access (if any) is required for consciousness (contrast Dehaene *et.al.* 2006 with Prinz 2012). In this paper, I am primarily concerned with whether there can be conscious content that is not *accessed,* as that has been the locus of debate. Much of what I say is relevant to the question of whether there can be *inaccessible* conscious content, but I leave this implicit here.

In section 2, I divide the natural kinds framework into four steps. I then outline the framework's difficulties by presenting a hypothetical but empirically realistic scenario, in which following the steps of the framework will lead to the incorrect answer to the methodological puzzle. The scenario rests on two ideas. First, the possibility of 'false negatives': cases where a particular behavioural or neuroscientific property is causally underpinned by consciousness, but we have misclassified it as being not causally underpinned by consciousness (section 3). The second idea is that consciousness may be associated with more than one natural kind property. In short, the problem (explored in section 4) is that the presence of false negatives can lead us to incorrectly classify an underlying natural kind property as *not* an instance of consciousness, when in fact it is. The fact that consciousness may be associated with multiple natural kinds prevents us from correcting this mistake. The result is that (in the hypothetical but empirically realistic scenario) we will reach the incorrect answer to the methodological puzzle (section 4). In section 5, I consider no-report paradigms, then discuss objections and replies (section 6). Finally, I draw parallels between the natural kinds framework and Block's own approach (section 7).

2: The natural kinds framework

2.1: Natural kinds

Natural kinds are groups of entities that support scientific inductions, projections and generalisations. By studying a phenomenon as a natural kind, a characteristic shift comes about from defining the phenomenon in terms of its readily observable properties to defining it in terms of its underlying nature. We start by characterising gold as a yellow and malleable metal, and then shift to defining it in terms of atomic number 79. This property (along with background theory) *explains* the observable properties with which our investigation began. We can then develop new ways to test for the underlying nature, and study cases that do not have the observable properties with which we began.

The dominant view of natural kinds takes them to be clusters of properties that reliably co-occur because of an underlying property or mechanism (Boyd 1989, 1999, Shea 2012, p.326, Taylor 2019, Kornblith 1993, Khalidi 2013 ch.4).[2] The properties in the cluster are not necessary and sufficient for kind

[2] There is debate over whether the view can account for all biological natural kinds (Ereshefsky 2010, Ereshefsky and Matthen 2005). I'm concerned with psychological l kinds, so I pass over these issues.

membership. Something can lack some properties and still be a member of the kind. A *natural kind property* is the property that causally underpins these properties, and thereby causally explains why the properties tend to cluster together (in the case of gold, this property is atomic number 79).

The application to consciousness proceeds in four steps (Shea 2012, Bayne 2018, Shea and Bayne 2010). The first I call the *marker assembly step*. Start by assembling the properties that are taken to be markers of consciousness. A 'marker' of consciousness is any property that is indicative of the presence of consciousness of some stimulus. The most obvious marker is explicit verbal report of the stimulus by the subject. If a subject with normal vision views a red ball in good lighting and says 'I see a red ball', that is an indicator that they had a conscious experience of the ball. Markers need not be limited to verbal report. They can be any functional, behavioural and/or neuroscientific properties that reliably indicate the presence of consciousness of some stimulus (Shea 2012, pp.329-330; Jack and Shallice 2001). Indeed, the set of markers can include folk claims about consciousness (e.g. 'being scared makes my heart rate increase').

The natural kinds approach involves treating this set of markers as the cluster of properties characteristic of a natural kind, and consciousness as the natural kind property that causally underpins the cluster. So, the markers are taken to be properties that are causally underpinned by a natural kind property, which is an instance of consciousness. We use causal modelling to identify this natural kind property. That is, we identify the property that causally underpins all of the marker properties that are associated with the presence of a conscious episode. Then (the framework claims) we would have reason to identify this property with consciousness.[3] Call this the *causal modelling step*.

To the framework, a *marker* of consciousness turns out to be a property that is indicative of the presence of consciousness because it is reliably causally underpinned by a natural kind property that is an instance of consciousness. Once we have identified this consciousness property, we examine its causal profile further, and discover other properties that are causally underpinned by the consciousness property, which we did not initially know were underpinned by consciousness. Our list of markers of consciousness

[3] I assume for simplicity that the natural kind property would be *identical* with consciousness, but my arguments apply if the natural kind property grounds (realises, subvenes, causes…) the consciousness property.

can then be expanded to include these novel properties. In this way, the method will allow us to develop new techniques for detecting the presence of consciousness. This is the *marker expansion step*. During this step, we can also abandon properties that we previously took to be markers of consciousness, but which turn out not to be causally underpinned by the consciousness property. This marker expansion step is crucial to the approach, as we will see.

2.2: Back to the methodological puzzle

Step four can be called the *methodological puzzle step,* as it is where the natural kinds framework aims to resolve the puzzle. Once we have isolated the underlying kind property (to be identified with consciousness), we can determine whether that property co-occurs with cognitive access or not. Suppose that our causal modelling reveals that there is one natural kind property underpinning our markers and that this natural kind property always correlates with cognitive access (figure 1). In figure 1, T_1-T_7 represent our known markers of consciousness, and K is the kind property that causally underpins them. In this case we would have evidence that the no access scenario (NAS) does not occur (Shea 2012).

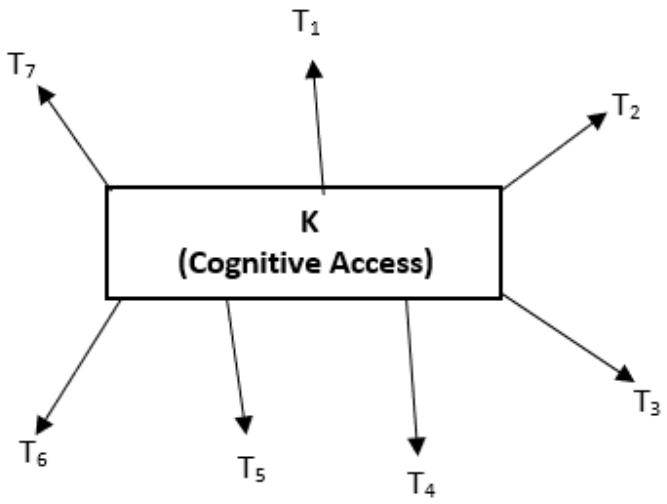

Figure 1: an 'easy' case where all our markers of consciousness converge on one kind property, which correlates with cognitive access.

Things are more complicated in a *two-property cluster* case, where our initial markers of consciousness are underpinned by two natural kind properties (figure 2).

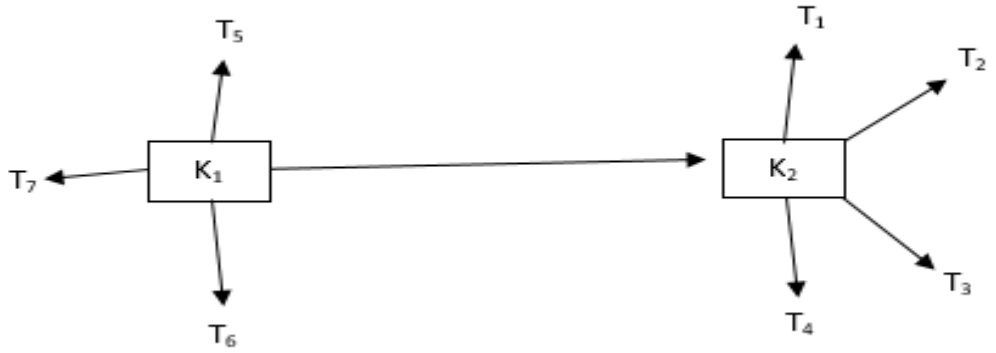

Figure 2: Two underlying natural kinds.

Suppose that K_2 is correlated with cognitive access. In this case, we have established that *some* of our markers of consciousness (T_1-T_4) are causally underpinned by a property that correlates with cognitive access, but some of them are causally underpinned by another underlying property (K_1), which is itself causally linked to K_2. Now the reasoning becomes complicated. Shea's interpretation (2012, p.337) is that, since K_1 is reliably correlated with some of the markers of consciousness, we have evidence that K_1 is an instance of consciousness, and K_2 is merely cognitive access to that conscious content. Therefore, there can be cases of consciousness without access. (This assumes that K_1 can occur without K_2, otherwise, K_1 would itself correlate with cognitive access. This assumption can be tested using further causal modelling). There are other interpretations of the situation in figure 2. Ian Phillips (2018, p.7) suggests that similar reasoning supports the claim that K_2 is the consciousness property. He notes that K_1 would be a good candidate for a 'preconscious' system, which consists of contents that are *potentially* conscious, but require attentional modulation to bring them into the global workspace, and boost them to consciousness (Dehaene *et.al.* 2006). Since the purpose of this section is only to explain how the natural kinds approach works in principle, I will not choose between these interpretations here. The problem I will raise for the approach is more general.

3: False negatives

The framework starts with the marker assembly step: identify an initial set of (behavioural/functional/neuroscientific) properties that are causally underpinned by the kind property that

is an instance of consciousness (these are the *markers* of consciousness). To do this, we must decide which properties to include in the set of markers, but also which to *exclude*. That is, we must decide which properties we *do not* take to be causally underpinned by consciousness. This raises a danger: that we may incorrectly exclude a property from the set. There may be a property that *is* causally underpinned by consciousness (and which should therefore be taken as a marker of the presence of consciousness) but that we have misclassified as *not* underpinned by consciousness. This would be a *false negative* because we have *falsely* concluded that the property is *not* causally underpinned by consciousness, when in fact it is.[4] As well as false negatives there will also be what I call *known markers:* properties that are causally underpinned by consciousness, that we correctly take to be indicative of the presence of consciousness. Known markers are the 'true positives'.

It is uncontroversial that our set of markers may contain false negatives. Even advocates of the framework accept that our initial set of markers will be imperfect. With this in mind, I will provide just two concrete examples of how false negatives could in principle arise. In section 4, I will show how this is problematic for the natural kinds framework.

3.1: Stem completion

In stem completion (Debner and Jacoby 1994), a prime word (e.g. 'HOSPITAL') is sandwiched between two other words (various presentation times are used). Then a stem (e.g. 'HOS-') is presented. In the 'exclusion' condition (which is the crucial one) subjects must complete the stem with any word *other* than the prime (e.g. 'HOST' would be correct, 'HOSPITAL' would be incorrect). If subjects succeed the exclusion condition (i.e. complete the stem word with a word other than the prime) they are 'insensitive' to the stem completion effect. If subjects fail (by completing the stem with the prime word) they are 'sensitive' to stem completion. The standard interpretation is that success at the exclusion condition (insensitivity to stem completion) indicates that the subject consciously perceived the prime word, whilst failure in the exclusion condition (sensitivity) does not involve conscious perception of it. On these grounds, advocates

[4] Note that a false negative is a property that is causally underpinned by the natural kind property that is identical with consciousness, but which we have incorrectly excluded from our set of markers. A false negative is not the underlying natural kind property itself.

of the natural kinds approach include insensitivity to stem completion in the set of markers that are causally underpinned by consciousness. Conversely, sensitivity to stem completion is not associated with conscious perception of the prime, and so it is not allowed into the set of markers indicative of consciousness.

However, another interpretation is possible, which is that in the sensitivity case, subjects *do* consciously perceive the prime word, but do not exclude it from consideration when they go to complete the stem. Snodgrass suggests that subjects may do this simply because they lack confidence in their identification of the prime (in signal detection theory terms, they have a conservative response criterion (2002, p.557)). This is supported by evidence that offering monetary rewards makes subjects more likely to successfully exclude the prime word, and complete the stem with another word, thus succeeding at the exclusion task (Visser and Merikle 1999).[5] The crucial point is that, if this latter interpretation is correct, then sensitivity to stem completion will actually be causally underpinned by conscious perception of the prime. In such a case, this property will be causally underpinned by consciousness, and we will have been incorrect to exclude it from our set of markers indicative of consciousness. It would be a false negative.

3.2: Trace conditioning/ delay conditioning

This is not a problem peculiar to stem completion. Turn to another paradigm that is suggested by the advocates of the approach as a source of markers of consciousness: trace conditioning and delay conditioning (Perruchet 1985, Perruchet *et.al.* 2006, Weidemann and Lovibond 2016). Subjects are exposed to a tone, which is sometimes followed by an air puff to the eye after the tone has ended (trace condition) and sometimes during the tone, after its onset (delay condition). Subjects were asked whether they expected the puff to follow the tone. In some versions of the experiment, subjects ranked their expectations using button presses on a scale of 1-7 (Perruchet 1985, p.165). In others, they filled in a questionnaire to indicate their expectations (Clark and Squire 1998). The paradigm tests whether their eye blinks correspond with this expectation (i.e. whether they only blinked when they expected the puff to follow the tone). In the trace condition, whether the subjects give an eye blink response is dependent on them expecting the puff to follow the tone. In the delay condition, their eye blink responses are not dependent on their expectations.

[5] Thanks to Ian Phillips for drawing my attention to this work.

The standard interpretation is that an eye blink in the *trace* condition indicates that the subject is conscious of the contingency between tone and puff. Therefore, an eye blink in trace conditioning can be included in the set of markers indicative of consciousness. Conversely, an eye blink in the delay conditioning case is not thought to indicate that the subject was conscious of the contingency between tone and puff (because this eye blink did not correlate with subjects' reported expectations about whether the puff would follow the tone). Therefore, an eye blink in the delay conditioning case is excluded from the set of markers of consciousness.

Again, there is an alternative interpretation, on which subjects are conscious of the contingency between tone and puff in the delay case. Lovibond and Shanks point out that experimenters placed a very high demand on what it takes for a subject to count as 'aware' of the contingency between tone and puff (2002, p.12). By setting the threshold for awareness lower, more participants in the delay conditioning case would have counted as 'aware' of the stimuli. For these reasons, it is possible that subjects in the delay conditioning case were conscious of the contingency between tone and puff (Lovibond and Shanks 2002, p.12). The experiment didn't use masking techniques to diminish conscious perception (Lovibond and Shanks 2002, p.12).[6] If this is the case, then eye blinks in cases of delay conditioning would in fact be causally underpinned by consciousness of the contingency between tone and puff. In which case, we will have made a mistake by not including eye blinks in delay conditioning from our list of the markers of consciousness. Eye blinks in delay conditioning would be another false negative.[7]

I'm not saying that sensitivity to stem completion and eye blinks in delay conditioning are causally underpinned by consciousness, and that we have incorrectly excluded them from the set of markers of consciousness (making them false negatives). Rather, I am raising this as an empirical possibility, in order to demonstrate how false negatives can in principle arise, and to show that we must take the possibility of

[6] Note that masking is designed to extinguish conscious perception of the *stimuli*, not the *contingency* between the stimuli.
[7] When masking techniques are used to prevent conscious perception of the stimuli, the conditioning effect disappears for *both* delay and trace conditioning (Skora *et.al* 2021). Initially, this looks like it supports my suggestion that delay conditioning requires consciousness of the contingency. However, caution is advised in inferring from these results to the case that has been my primary point of discussion, as they are different in at least two ways. First, the Skora *et.al.* experiment used visual stimuli (not air puffs and tones), so there may be differences across sense modalities. Second, the kind of conditioning involved in the Skora *et.al.* experiment was more complex (it required effortful co-ordination of action on the part of the participant, rather than just an eye blink).

false negatives seriously. Here I have given only two plausible examples of where false negatives may arise, to give a sense of how issues may arise, but there will of course be many more.

4. False negatives and the methodological puzzle

So far I have only argued that false negatives are a realistic empirical possibility. The advocate of the natural kinds approach will certainly accept the possibility of false negatives in principle. However, they will reply by deferring to the marker expansion step. There may indeed be properties that are causally underpinned by consciousness, but that we have excluded from our set of markers (false negatives). However (the advocate will claim) over time we will come to realise that they are causally underpinned by consciousness, and then include them in our set of markers of consciousness. The false negatives will disappear. So the advocate of the approach will claim.

In this section, I show that this does not work. I present a hypothetical (but empirically realistic) scenario, and show that, by following the steps of the natural kinds framework, we *will* be led to the incorrect answer to the methodological puzzle. Specifically, in the scenario, (NAS) does occur, but the framework would tell us it doesn't. I show that this cannot be remedied by the marker expansion step. The scenario relies on two core ideas. First, that there may be false negatives. Second, that it is at least a realistic empirical possibility that consciousness may be associated with more than one natural kind property (I argue for this in 4.1). In the scenario, the false negatives issue leads us to misclassify an instance of consciousness as *not* an instance of consciousness. The multiple kinds issue then prevents us from correcting our false negatives. The result is that (in this empirically realistic scenario) the framework would lead us to the incorrect answer to the methodological puzzle.

4.1: A multiple kinds view of consciousness.

The claim that consciousness may be associated with more than one natural kind is not just the claim that different instances of consciousness are different from each other in certain ways. Rather, it is that consciousness may be like jade (Kim 1992). All instances of jade are similar in certain ways (in colour). However, jade is underpinned by two distinct natural kind properties (jadeite and nephrite), each of which have distinct chemical compositions, and hence distinct causal profiles. To say that consciousness is associated with multiple kinds is to say that, similar to jade, different instances of consciousness are

associated with separate underlying natural properties, which have different causal profiles. In this context, to say that two natural properties have *different* causal profiles means that they causally underpin separate clusters of properties. In such a case, the presence of properties in one cluster does not allow us to infer the presence of properties in the other (they are not 'co-projectible'), and the two natural kind properties feature in separate inductive statements. This reflects the role assigned to natural kinds of supporting induction and projection (Shea 2012, p.331; Taylor 2020, p.2083; Khalidi 2013, pp.83-92).[8] Here I do not claim that consciousness definitely is associated with more than one natural kind property. Rather, I claim only that it is a realistic empirical possibility, which must be taken seriously. In support of this, consider that many posits in cognitive science have been discovered to be associated with more than one natural kind property (Machery 2009, Griffiths 1997). At the very least, anyone attracted to the idea that consciousness is associated with one natural kind property must also take seriously the suggestion that it is underpinned by *more* than one.

A multiple kinds view of consciousness may be resisted on the grounds that all instances of consciousness share similarities (they all have phenomenal character, for example). This may be taken to imply that they are all underpinned by the same kind property. The jade example shows the faultiness of this reasoning. A phenomenon can be associated with more than one natural kind, even if all instances of it share some similarities. Jadeite and nephrite are similar in colour, but are underpinned by different natural kind properties.

Someone may worry that a multiple kinds view leads to eliminativism about 'consciousness' as a scientific concept (cf. Irvine 2013). There are several things to be said about this worry. First, the inference from a multiple kinds view to eliminativism about 'consciousness' is anything but straightforward, and can be resisted (Taylor and Vickers 2017, p.35). Some scientific concepts fail to refer to natural kinds, but still play an important role in a science.[9] Second, even if we were to accept the inference from a multiple kinds view to eliminativism, that is not good reason to refuse to even take a multiple kinds view seriously as an

[8] It might be vague when the causal profiles of two properties are different enough to constitute separate natural kind properties. No worries here. Vagueness is a well-known feature of natural kinds in cognitive science (Taylor 2020, Craver 2009).

[9] One candidate is 'hardness' in materials science and 'cortical column' in neuroscience (Haueis forthcoming).

empirical possibility. Eliminativism about 'consciousness' as a scientific concept is a respectable position, worthy of serious consideration (Irvine 2013), not something to be dodged in advance by strategically avoiding commitment to the claims that we think might lead us there.

4.2 Multiple kind properties and false negatives.

I have argued for two core ideas. First, the possibility of properties that are causally underpinned by consciousness, but which we have incorrectly judged to not be causally linked to consciousness, and thereby excluded from our set of markers of consciousness (false negatives). Second, for the empirical possibility that consciousness is associated with more than one natural kind property. By drawing together these ideas, we can generate an empirically realistic scenario, and demonstrate (by following the four steps of the natural kinds framework) that the framework will give the wrong answer to the methodological puzzle in this scenario.

Hypothetically, suppose that there are multiple kind properties that are instances of consciousness (K_1 and K_2). Now, suppose we assemble a list of accepted initial markers of consciousness (T_1-T_7). Insensitivity to stem completion and eye blinks in trace conditioning would be in this set. However, suppose some properties (T_8-T_{12}) are false negatives. They are *also* causally underpinned by consciousness, but we have excluded them from our set of markers of consciousness, for the reasons explained above. Suppose the actual causal situation is as follows (figure 3):

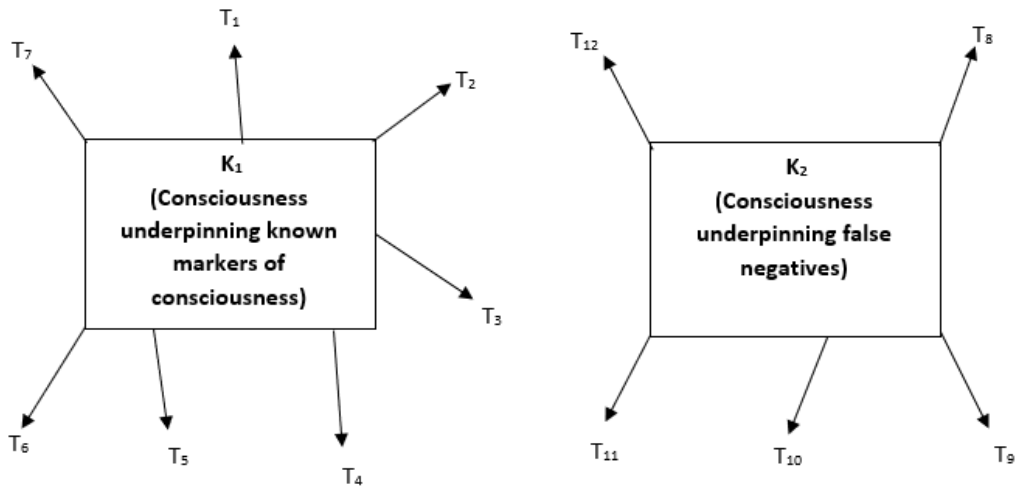

Figure 3: T_1-T_7 represent the widely accepted markers of consciousness (e.g. eye blinks in trace conditioning), which are underpinned by one instance of consciousness (K_1). T_8-T_{12} represent the false negatives: properties that are causally underpinned by consciousness, but which we have excluded from our set of markers (e.g. eye blinks in delay conditioning).

Some clarifications: figure 3 does *not* depict a scenario in which our known markers of consciousness are underpinned by two distinct kind properties. Rather, only T_1-T_7 represent our known markers of consciousness, whilst T_8-T_{12} represent false negatives. Here, our known markers of consciousness (T_1-T_7) all converge on one natural kind property (K_1) but there is another kind property (K_2) that is also an instance of consciousness, and which underpins a set of properties (T_8-T_{12}) that we have incorrectly judged to *not* be causally underpinned by consciousness, and hence incorrectly excluded from our set of markers of consciousness (these are false negatives).

Now in this scenario, K_1 and K_2 are in fact instances of consciousness. However, someone taking the natural kinds approach to consciousness would not realise this. As far as they are concerned, they have taken all the known markers of consciousness (T_1-T_7) and found one natural kind property that underpins them (K_1). Crucially, according to the framework, *we infer that a particular kind property is an instance of consciousness if that property causally underpins known markers of consciousness.* Because K_1 underpins the known markers of consciousness, the framework would then allow us to infer that K_1 is an instance of consciousness. So far so good. But here is where the false negatives issue becomes problematic. In the scenario above, we have misclassified T_8-T_{12} as *not* being causally underpinned by consciousness, and hence incorrectly excluded them from our set of known markers of consciousness. Because they have been incorrectly excluded from the set of known markers, then as far as someone taking the natural kinds approach is concerned, there will be no evidence that the property that underpins them (K_2) is also an instance of consciousness.

If we had (correctly) included T_8-T_{12} into the set of markers of consciousness, then the framework would allow us to infer that K_2 is an instance of consciousness, on the grounds that this property causally underpins at least some of the markers of consciousness. However, since the framework allows us to infer that a kind property is an instance of consciousness if that property underpins our known markers of consciousness, *and* since we have excluded T_8-T_{12} from our set of markers, the framework provides us with no way to infer that K_2 is an instance of consciousness. This is the problem caused by false negatives: that

any natural kind property underpinning a false negative will be incorrectly misclassified as *not* an instance of consciousness, when in fact it is.

The situation we are in is that false negatives will prevent us from recognising that K_2 (which underpins our false negatives T_8-T_{12}) is an instance of consciousness. An advocate of the approach will reply by noting that the marker expansion step will expand our set of markers of consciousness, and fix the false negatives. We may hope that the marker expansion step will expand our set of markers to *include* T_8-T_{12}. That is, to recognise that T_8-T_{12} *are* causally underpinned by consciousness, and thereby provide reason for us to infer that K_2 is an instance of consciousness.

It is here that the presence of multiple natural kind properties becomes problematic, because it prevents the marker expansion step from recognising that T_8-T_{12} are causally underpinned by consciousness, and therefore prevents us from recognising that K_2 is an instance of consciousness, and fixing our mistake. The problem lies in how the marker expansion step works. According to the framework, we take our set of initial markers of consciousness: T_1-T_7 in figure 3 (marker assembly step).[10] We use causal modelling to identify the kind property that causally underpins those markers, and infer that this property is an instance of consciousness (causal modelling step). In this way, we could establish that K_1 in figure 3 is an instance of consciousness. The crucial point is as follows: the marker expansion step works *by discovering novel properties that are causally underpinned by the kind property that has already been established as an instance of consciousness* (again, this is K_1 in figure 3). So the marker expansion step can only lead us to discover novel properties that are underpinned by K_1. It does not extend to properties that are underpinned by *separate* natural kind properties from the one that has already been established as an instance of consciousness. In the situation in figure 3, T_8-T_{12} (our false negatives) are underpinned by a *distinct* natural kind property from the one that we have established as an instance of consciousness (that is, T_8-T_{12} are underpinned by K_2, which is separate from the one that has been established as an instance of consciousness, which is K_1). Because they are underpinned by a separate kind property, the marker expansion step cannot help us

[10] Recall that T_8-T_{12} are not initial markers of consciousness, they are properties we have incorrectly excluded from our set of markers of consciousness, even though they are in fact causally underpinned by consciousness (they are false negatives).

discover that consciousness is also associated with a *separate* natural kind property (K_2), which underpins its own separate cluster of properties (T_8-T_{12}). For this reason, the framework provides us with no evidence for the claim that T_8-T_{12} are causally linked to consciousness. Without realising that T_8-T_{12} are causally linked to consciousness, we have no evidence that the kind property underpinning them (K_2) is itself an instance of consciousness.

If T_1-T_7 and T_8-T_{12} were both underpinned by the same natural kind property, K_1 then things would be simpler. We would infer that K_1 was an instance of consciousness (on the grounds that it causally underpins our known markers of consciousness, T_1-T_7). We could then infer *based on the fact that the very same natural kind property also underpins T_8-T_{12}* that T_8-T_{12} are also underpinned by consciousness. We could then infer that these properties should be taken to be markers of consciousness, and correct our mistake. However, this hinges on T_1-T_7 and T_8-T_{12} both being underpinned by the same natural kind property, which is not the case in the scenario above.

4.3: The methodological puzzle.

The presence of false negatives is what leads us to misclassify K_2 as not an instance of consciousness. The fact that consciousness is associated with multiple kind properties is what prevents us from fixing this mistake by recognising that T_8-T_{12} are causally associated with consciousness, and thereby inferring from that K_2 causally underpins T_8-T_{12} to the conclusion that K_2 is not an instance of consciousness. In short, the core problem caused by the interaction of the false negatives and multiple kinds ideas is that *the framework provides no way for us to ascertain that K_2 is an instance of consciousness.*

Application to the methodological puzzle is now straightforward. Recall that the methodological puzzle step is the last step in the framework: we examine whether the underlying kind property coincides with cognitive access, and if it does, we conclude that (NAS) does not occur. Now, suppose that K_1 coincides with cognitive access, whilst K_2 can occur without cognitive access (figure 4). So in this *hypothetical* situation, consciousness without access does occur.

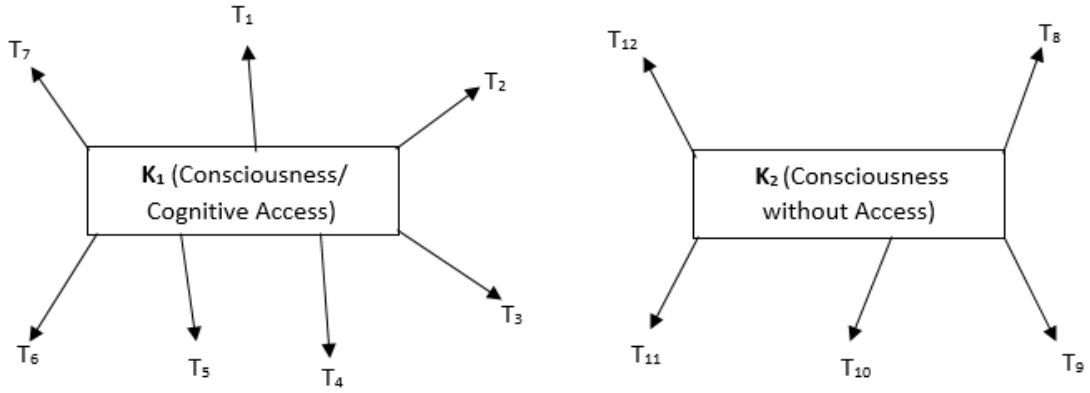

Figure 4: K_1 coincides with cognitive access, and underpins what are widely accepted to be markers of consciousness.K_2 is a case of consciousness without access, but it underpins a set of properties (T_8-T_{12}) that we have incorrectly excluded from our set of properties that are underpinned by consciousness (false negatives).

Someone following the framework would trace all of our known markers of consciousness back to one kind property (K_1), and find that this property coincides with cognitive access. The conclusion that the methodological puzzle step of the framework sanctions is clear: that consciousness coincides with access. Indeed, to such a researcher, the situation in figure 4 would look just like the 'easy' solution to the methodological puzzle represented in figure 1: one where all our known markers of consciousness are underpinned by a natural kind property that correlates with cognitive access. This would of course be wrong because there would be another natural kind property that is also an instance of consciousness, and which can occur without access (K_2). Unfortunately, we couldn't have recognised K_2 as an instance of consciousness, so the natural kinds approach would provide us with no reason to think that K_2 is a case of consciousness.

Let's take a concrete application. We can usefully think of K_1 as the global workspace system that underpins cognitive access. We can think of K_2 as the recurrent processing property in areas V1-V4, which recurrent processing theorists argue underpins consciousness. Suppose (for the sake of the example) that recurrent processing does underpin consciousness, but that we have incorrectly excluded the causal upshots of recurrent processing from the set of markers of consciousness (they are false negatives). Such properties would be like T_8 in figure 4: a case where a property that is causally linked to consciousness has not been recognised as such (a false negative). We have erroneously excluded T_8 from the set of properties causally

linked to consciousness, so we would have no reason to think that the recurrent processing property that underpins it (K_2) is an instance of consciousness. All of our known markers of consciousness coincide on the global workspace system (which correlates with cognitive access), meaning that we will erroneously conclude that consciousness always coincides with cognitive access.[11]

Notice that this problem has nothing to do with causal screening off (Shea 2012). In figure 4, the effects of K_2 are not screened off by the effects of K_1. The causal effects of K_2 (T_8-T_{12}) are fully detectable, but have been incorrectly excluded from the list of properties that are markers of consciousness (they are false negatives).

I do not take the situation in figure 4 to be actual. Rather, it is a hypothetical but empirically realistic scenario. The problem isn't that the framework *might* for all we know deliver the incorrect answer, in this scenario. It's that, in the scenario, following the four steps of the natural kinds framework *will inevitably* lead to the incorrect conclusion. Finding that K_1 underpins our known markers of consciousness is sufficient (according to the framework) for us to infer that that property is an instance of consciousness. Finding that this property coincides with cognitive access is (according to the framework) sufficient to conclude that consciousness and cognitive access overlap. The only way that this incorrect conclusion could be avoided is if there were some way to ascertain that K_2 is also an instance of consciousness. As I have shown, the framework cannot do this.

This is by no means the *only* scenario in which similar problems would lead to the incorrect answer to the methodological puzzle. For example, it is inessential to the problem that there are only *two* kind properties involved. If the known markers of consciousness (T_1-T_7) were underpinned by two natural kind properties (or three, or four…) and the false negatives (T_8-T_{12}) were also underpinned by two natural kind properties (or three, or four…) then the problem would remain. In line with the framework, we would infer that all of the kind properties that causally underpinned (T_1-T_7) were instances of consciousness, and all of

[11] Contrast figures 3-4 with figure 2. Figure 2 represents a case where our initial markers of consciousness are underpinned by two distinct kind properties (K_1 and K_2), both of which are candidates for consciousness. In figures 3-4, our markers of consciousness (T_1-T_7) *all* converge on a single natural kind property (K_1). The problem is not that our initial markers of consciousness (T_1-T_7) are underpinned by multiple natural kind properties, but that there are many *more* properties that are causally underpinned by consciousness (including T_8-T_{12}) but which have not been recognised as underpinned by consciousness (false negatives).

the kind properties that underpinned (T_8-T_{12}) were not, and the rest of the issue would remain as outlined above.

4.4: Comparison with Phillips.

It will be helpful to compare my criticism of the natural kinds framework with Ian Phillips'. Phillips objects to the natural kinds approach by pointing out that there is a lack of widespread agreement about the markers of consciousness (2018, pp.5-6). He notes the difference between subjective and objective measures of consciousness, suggesting that some proposed properties will count as markers of consciousness on one metric, but not on others (p.6). This is a problem, he argues, because which set of markers we take as our starting point may dramatically alter the course of our future investigation.

The problem that I have raised for the natural kinds approach is significantly different from Phillips' in several ways. First, though I certainly agree with Phillips that markers are more problematic than advocates of the approach admit, my objection does not rely simply on *disagreement* amongst consciousness researchers about the markers of consciousness, but on the empirically plausible assumption that there may be a lot of false negatives. Second, whilst Phillips raises the possibility that markers *may* be problematic for the framework, I demonstrate that following the steps of the natural kinds framework *will* lead us to the incorrect answer to the methodological puzzle in certain empirically realistic scenarios. Third, I have shown how this problem cannot be remedied by the marker expansion step, which is a core part of the natural kinds framework.

5: No-report paradigms

The following suggestion may be made. In figures 3-4, the basic problem is that K_2 is an instance of consciousness, but that the natural kinds approach cannot identify it as an instance of consciousness. So, this problem could be resolved with the help of a paradigm that is independently capable of identifying K_2 as an instance of consciousness. In the hypothetical scenario, K_2 would be a case of consciousness that can occur without cognitive access, so a good candidate paradigm would be one designed to test for consciousness without cognitive access. No-report paradigms may be suggested to do the job.

There are two problems. The first is specific to the no-report paradigm. In one example, subjects are shown images of animals, and everyday objects (Cohen *et.al.* 2020). In some trials, masks appear directly

before and after the image, preventing it from being consciously perceived. In the 'conscious' condition, there is a gap of 200ms between the mask, the image, and the second mask, allowing the image to be consciously seen. This 'conscious' version itself had two conditions, one in which subjects reported the identity of the images, and a 'no-report' condition where they did not. Subjects' neural activity was measured using EEG. By isolating the activity that was present in the report condition, and absent in the no-report condition, we can identify the neural correlates of report as opposed to neural correlates of consciousness (the P3b event related potential is presented as a candidate for the neural correlates of report (Cohen *et.al.* 2020, p.10)). This may be suggested as a way to establish consciousness in the absence of cognitive access, and thus a way of positively identifying a property like K_2 as an instance of consciousness.

The trouble is that this paradigm is designed to distinguish *reporting* from consciousness, not to distinguish cognitive access from consciousness (Block 2019). The subjects in the no-report condition will still presumably cognitively access the visible stimuli in some way, and so the no-report condition does not represent a case of successfully distinguishing consciousness of a stimulus from cognitive access to that stimulus.[12]

The more general problem with this suggestion is that it relies on a paradigm that can establish a case of consciousness without cognitive access. But any such paradigm would have solved the methodological puzzle already! It would have supplied an example of the No Access Situation (NAS) with which we started. In order to remedy the problem with the natural kinds framework outlined above, we would have to introduce a paradigm that can simply solve the methodological puzzle on its own. But of course, if a paradigm can achieve this, then there would be no work left for the natural kinds approach to do. This point is not particular to no-report paradigms. *Any* paradigm that could possibly give us independent reason to believe that K_2 is an instance of consciousness, would by definition be a paradigm that can positively identify cases of consciousness that can occur without cognitive access (that is what K_2 is, after all). Any such paradigm render the natural kinds approach otiose.

[12] The authors do not claim that it does (Cohen *et.al.* 2020, p.2). Block (2019) suggests 'no-cognition' paradigms such as Brascamp *et.al.* (2015) as an example of a paradigm that can, but as Phillips and Morales (2020) point out, this does not entirely dissociate consciousness from access.

6: Objections and replies

6.1: Is the scenario empirically realistic?

Objection: the natural kinds framework cannot *always* supply the correct answer, but only in more favourable circumstances. This is good enough.[13]

Reply: it would be unreasonable to expect the framework to deliver the correct answer in extravagant scenarios. However, I take it to be a requirement that the framework deliver the correct answer in empirically realistic scenarios: scientifically respectable possibilities that are not rendered unlikely by our current knowledge. The scenario outlined above rests on two claims: that our set of markers may contain a lot of false negatives, and that consciousness might be associated with multiple natural kinds. Everyone (including advocates of the framework) will accept the possibility of false negatives, and it is rendered plausible by alternative interpretations of paradigms like stem completion and trace/delay conditioning. The multiple kinds claim is rendered realistic on the grounds that many other faculties in cognitive science have been discovered to be like this. At present, we have no more reason to think that consciousness is associated with one natural kind property than to think it is associated with many. For these reasons, the two assumptions at the heart of the scenario are both empirically realistic possibilities, and the scenario itself must be taken seriously. Of course, we do not need to accept that these assumptions are true, only that they are scientifically respectable possibilities. The fact that the framework's core steps would lead us to the incorrect answer in such a scenario is enough to undercut trust in it.

An opponent may reply by saying that we should first rule out the view that consciousness is associated with multiple kinds on independent grounds, which would allow us to reject the problematic case outlined above (figures 3-4). The problem is that the methodological puzzle itself presents an obstacle to this. This suggestion would involve identifying how many causally significant properties underpin conscious experiences. But different answers to the methodological puzzle will give us different answers to where conscious experience is located, and hence where we should look for underlying natural kind properties. For example, if global workspace theory is correct, then we should only be trying to ascertain which natural

[13] Thanks to an anonymous referee.

kind properties underly activations in the global workspace. If recurrent processing theory is true, we should also be looking for natural kind properties that underpin recurrent activations in many other areas of the brain, such as visual cortex. Obviously, because these starting points are so radically different, they are likely to deliver different answers to the question of how many natural kind properties underpin consciousness. For these reasons, we cannot first establish how many natural kind properties underpin consciousness *and then* turn to the methodological puzzle. The puzzle has to come first.

An opponent may resist my argument on the grounds that it is unlikely that there are *as many* false negatives as correct markers of consciousness.[14] It is unclear how we might decide how likely it is that our paradigms contain as many false negatives as known markers. However, we can sidestep these issues here because it is not essential to the scenario. In figures 3-4, I have depicted a similar number of false negatives (T_8-T_{12}) as known markers of consciousness (T_1-T_7), for simplicity. However, imagine a scenario just like figure 4, except that there were hundreds of known markers of consciousness, underpinned by K_1, and only two false negatives, underpinned by K_2. The same problem would arise: we would infer (based on the fact that K_1 underpins our known markers of consciousness) that K_1 was an instance of consciousness. However, because the false negatives have been incorrectly classified as not underpinned by consciousness, we would conclude that the property underpinning them (K_2) was not an instance of consciousness, when in fact it is. The rest of the reasoning from the scenario outlined would then apply in the same way, leading to the incorrect answer to the methodological puzzle. The issue is not solved by supposing fewer false negatives.

6.2: Taking all the markers

Objection: the false negatives problem arises because we may have excluded some properties that are causally underpinned by consciousness from the set of markers of consciousness (these are false negatives). So our default assumption should be inclusion. We should err on the side of including *more* markers in the set. For example, we should assume that eye blinks in *both* delay and trace conditioning cases are causally underpinned by consciousness.

[14] Thanks to an anonymous referee for suggesting this.

Reply: this risks assuming that some properties are causally underpinned by consciousness when they actually are not (false positives). Suppose we collect together our large set of markers (T_1-T_{12}), but some of them (say, T_7-T_{12}) are not causally underpinned by consciousness, whereas we have assumed that they are underpinned by consciousness (false positives). Then, suppose our causal modelling reveals that T_1-T_6 are underpinned by the global workspace system underpinning cognitive access, whilst T_7-T_{12} are underpinned by a separate kind property, which can occur without cognitive access. We would then erroneously conclude that this separate kind property is an instance of consciousness, based on the fact that it is the property that underpins many of our initial markers of consciousness. We would then infer that consciousness without cognitive access occurs (that NAS happens). This would of course be incorrect, because in this scenario, T_7-T_{12} are not really causally underpinned by consciousness (the property that underpins them is not actually an instance of consciousness).

7. Overflow/mesh and natural kinds

Readers may wonder about the connections between this discussion, and Block's own inference to the best explanation arguments concerning the methodological puzzle.[15] I close by highlighting some points of contact. This will serve to reinforce one conclusion of this paper: that deciding between alternative markers of consciousness is a more problematic step than has been realised.

Block's argument can be broken into two smaller arguments. The first is the overflow argument, which uses partial report paradigms (Sperling 1960, Landman *et.al.* 2003, Bronfman *et.al.* 2014). Subjects are briefly presented with an array of letters (sometimes shapes). After the offset of the letters, a visual or auditory cue instructs subjects to recall a particular row of letters (or particular shape). Subjects do this reliably, but cannot recall the entire array. The explanation is that there is a pre-workspace visual memory system (known variously as iconic memory, or fragile visual short-term memory (Sligte *et.al.* 2008, 2009)) in which information about the entire array is stored. This system has a higher capacity than the global workspace system that underpins cognitive access. Because information about the whole array is stored in this pre-workspace system, any one row can be retrieved by attentional amplification and reported; but the whole

[15] Thanks to an anonymous referee for pushing me on this.

array cannot, because of the capacity limitations of the workspace. Block argues that the representations stored in this pre-workspace memory system are conscious, including the uncued rows (2007, 2014, 2019). Since the uncued rows are not accessed, consciousness overflows cognitive access. So goes the first of Block's arguments.

The second is the 'mesh' argument. He refers to neuroscientific data showing that coalitions of neurons located in visual cortex 'compete' with each other. The winners get attentional amplification that results in their information being accessed by prefrontal areas associated with the workspace. The losers do not (2007, pp.496-498). He says:

> If we assume that the strong but still losing coalitions at the back of the head are the neural basis of phenomenal states (so long as they involve recurrent activity) then we have a neural mechanism that explains why phenomenology has a higher capacity than the global workspace' (2007, p.498).

My concern is to compare Block's arguments to the natural kinds framework. The overflow argument is structurally similar to the marker assembly step. It involves identifying a property of consciousness that calls for underlying explanation. In our terminology, the overflow property can be considered one of the *markers* of consciousness. Block's second argument is similar to the step in which you identify a property that causally underpins the markers. In Block's case, a property is identified (neural coalitions in visual cortex), which explains the marker property (that consciousness overflows access). The similarities to the natural kinds framework are clear.[16]

Almost all of the debate has focussed around the overflow argument. Critics argue that subjects' conscious experience of the whole array lacks the detail to identify any one letter. They claim that attentional amplification (triggered by the cue) leads to some objects being raised from the pre-workspace system to the level of conscious detail required for report (Phillips 2018, Cohen and Dennett 2011). Therefore (they claim) there is no reason to think that consciousness overflows the workspace. This is interesting for our purposes. Advocates of the natural kinds framework take the marker assembly step to be only the first step on the path to resolution of the methodological puzzle. I have argued that this step is problematic, because

[16] There are some dissimilarities. Block uses inference to the best explanation (not causal modelling) to establish that neural coalitions are the basis of phenomenology. Furthermore, Block's argument does not require that neural coalitions be a *natural kind* property in the sense of supporting scientific projection and induction.

false negatives have the potential to lead us directly to the incorrect answer to the puzzle. The controversy over the overflow argument reveals another way that markers problematic. The locus of this debate has concerned the overflow argument, which is similar to the marker assembly step. What this shows is that the first step of the natural kinds framework (the marker assembly step) is not just a preliminary step to be gotten over in pursuit of a solution to the puzzle. It is the sticking point of the entire debate.

The point is: markers are not so simple. To look for a marker of consciousness to make decisions about which properties are causal expressions of consciousness, which need to be explained by an underlying causal property. But to make these decisions is far from theory-neutral. In the case of the overflow argument, they are what the entire debate is about. In the argument of this paper, our choice over which markers to use can make the difference between a correct and an incorrect answer to the puzzle. Markers are not the first step to resolving the puzzle. They are the puzzle.[17]

[17] Thanks to Tim Bayne, Ian Phillips, Nick Shea, the editor and two anonymous referees for challenging but sympathetic comments on previous drafts. Thanks to Ned Block, Alexandria Boyle, Elizabeth Irvine, Bob Kentridge, Maja Spener and Cecily Whiteley for valuable discussion.

References

Bayne, T. 2018. On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness.* 4 (niy007): 1-8

Bayne, T. and Shea, N. forthcoming. Consciousness, concepts and natural kinds. *Philosophical Topics.*

Block, N. 2007. Consciousness, accessibility and the mesh between psychology and neuroscience. *Behavioural and Brain Sciences.* 30: 481-548.

Block, N. 2014. Rich conscious perception outside focal attention. *Trends in Cognitive Science.* 18: 445-447

Block, N. 2019. What is wrong with the no report paradigm and how to fix it. *Trends in Cognitive Sciences.* 23: 1003-1013.

Boyd, R. 1989. What realism implies and what it does not. *Dialectica.* 43: 5-29.

Boyd, R. 1999. Kinds, complexity and multiple realization: comments on Millikan's "historical kinds and the special sciences". *Philosophical Studies*, 95: 67-98.

Brascamp, J., Blake, R., Knapen, T. 2015. Negligible fronto-parietal BOLD accompanying unreportable switches in bistable perception. *Nature Neuroscience.* 18: 1672-1681.

Bronfman, Z., Brezis, N., Jacobsen, H. and Usher, M. 2014. We see more than we can report: 'cost free' colour phenomenology outside focal attention. *Psychological Science.* 25: 1394-1403.

Clark, R. and Squire, L. 1998. Classical conditioning and brain systems: the role of awareness. *Science.* 280: 77-81.

Cohen, M. and Dennett, D. 2011. Consciousness cannot be separated from function. *Trends in Cognitive Sciences.* 15: 358-364.

Cohen, M., Ortego, K., Kyroudis, A. and Pitts, M. 2020. Distinguishing the neural correlates of perceptual awareness and post-perceptual processing. *The Journal of Neuroscience.* 40: 4925-4935.

Craver, C. 2009. Mechanisms and natural kinds. *Philosophical Psychology.* 22: 575-594

Debner, J. and Jacoby, L. 1994. Unconscious perception: attention, awareness and control. *Journal of Experimental Psychology.* 20: 304-317.

Dehaene, S., Changeux, J., Naccache, L., Sackur, J., Sergent, C. 2006. Conscious, preconscious, subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences.* 10: 204-211.

Denison, R., Block, N. and Samaha, J. 2020. What do models of visual perception tell us about visual phenomenology? In F. de Brigard & W. Sinnot-Armstrong (Eds.), Neuroscience and philosophy. (Cambridge, MA: MIT Press).

Ereshefsky, M. 2010. What's wrong with the new biological essentialism. *Philosophy of Science.* 77: 674-685.

Ereshefsky, M. and Matthen, M. 2005. Taxonomy, polymorphism, and history. *Philosophy of Science.* 72: 1-21.

Griffiths, P. 1997. *What Emotions Really Are.* (Chicago: Chicago University Press).

Haueis, P. forthcoming. A generalized patchwork approach to scientific concepts. *The British Journal for the Philosophy of Science.*

Irvine, E. *Consciousness as a Scientific Concept.* (Dorecht: Springer).

Jack, A. and Shallice, T. 2001. Introspective physicalism as an approach to the science of consciousness. *Cognition.* 79: 161-196.

Kim, J. 1992. Multiple realization and the metaphysics of reduction. *Philosophy and Phenomenological Research.* 52: 1-26.

Kornblith, H. 1993. *Inductive Inference and its Natural Ground.* (USA: MIT Press).

Khalidi, M. 2013. *Natural Categories and Human Kinds.* (Cambridge: CUP).

Lamme, V. 2004. Separate neural definitions of visual consciousness and visual attention; a case for phenomenal awareness. *Neural Networks.* 17: 861-872.

Lamme, V. 2006. Towards a true neural stance on consciousness. *Trends in Cognitive Sciences.* 10: 494-501.

Landman, R., Spekpeijse, H. and Lamme, V. 2003. Large capacity storage of integrated objects before change blindness. *Vision Research.* 43: 149-64

Lovibond, P. and Shanks, D. 2002. The role of awareness in Pavlovian conditioning: empirical evidence and theoretical implications. *Journal of Experimental Psychology.* 28: 3-26.

Machery, E. 2009. *Doing Without Concepts.* (New York: OUP).

Michel, M. 2019. Consciousness science underdetermined. *Ergo.* 6: 10.3998/ergo.12405314.0006.028

Millikan, R. 2000. Historical kinds and the "special sciences." *Philosophical Studies.* 95: 45-65.

Peruchett, P. 1985. A pitfall for the expectancy theory of human eyelid conditioning. *Pav. J. Biol. Sci.* 20: 163-170.

Peruchett, P., Cleermans, A., Destrebecqz, A. 2006. Dissociating the effects of automatic activation and explicit expectancy on reaction times in a simple associative learning task. *Journal of Experimental Psychology.* 32: 1-11

Phillips, I. 2018. The methodological puzzle of phenomenal consciousness. *Phil. Trans. R. Soc. B.* 373: 20170347.

Phillips, I. and Morales, J. 2020. The fundamental problem with no cognition paradigms. *Trends in Cognitive Sciences.* DOI: 10.1098/rstb.2017.0347

Prinz, J. 2012. *The Conscious Brain.* (New York: OUP).

Sandberg, K. and Overgaard, M. 2015. Using the perceptual awareness scale. In M. Overgaard (ed.) *Behavioral Methods in Consciousness Research.* (OUP).

Sergent, C., Baillet, S. and Dehaene, S. (2005) Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience.* 8: 1391-1400

Shea, N. 2012. Methodological encounters with the phenomenal kind. *Philosophy and Phenomenological Research.* 84: 307-344.

Shea, N and Bayne, T. 2010. The vegetative state and the science of consciousness. *The British Journal for the Philosophy of Science.* 61: 459-484.

Skora, L., Yeomans, M., Crombag, H. and Scott, R. Evidence that instrumental conditioning requires consciousness in humans. *Cognition.* 208: 104546

Sligte, I., Scholte, H., Lamme, V. 2008. Are there multiple short term memory stores? *PLoS ONE.* 3 e1699: 1-9.

Sligte, I., Scholte, S. and Lamme, V. 2009. V4 activity predicts the strength of visual short-term memory representations. 24: 7432-7438

Snodgrass, M. 2002. Disambiguating conscious and unconscious influences: do exclusion paradigms demonstrate unconscious perception? *American Journal of Psychology.* 115: 545-579.

Sperling, G. 1960. The information available in brief visual presentations. *Psychological Monographs.* 74: 1-29.

Taylor, H. and Vickers, P. 2017. Conceptual fragmentation and the rise of eliminativism. *European Journal for the Philosophy of Science.* 7: 17-40.

Taylor, H. 2019. Fuzziness in the mind: can perception be unconscious? *Philosophy and Phenomenological Research.* doi: 10.1111/phpr.12592

Taylor, H. 2020. Emotions, concepts and the indeterminacy of natural kinds. *Synthese.* 197: 2703-2093.

Visser, T. and Werikle, P. 1999. Conscious and unconscious process: The effects of motivation. *Consciousness and Cognition.* 8: 94-113.

Weidemann, G. and Lovibond, P. 2016. The role of US recency in the Perruchet effect in eyeblink conditioning. *Biological Psychology.* 119: 1-10.