PhiMiSci
Philosophy and the Mind Sciences

Finding the neural correlates of consciousness will not solve all our problems

Morten S. Overgaard[a,b,c] (morten.storm.overgaard@cfin.au.dk)
Asger Kirkeby-Hinrup[d] (asger.kirkeby-hinrup@fil.lu.se)

Abstract
Subjective experience has often taken center stage in debates between competing conceptual theories of the mind. This is also a central object of concern in the empirical domain, and especially in the search for the neural correlates of consciousness (NCCs). By now, most of the competing conceptual theories of consciousness have become aligned with distinct hypotheses about the NCCs. These hypotheses are usually distinguished by reference to their proposed location of the NCCs. This difference in hypothesized location of the NCCs has tempted participants in these debates to infer that evidence indicating the location of the NCCs in one or the other brain region can be taken as direct evidence for or against a given conceptual theory of consciousness. We argue that this is an overestimation of the work finding the NCCs can do for us, and that there are principled reasons to resist this kind of inference. To show this we point out the lack of both an isomorphism and a homomorphism between the conceptual frameworks in which most theories are cached, and the kind of data we can get from neuroimaging. The upshot is that neural activation profiles are insufficient to distinguish between competing theories in the conceptual domain. We suggest two ways to go about ameliorating this issue.

Keywords
Behavioural methods · Correlations · Homomorphism · Isomorphism · Localization· Neural correlates of consciousness · Theories of consciousness

This article is part of a special issue on "The Neural Correlates of Consciousness," edited by Sascha Benjamin Fink and Ying-Tung Lin.

[a]Professor, Head of Cognitive Neuroscience Research Unit (CNRU), Aarhus University
[b]Department of Clinical Medicine, Center of Functionally Integrative Neuroscience (CFIN), Aarhus University
[c]Aarhus Institute of Advanced Studies, Aarhus University, Denmark
[d]Department of Philosophy and Cognitive Science, Theoretical Philosophy, Lund University, Sweden

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5.
https://doi.org/10.33735/phimisci.2021.37

1 Introduction

By now, most researchers concerned with the study of consciousness believe that the mind can be naturalized, for instance in the sense that consciousness depends on activity in the brain. For this reason, the project of identifying the Neural Correlates of Consciousness (NCCs) has attracted a lot of attention and precipitated interdisciplinary interactions to clarify the mesh between competing conceptual theories of consciousness, particularly (but not exclusively) from the domain of philosophy[1], and findings from the empirical domain.

Subjective experience—the peculiar first-person aspects of being conscious, the nature of which often has taken centre stage in debates between competing conceptual theories of the mind—is also a central object of concern in the empirical domain, and especially in the search for the NCCs. One upshot of this shared explanandum is that most of the competing conceptual theories of consciousness have become aligned with distinct hypotheses about the NCCs. For instance, many proponents of higher-order thought (HOT) theories of consciousness have embraced the hypothesis that the NCCs are located in the prefrontal cortex (PFC). Furthermore, proponents of HOT theories argue that conscious access and phenomenal consciousness may coincide (a sentiment shared by proponents of other theories, see e.g. Naccache, 2018; see also Overgaard, 2018). A main competitor to the HOT theories, the so-called recurrency theory, argue that the PFC is required only for conscious access, whereas the NCCs for phenomenal consciousness rely on recurrent activations in early sensory areas, such as the primary visual cortex (V1) in the case of visual phenomenology (Lamme, 2018; Pinto et al., 2017). This difference in views has sparked extensive debate (importantly, this debate is not limited to proponents of HOT and recurrency theories, see e.g. Peters et al., 2017; Railo et al., 2015). Furthermore, the difference in hypothesized location of the NCCs has tempted participants in these debates to infer that evidence indicating the location of the NCCs in one or the other brain region can be taken as direct evidence for or against a given conceptual theory of consciousness.

Arguing in this way seems to be an overestimation of the work the NCCs can do for us, when it comes to theories in the conceptual domain, and we will argue that there are principled reasons to resist these kinds of argument. Importantly, and to be clear, finding the NCCs almost certainly will allow us to make significant headway in our understanding of consciousness, and how it may arise from the brain. This is not the object of our concern. Instead, what we highlight is that

[1] Throughout, we use 'conceptual theories' and 'the conceptual domain' broadly to refer to the various conceptual frameworks aiming to explain consciousness, such as higher-order thought theories, reflexive theories, panpsychic theories, workspace theories, to name a few. Importantly, while many existing conceptual theories are now closely tied to empirical work, almost all contain theory-crafting in their deployment of concepts (e.g. 'mental state') and posits (e.g. 'recurrent processing are necessary for phenomenal consciousness') that are not essentially empirical concepts (as opposed to e.g. 'spiketrains' or the 'prefrontal cortex'). It is the relation between these concepts and posits, on the one hand, and empirical data on the other hand, that is our main concern here.

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5.
https://doi.org/10.33735/phimisci.2021.37

there are reasons to doubt that finding the NCCs—in and of itself—may resolve the conceptual debates by supporting the empirically plausibility of one conceptual theory over the other. This problem turns on two issues. The first issue concerns a missing isomorphism—the lack of a 1-to-1 correspondence—between theoretical posits and neural data.[2] The second concerns the lack of homomorphism—the lack of a structure-preserving mapping—between conceptual frameworks deployed in the theoretical and empirical domain respectively.

As we will argue, the consequence of the missing isomorphism is that findings or predictions about activity localized to specific brain regions, by themselves, are insufficient to distinguish between competing theories in the conceptual domain. To boot, the consequence of the missing homomorphism is that the concepts deployed in the conceptual debates do not map straightforwardly onto the kind of data generated by the brain sciences, and *vice versa*.[3] Therefore, significant liberties are often taken when mapping an empirical finding onto a conceptual theory. It is worth noting that the way we operationalize our conception of the NCCs may

[2]The notion of isomorphism we deploy throughout the paper is similar to the one found in Pessoa et al. (1998), which they call *analytic isomorphism*. Pessoa and colleagues conceive of analytic isomorphism as "essentially a conceptual or methodologic doctrine about the proper form of explanation in cognitive neuroscience." (Pessoa et al., 1998, p. 726). However, while Pessoa and colleagues seek to apply the isomorphism relation to neural data and phenomenal experience, we instead (for reasons related to correlation explained in the next section) apply the isomorphism relation to neural data and theoretical posits. One reviewer suggested isomorphic relations cannot hold between entities of different cardinality since this would be a category mistake. We do not think this is the case and note that application of the isomorphism relation to entities of different cardinality is rife within all areas of the scientific literature. For instance, isomorphism is discussed as a relation between thought and reality (Glock, 1997) and in models in the philosophy of science (Da Costa & French, 1990).

[3]Clearly, some theories from the conceptual domain are better equipped to cater to a conceptual mapping between their theoretical posits and the kind of data generated by the brain sciences. To some extent, it seems that theories developed at the intersection between philosophy and the empirical sciences, fare somewhat better in this regard. However, we do not put much weight on this observation, and nothing hinges on it, since we believe what we have to say is pertinent to most if not all theories. Thus, our use of HOT theory and recurrent processing theory as examples in this paper does not mean that these are the only conceptual frameworks that have issues with the conceptual mapping task (see also footnote 1). A secondary question pertains to whether the mapping issues we address in this paper find their root cause in the conceptual frameworks themselves or in the conceptual background assumptions of the varying proponents of the competing conceptual frameworks. While our concern here is mainly about the conceptual frameworks themselves, it seems obvious that there is an interplay between these two factors. Certainly, the (pre-theoretic) predilections and conceptual commitments of researchers are a key factor in the development of theories. Consequently, the interpretations of the concepts deployed by a given theory influences the conceptual mapping to the empirical domain. Additionally, we acknowledge that much debate in the theoretical domain is driven by longstanding disagreements about what the central concepts actually mean ("consciousness" is a clear cut example), and there is an overhanging risk (and some evidence) that these disagreements bleed into the work in the empirical domain. We see this as all the more reason to meticulously and critically scrutinize the practices involved in conceptual mapping, which is exactly our objective with this paper. We thank an anonymous reviewer to prompting us to clarify this.

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5.
https://doi.org/10.33735/phimisci.2021.37

compound (or, in the best case, alleviate) these issues (see e.g. discussion in Fink, 2016). Be that as it may, we here advertise two ways to go about ameliorating these issues. Before we move on, it is necessary to briefly summarize the foundation for our argument, starting with a reminder about the nature of empirical investigations of consciousness.

2 Correlations, subjective and objective measures

The most fundamental distinction in empirical methods applied in consciousness research is between subjective and objective measures (Koch et al., 2016; Timmermans & Cleeremans, 2015). Subjective methods tap into the assumed privileged access an individual has to her own conscious states. Usually, subjective methods rely on subjects to introspectively determine what they are consciously experiencing. Objective methods, on the other hand, rely on behavioural data (e.g. eye movements, reaction times) or measurements of events in the brain (e.g. blood oxygen level, event related potentials, spike trains). Strictly speaking, objective methods are not measuring *consciousness per se*. Rather, objective measures concern behaviour or neural events that we think are somehow related to the presence or absence of conscious states. So, our objective data is not *about* consciousness *per se*. Subjective measures, conversely, are (at least *prima facie*) *about* conscious states. However, subjective measures suffer from a range of other issues. First and foremost, subjective methods do not sit well with standard conceptions of scientific rigor such as third person access. In contrast to this, objective methods can cater to this requirement. Thus, it seems a combination of subjective and objective methods may be the best solution to satisfy, one the one hand, the desideratum that our data are *about* consciousness, and on the other hand, the desideratum that our data are third person accessible, i.e., can be verified by an external observer. Still, it is worth remembering that, even in a best case scenario, neither consciousness nor the neural processes presumably underpinning it actually are what we *observe*. The aim is to establish a *correlation* between subjective measures (that we assume reflect consciousness more or less accurately), and measurements of neural (or behavioural) events (that we assume are related to the presence or absence of conscious states by proxy of the neural states or processes that underpin them). Importantly, the ways we establish these correlations themselves depend on the assumption that other correlations have been reliably established (e.g. that activity in LGN and V1 is correlated with the presentation of a visual stimulus). This means that reaching our objective consists in establishing a correlation between two distinct correlations. On the one hand, we assume that our neural or behavioural data correlate with the NCCs. On the other hand, we assume that our subjective measures correlate with consciousness. The result, then, is a correlation between our neural/behavioural data and our data from our subjective measures (Overgaard, 2015).

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5.
https://doi.org/10.33735/phimisci.2021.37

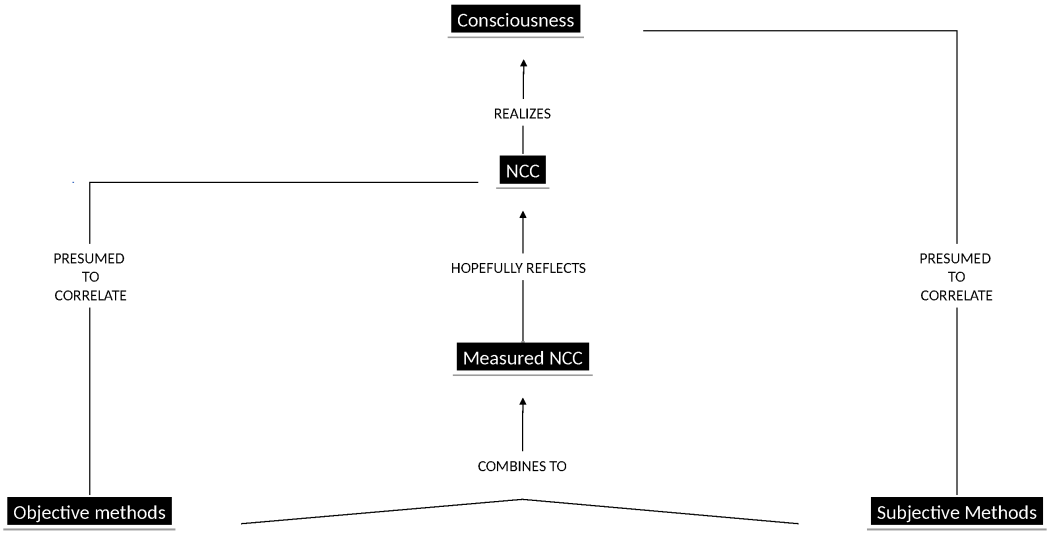

Figure 1: A schematic view of the role of correlations in the search for the NCCs

Suppose we establish the desired correlation between subjective and objective measures, how will this impact competing theories at the conceptual level? Theories in the conceptual domain predict more or less the same with respect to subjective measures and are mute on objective measures. Objective measures, by virtue of being nothing more than visual or mathematical representations of measurements, reciprocally are mute on conceptual theories. Finally, to the extent that subjective measures have anything to say about conceptual theories, it is by appeal to how things appear to a single individual through introspection. Such introspective reports, as noted above, have trouble with third person verification, and, somewhat worse, appear arbitrary with respect to the conclusions individuals take them to warrant with respect to conceptual theories (something that should be obvious from the last 50 or so years of philosophical debate on consciousness) (Overgaard, 2017).

3 The missing isomorphism

The idea that—when identified—the neural correlates of consciousness (NCCs) will be sensitive to the nuances of competing conceptual theories, in a way that allows us to distinguish between them, and confirm one over the others, presupposes an isomorphism between the theoretical posits and our empirical data about the NCCs. That two entities are isomorphic means that there is a 1:1 correspondence between the two entities with respect to the aspect(s) under consideration. The kind of isomorphism relevant in the current context is supposed to hold between posits of conceptual theories of consciousness and our (eventual) evidence for the NCCs. As mentioned above, the work discovering the NCCs is supposed to do for

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5.
https://doi.org/10.33735/phimisci.2021.37

us with respect to the debate between competing theories requires it to be *sensitive to the nuances* of competing theories. Effectively, this means that, in order for the NCCs to be sensitive to the nuances of distinct theories, these nuances need to show up unequivocally in the conceptual mapping. Isomorphism between a conceptual framework and the kind of evidence we can get for the NCCs would constitute the strongest possible mapping. Anything short of isomorphism will leave significant conceptual wiggle room and likely fail to settle the debate. To boot, issues stemming from conceptual wiggle room, are among the problems that finding the NCCs is supposed to solve for us.

As mentioned above, one prominent debate pitches higher-order thought (HOT) theories of consciousness against recurrency theory. The two theories are now closely aligned with two distinct hypotheses about the neural correlates of consciousness (NCC). On the one hand, HOT theories are aligned with the idea that consciousness is generated by *late* processes, tentatively attributed to the prefrontal cortex (PFC). On the other hand, recurrency theories have aligned themselves with the hypothesis that consciousness depends on *early* processes, for instance recurrent processing in the primary visual cortex (V1) in the case of conscious visual awareness.

However, there is no isomorphism supporting evidential transitivity between activations in this or that brain region and one or another theory. For example, higher-order theories hypothesize that consciousness generating higher-order thoughts may involve prefrontal activity, but this does not entail that findings where conscious subjects do not exhibit significant prefrontal activations is evidence against the higher-order theories as conceived of in the conceptual domain. Nevertheless, many seem to engage with this kind of argument (Kozuch, 2014; Odegaard et al., 2017; Sebastián, 2014). In one recent example of this kind of endeavour, Michel and Morales, in a nice paper (2020), provide extensive arguments to the effect that in many studies prefrontal activations correlate with consciousness, as opposed to reports. According to Michel and Morales, if this can be shown to be the case, it would support so-called prefrontal theories against their competitors. Importantly, (at least in the present context) we are not contending the content of this paper. What we are sceptical about is the premise of the paper that links certain theories to certain regions of the brain. Monikers such as 'prefrontal theories' referring to a group of theories in the conceptual domain (e.g. higher-order thought theory, global workspace and others) and 'local recurrency theory' generally used to refer to the first-order reflexive theory associated with Ned Block, tacitly indicate that somehow the viability of these theories from the conceptual domain depends on where in the brain an eventual discovery locates the NCCs. It is important to remember that even if the NCCs are eventually located in the early sensory regions, this does not entail the falsification of HOT theory. Similarly, the reflexive theories are not automatically wrong, in case the NCCs happen to be in the PFC. It is exactly the implicit assumption (as found in e.g. Michel and Morales' paper) that evidence of the NCCs being in one brain region or another automatically supports this or that theory in the conceptual domain of which we are sceptical.

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5.
https://doi.org/10.33735/phimisci.2021.37

4 The missing homomorphism

In addition to the missing isomorphism between brain regions and conceptual theories, there are a plethora of technical terms from the conceptual debates that have no obvious and/or measurable empirical counterpart. Prominent examples are the notions of phenomenality and subjectivity. These notions are prevalent in theoretical debates, but we have no good idea of how to cash them out in measurements amenable to third person scrutiny. Thus, in addition to the missing isomorphism presenting a problem on the general level when connecting theories with brain regions (a significant problem in itself) there is a lack of homomorphism between much of the conceptual and empirical terminology and overall frameworks. For instance, a central theoretical construct scaffolding both HOT and reflexive theories is *mental states*. To illustrate, the HOT theory posits that a mental state, such as a sensation, is conscious when it is the intentional object of another (higher-order) mental state. Similarly, reflexive theories hold that a mental state has the property of being conscious by instantiating a special reflexive relation to itself. Both theories are similar in the sense that they both seek to explain consciousness by reference to properties of—or relations between—mental states. However, when it comes to the brain, the notion of mental state does not straightforwardly apply. It is true that neuroscientists may deploy the concept of a mental state, for instance when referring to representations in the brain incurred by the introduction of a stimulus. However, this usage is merely a proxy referring to a collection of signals and processes propagating through different brain areas. So, while the notion of mental states, conceived of as ontologically isolatable entities is conversationally useful and theoretically harmless, it is—strictly speaking—empirically mistaken. On the neural level states are *signals* or patterns of neural activation, and this is what we measure empirically. Signals move through brain regions—and patterns unfold—over time. We can trace a signal and its interactions as it propagates through the brain and reaches different stages of processing. And speaking loosely, we may agree that the signal somehow corresponds to, say, a perceptual representation of a visual stimulus. Such loose talk is useful, and for most purposes more than sufficient. But, in virtue of being a general gesture toward an underspecified group of phenomena, loose speak fails to provide a specific answer to what and where the mental state actually is. And this is what matters when we need to individuate a mental state. There is no non-arbitrary way to fixate upon an exact point in time and space to delineate the boundaries of a mental state, which is what is necessary for individuation. To see why this is important, suppose the signal from a visual stimulus arrives at V1, after which the signal propagates through the visual system as normal, and makes its way around the frontoparietal network, and eventually the subject reports being conscious of the stimulus. According to reflexive theories, the signal, through processing, acquired some reflexive relationship to itself (supposedly from recurrent processing). According to HOT theory, the signal arrived at some location in the PFC (supposedly the dorso-

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5.
https://doi.org/10.33735/phimisci.2021.37

lateral prefrontal cortex; dlPFC), at which point it got represented by a HOT. So, in the first case, we have one mental state (with fuzzy boundaries), the processing of which yields consciousness. Whereas in the latter case we have a mental state (with fuzzy boundaries) arriving at some location and triggering another mental state (the HOT). But how do we determine from the neural data, whether there are one or two states? In light of the debate between competing theories, it matters a great deal, whether we are talking about one or two states. There are, of course, plenty of timeslices we can pick out and lines we can draw around the signal. For instance, one could argue that a particular development in the signal, such as a change in distribution, route, amplitude, or events such as arrival at specific cortical areas, changes in encoding, transfer to different kinds of memory-caches and so, should form the basis for the individuation. However, such a decision is ultimately arbitrary. The upshot of this is that there is no non-arbitrary structure preserving mapping, (i.e., homomorphism) between many theoretical frameworks and empirical data.

5 Activation profiles are insufficient to distinguish between theories

As noted above, most of the competing conceptual theories of consciousness have become aligned with distinct hypotheses about the NCCs. Essentially, what distinguishes one hypothesis about the NCCs from another is the location and the nature of neural processes it posits as *necessary* and/or *sufficient* for consciousness to occur. In particular, the localization of the NCCs has been the topic of much debate. Probably, the hypothesis that *early* processes in occipital cortex constitute the NCCs is sufficiently distinct from the hypothesis that the NCCs are to be found in *late* processes in the front of the brain, to allow us to distinguish *empirically* between the two. However, the possibility of distinguishing empirically between competing hypotheses of the NCCs is not the object of concern here. What we doubt is whether the empirical confirmation of any particular NCC hypothesis allows the confirmation of any particular conceptual theory (e.g. HOT) over another. To understand this worry, it is informative to reflect on how conceptual theories have come to be aligned with respective NCC hypotheses.

Deploying HOT theory as an example, the consensus that the PFC is the brain region most likely to underpin higher-order thoughts is derived from perceived similarities between how HOTs are conceived of in the conceptual domain and a range of metacognitive abilities which evidence suggests depend on prefrontal processes (Beeckmans, 2007; Brown, 2012; Kozuch, 2014; Kriegel, 2007; Lau, 2007; Lau & Brown, 2019; Lau & Rosenthal, 2011). For instance, some theoretical posits of HOT theory are conceptually or functionally similar to metacognitive abilities that are thought to be located in the PFC, such as self-monitoring and judgements of (own) performance (Fleming & Dolan, 2012), cognitive control and planning

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5. https://doi.org/10.33735/phimisci.2021.37

(Lau, 2011) and theory of mind (Frith & Frith, 2006). However, we came to associate each of these metacognitive abilities with the PFC because we have empirical reasons—from *behavioural* measures—to think they depend on PFC activity (see e.g. Kirkeby-Hinrup, 2020, p. 142). In contrast to this, the putative reasons to think that the viability of e.g. HOT theory depends on whether the NCCs are found in the PFC, is extrapolated from data obtained from investigations *not of consciousness, but of other cognitive phenomena.*

6 Facing up to the missing isomorphism

Our first advice to ameliorate the troubles incurred by the missing isomorphism concerns how to establish a better mesh between our conceptual theories and the kind of data we can collect about the brain. The physical and functional nature of the brain constrains the kinds of data we can obtain from it. We currently have a decent preliminary grasp of these kinds of data. Even if new neuroscientific methods become available, we would not expect these to fundamentally solve the conceptual divide between theory and data. And (barring the unlikely discovery of non-physical causally efficacious phenomena) the data we can obtain about the brain will be constrained by physics and will reflect the physical nature of the processes and mechanisms in the brain. This means that the empirical data we collect, while possibly crude compared to future technologies, will retain certain shared fundamental characteristics with data collected in the future. Given the constraints on the nature of the empirical data we can collect, it does not appear we can remedy the missing isomorphism through changes in the data we treat or the way we collect it. This is because the missing isomorphism, essentially, is a conceptual disconnect between the way we conceive of our theories and the nature of the data about neural instantiations supposed to exemplify or instantiate the theoretical posits.[4] Therefore, if we cannot bridge this disconnect by altering the data, it seems the only option is to revise the conceptual frameworks in which theories are cached, i.e., bottom up data driven theoretical revision based on the kind of data we can collect (a similar sentiment can be gleaned from Genon et al., 2018). In short: because we cannot make the data fit our theories, we should make our theories fit the data. This of course does not mean throwing out the baby with the bathwater such as a wholesale discarding of theories or giving up on the central explanandum (consciousness). Rather, it means reformulating or updating theories to be sensitive to the kind of data we can collect, and—importantly—be amenable to testing by behavioural methods, as discussed in the next section.[5]

[4]We appreciate that there are historical parallels to this problem, for instance in the philosophical literature on identity theory. Given the prevalence of this kind of argument we see value in restating the problem and applying it in the context of contemporary debates. We thank an anonymous reviewer for reminding us of this.

[5]To elaborate, the goal here is neither to engage in conceptual revision of the kind proposed by eliminativists (Churchland, 1981), nor do we encourage putting our hopes in the discovery of

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5.
https://doi.org/10.33735/phimisci.2021.37

To illustrate these issues, we offer the following examples. The first example concerns localization. In relation to the discussion of higher order theories above, some versions of HOT (Lau & Rosenthal, 2011) posit that specific neural regions are necessary for conscious experience. Particularly, the dorsolateral Prefrontal Cortex (dlPFC) has received a lot of attention. However, the proposal is not merely underdetermined by empirical evidence, it is challenged by seemingly obvious counterevidence. For instance, during REM sleep, prefrontal cortex activity is often low compared to awake states, even though the experienced content may be just as vivid (for some discussion see e.g. Sebastián, 2014; Weisberg, 2014). Similarly, patients undergoing generalized seizures have increased blood flow in prefrontal areas, even though they are considered unconscious.

An energetic debate as of recent stems from the advent of so-called no-report paradigms (e.g. Block, 2019). This debate concerns attempts—without the use of reports—to contrast conditions in which the subject is (presumed to be) conscious of a stimulus with conditions where the subject is (presumed to be) unconscious of the stimulus. A range of issues have been raised with this methodology. One of the more significant issues, is that, in virtue of the lack of reports, no-report paradigms have no methodological control for whether a stimulus was in fact consciously seen (Overgaard & Fazekas, 2016). Nevertheless, *prima facie*, they appear to provide strong evidence that the frontal activations found in many experiments in fact are correlates of reporting, task preparation and execution rather than the conscious experience *per se*.

To illustrate that issues of this kind are not exclusive to higher-order theory, let us consider another example. The main opponents of higher-order theory generally endorse the hypothesis that consciousness is related to early activity in temporo-parietal-occipital networks, rather than late activity in the fronto-parietal networks. One obvious issue with so-called *early* theories is that the varying theoretical proposals suggesting a temporo-parietal-occipital network—while not necessarily mutually exclusive—are not particularly compatible. To boot, it is very unclear that even if we succeeded in locating the NCCs in *early* temporo-parietal-occipital processes, it would be possible to distinguish further between the competing theories available. To illustrate, Victor Lamme argues that all brain regions are unconscious during a feed-forward sweep of information, whereas conscious experience happens when information feeds back to occipital regions (Lamme, 2006). Other views, e.g. the Visual Awareness Negativity hypothesis, makes no

an adequate vocabulary sometime in the future as Nagel (1974) ponders. What we suggest is for theories to be reformulated or translated so as to be sensitive to the kind of data we can collect, *while* maintaining their core hypotheses. To exemplify, we need an explanation of what the difference is between an information carrying signal (mental state) to reach, say, the PFC and become higher-order represented, as opposed to the signal merely reaching the PFC and becoming transformed. We need an answer to why we should think there are two states as opposed to just one that changes, and we need a way to discern this difference either from a dataset that depicts continuous activations or from behavioural measures. We thank an anonymous reviewer for pushing us to clarify this.

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5.
https://doi.org/10.33735/phimisci.2021.37

such claim, but argues that consciousness correlates with activity in the "VAN range" (Koivisto & Revonsuo, 2010). Although the feed-back view and the VAN view both posit that consciousness is to be found in primary sensory regions, they disagree for instance with regards to timing. The feed-back view argues that information must first travel through a feed-forward sweep before it "returns," whereas the VAN view expects the feed-forward sweep to be conscious. Very few experiments have directly attempted to disentangle these views (Crouzet et al., 2014, but see, e.g., 2017), and whether this can be done conclusively is very much an open question.

These problems stem from the fact that localizing the NCC can only tell us so much. Simply pointing to a brain region does not move the needle much. In order properly to bridge the gap between the conceptual domain and our neurological data, we need to at least know about implementation. Localization is not enough in itself. That this is the case has already been acknowledged in discussions of attempts to localize (other) cognitive functions, where merely aiming for localization has been criticised repeatedly as being insufficient in order to provide a deeper understanding (e.g. Carandini, 2012; Mogensen & Overgaard, 2018; Overgaard & Mogensen, 2011). According to this criticism, it is insufficient to point to a correspondence between a 'function' at the mental level and a given structure at the neural level. There is a need for an 'intermediate' computational level. Without a computational level (or understanding the implementation), pinpointing the location of an NCC does little to improve our *understanding* of consciousness. It is worth mentioning that a few of the current models of the NCC are closer to addressing the issue regarding these missing computational processes than others. Among these are the Integrated Information Theory (e.g. Tononi et al., 2016) and the Reorganization of Elementary Functions (REF) framework in the form of the REFCON and REFGEN models proposed by Overgaard and Mogensen (Mogensen & Overgaard, 2018; Overgaard, 2015). In these models the suggested computational processes may point to mechanisms which eventually may yield new empirically testable predictions and thereby contribute to answers in the context of problems of consciousness.

The examples we have provided here are just a sample to illustrate some of the issues that arise from the current practice of trying to fit the evidence to the theory rather than the other way around. Although the debate is far from resolved, we take the issues presented here to be good reason to at least consider alternatives to this practice. One such alternative concerns behavioural methods. We will consider that next.

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5.
https://doi.org/10.33735/phimisci.2021.37

7 Back to behavioural methods

Our second advice concerns shifting focus away from topography. The focus on topography is a trend that has been prevalent in the early-late debate, i.e., the debate concerning whether the NCC are to be found in early processes in the back of the head, or late processes in the front of the head (Boly et al., 2017). Indeed, the debate itself is often framed in terms of topography (either as the occipital-frontal debate, the front-back debate, or by proxy of early-late). As we highlighted above, there is no *necessary* connection between the theoretical posits of higher-order theory and the PFC. Similarly, recurrent processing in the early visual system is, absent radical interpretation, not isomorphic with the ideas of reflexive theories of consciousness. Nevertheless, *pace* disclaimers (Brown et al., 2019; Morales & Lau, forthcoming) activations in specific brain regions have come to be seen as reliable indicators of whether a given empirical finding meshes better with higher-order theory or reflexive theories of consciousness. For instance, arguing that if subjects with lesions to the PFC retained a capacity for consciousness, this would count against the higher-order theories reflects this focus on topography, as does the defence mounted against the lesion data by proponents of higher-order theories (Odegaard et al., 2017) in spite of their own disclaimers (Brown et al., 2019).

In our opinion, focusing on topography is premature. It is premature because the way various brain regions contribute to the processes underpinning consciousness has yet to be determined with any inkling of consensus. That this is true can be seen straightforwardly from the fact that there is still widespread disagreement about where to even locate the NCCs. This disagreement itself is part of what forms the foundation for the debate in the first place. Some may here object that everyone involved is well aware of this and investigating the competing hypotheses involves exactly making predictions about their location. That is standard scientific practice. It is the testing of these predictions that in the end (sometime far into the future) will allow us to distinguish between the hypotheses empirically and determine the real NCCs and the nature of consciousness.

The extrapolation of topographical regions of interest based on conceived similarity in functional characteristics from data obtained in behavioural paradigms is not *per se* misguided. At least to the extent that this practice merely serves to inform new paradigms. What, on the other hand, is misguided is thinking that topographical data obtained through behavioural studies of other phenomena warrants bypassing further behavioural studies when it comes to consciousness, i.e., we cannot let the default be reasoning directly from our (perceived) similarity in functional characteristics of one process to the claim that another process is likely to (or must) reside in the same brain region. This problem is further aggravated if we allow ourselves to count empirical findings as evidence for or against a conceptual theory simply because some or other brain region is involved. If we want to truly deploy empirical sciences to establish the NCCs and distinguish between competing theories of consciousness, we must do so on the basis of differing behavioural predictions rather than predicting activation profiles in brain regions.

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5.
https://doi.org/10.33735/phimisci.2021.37

Experiments that attempt to replace subjective reports with objective behaviour, and behaviour with neural activation, face specific problems. If one is to avoid using subjective reports, one is forced to answer the obvious question: without relying on reports, how do we know that a particular no-report paradigm measures consciousness and not something else? What reasons could one have to think of, say, a binocular rivalry experiment without report as a method to investigate conscious experience per se in the first place (Frässle et al., 2014; Overgaard & Fazekas, 2016)? Arguably, the only reason is that one introspectively attends to what it feels like to experience binocular rivalry. In other words, the only way to invent a measure of consciousness without asking others to report seems to depend on the scientist's own intuitions, which could hardly be said to be any less based on introspection. Another answer could be that if a particular type of behaviour or phenomenon has been found to be associated with subjective experience in previous experiments using subjective reports, this behaviour or phenomenon can be used as an 'objective measure' to replace the report. However, this would carry any methodological weakness in reporting forward—as the identified measure is only a reliable measure under the condition that the correlation with the subjective report was reliable in the first place. Ironically, this requires the subjective measure to be a reliable type of measure, which would eliminate the reason for finding other measures to begin with.

If two competing theories of consciousness have different predictions about neural correlates of consciousness yet identical predictions about behaviour, the two theories will be difficult to compare. This is in fact typically the case in current debates between first and higher-order theories of consciousness. Based on the reasoning above, consciousness theories should be able to propose specific outcomes of behavioural experiments to be empirically compared directly.

8 Concluding remarks

Given the shortcomings discussed above, changes in the interdisciplinary practice are warranted to ensure a fit between neural evidence about the brain and conceptual theories about consciousness. We have offered advice indicating two different ways to go about this.

Our first advice concerned how to bridge the missing isomorphism between theories in the conceptual domain and neural data. We argued that to bridge the conceptual disconnect between the empirical and conceptual domains, bottom-up revision of the conceptual theories is required. This is captured in the dictum *because we cannot make the data fit our theories, we should make our theories fit the data.* Our second piece of advice concerns how to test competing conceptual theories. With regard to this, we advocated the necessity of competing conceptual theories delivering differing behavioural predictions.

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5.
https://doi.org/10.33735/phimisci.2021.37

We do not purport that these are the only ways forward and acknowledge that each requires further refinement than we can provide in this limited space. However, the two different ways may serve as the foundation for further development and facilitate future research in this area.

Acknowledgments
AKH is funded by the Swedish Research Council (Grant # 2018-06595). Morten Overgaard was supported for this work by a Jens Christian Skou fellowship from Aarhus Institute of Advanced Studies.

References

Beeckmans, J. (2007). Can higher-order representation theories pass scientific muster? *Journal of Consciousness Studies*, *14*(9-1), 90–111.

Block, N. (2019). What is wrong with the no-report paradigm and how to fix it. *Trends in Cognitive Sciences*, *23*(12), 1003–1013. https://doi.org/10.1016/j.tics.2019.10.001

Boly, M., Massimini, M., Tsuchiya, N., Postle, B. R., Koch, C., & Tononi, G. (2017). Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and neuroimaging evidence. *Journal of Neuroscience*, *37*(40), 9603–9613. https://doi.org/10.1523/JNEUROSCI.3218-16.2017

Brown, R. (2012). The brain and its states. In S. Edelman, T. Fekete, & N. Zach (Eds.), *Being in time: Dynamical models of phenomenal experience* (Vol. 88, pp. 211–238). https://doi.org/10.1075/aicr.88.10bro

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, *23*(9), 754–768. https://doi.org/10.1016/j.tics.2019.06.009

Carandini, M. (2012). From circuits to behavior: A bridge too far? *Nature Neuroscience*, *15*(4), 507–509. https://doi.org/10.1038/nn.3043

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, *78*(2), 67–90. https://doi.org/10.2307/2025900

Crouzet, S. M., Kovalenko, L. Y., Del Pin, S. H., Overgaard, M., & Busch, N. A. (2017). Early visual processing allows for selective behavior, shifts of attention, and conscious visual experience in spite of masking. *Consciousness and Cognition*, *54*, 89–100. https://doi.org/10.1016/j.concog.2017.01.021

Crouzet, S. M., Overgaard, M., & Busch, N. A. (2014). The fastest saccadic responses escape visual masking. *PloS One*, *9*(2), e87418. https://doi.org/10.1371/journal.pone.0087418

Da Costa, N., & French, S. (1990). The model-theoretic approach in the philosophy of science. *Philosophy of Science*, *57*(2), 248–265. https://doi.org/10.1086/289546

Fink, S. B. (2016). A deeper look at the "neural correlate of consciousness." *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01044

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1338–1349. https://doi.org/10.1098/rstb.2011.0417

Frässle, S., Sommer, J., Jansen, A., Naber, M., & Einhäuser, W. (2014). Binocular rivalry: Frontal activity relates to introspection and action but not to perception. *Journal of Neuroscience*, *34*(5), 1738–1747. https://doi.org/10.1523/JNEUROSCI.4403-13.2014

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, *50*(4), 531–534. https://doi.org/10.1016/j.neuron.2006.05.001

Genon, S., Reid, A., Langner, R., Amunts, K., & Eickhoff, S. B. (2018). How to characterize the function of a brain region. *Trends in Cognitive Sciences*, *22*(4), 350–364. https://doi.org/10.1016/j.tics.2018.01.010

Glock, H. (1997). Kant and Wittgenstein: Philosophy, necessity and representation. *International Journal of Philosophical Studies*, *5*(2), 285–305. https://doi.org/10.1080/09672559708570857

Kirkeby-Hinrup, A. (2020). A higher-order faculty and beyond. In Overgaard, Morten and Mogensen, Jesper and Kirkeby-Hinrup, Asger (Ed.), *Beyond the neural correlates of consciousness* (pp. 131–152). Psychology Press, Routledge.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience*, *17*, 307. https://doi.org/10.1038/nrn.2016.22

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5. https://doi.org/10.33735/phimisci.2021.37

Koivisto, M., & Revonsuo, A. (2010). Event-related brain potential correlates of visual awareness. *Neuroscience & Biobehavioral Reviews*, *34*(6), 922–934. https://doi.org/10.1016/j.neubiorev.2009.12.002

Kozuch, B. (2014). Prefrontal lesion evidence against higher-order theories of consciousness. *Philosophical Studies*, *167*(3), 721–746. https://doi.org/10.1007/s11098-013-0123-9

Kriegel, U. (2007). A cross-order integration hypothesis for the neural correlate of consciousness. *Consciousness and Cognition*, *16*(4), 897–912. https://doi.org/10.1016/j.concog.2007.02.001

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494–501. https://doi.org/10.1016/j.tics.2006.09.001

Lamme, V. A. F. (2018). Challenges for theories of consciousness: Seeing or knowing, the missing ingredient and how to deal with panpsychism. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755). https://doi.org/10.1098/rstb.2017.0344

Lau, H. (2007). A higher order bayesian decision theory of consciousness. *Progress in Brain Research*, *168*, 35–48. https://doi.org/10.1016/s0079-6123(07)68004-2

Lau, H. (2011). Theoretical motivations for investigating the neural correlates of consciousness. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(1), 1–7. https://doi.org/10.1002/wcs.93

Lau, H., & Brown, R. (2019). The emperor's new phenomenology? The empirical case for conscious experiences without first-order representations. In A. Pautz & D. Stoljar (Eds.), *Blockheads! Essays on Ned Block's philosophy of mind and consciousness* (pp. 171–197). MIT Press.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*(8), 365–373. https://doi.org/10.1016/j.tics.2011.05.009

Michel, M., & Morales, J. (2020). Minority reports: Consciousness and the prefrontal cortex. *Mind & Language*, 493–513. https://doi.org/10.1111/mila.12264

Mogensen, J., & Overgaard, M. (2018). Reorganization of the connectivity between elementary functions as a common mechanism of phenomenal consciousness and working memory: From functions to strategies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170346. https://doi.org/10.1098/rstb.2017.0346

Morales, J., & Lau, H. (forthcoming). The neural correlates of consciousness. In U. Kriegel (Ed.), *The Oxford Handbook of the Philosophy of Consciousness*. Oxford University Press.

Naccache, L. (2018). Why and how access consciousness can account for phenomenal consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170357. https://doi.org/10.1098/rstb.2017.0357

Nagel, T. (1974). What is it like to be a bat. *Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914

Odegaard, B., Knight, R. T., & Lau, H. (2017). Should a few null findings falsify prefrontal theories of conscious perception? *Journal of Neuroscience*, *37*(40), 9593–9602. https://doi.org/10.1523/JNEUROSCI.3217-16.2017

Overgaard, M. (2015). *Behavioral methods in consciousness research*. Oxford University Press.

Overgaard, M. (2017). The status and future of consciousness research. *Frontiers in Psychology*, *8*(1719). https://doi.org/10.3389/fpsyg.2017.01719

Overgaard, M. (2018). Phenomenal consciousness and cognitive access. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170353. https://doi.org/10.1098/rstb.2017.0353

Overgaard, M., & Fazekas, P. (2016). Can no-report paradigms extract true correlates of consciousness? *Trends in Cognitive Sciences*, *105*(20(4)), 241–242. https://doi.org/10.1016/j.tics.2016.01.004

Overgaard, M., & Mogensen, J. (2011). A framework for the study of multiple realizations: The importance of levels of analysis. *Frontiers in Psychology*, *2*, 79. https://doi.org/10.3389/fpsyg.2011.00079

Pessoa, L., Thompson, E., & Noë, A. (1998). Finding out about filling-in: A guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences*, *21*(6), 723–748. https://doi.org/10.1017/s0140525x98001757

Peters, M. A., Kentridge, R. W., Phillips, I., & Block, N. (2017). Does unconscious perception really exist? Continuing the ASSC20 debate. *Neuroscience of Consciousness*, *3*(1). https://doi.org/10.1093/nc/nix015

Pinto, Y., Vandenbroucke, A. R., Otten, M., Sligte, I. G., Seth, A. K., & Lamme, V. A. (2017). Conscious visual memory with minimal attention. *Journal of Experimental Psychology: General*, *146*(2), 214–226. https://doi.org/10.1037/xge0000255

Railo, H., Revonsuo, A., & Koivisto, M. (2015). Behavioral and electrophysiological evidence for fast emergence of visual consciousness. *Neuroscience of Consciousness*, *2015*(1). https://doi.org/10.1093/nc/niv004

Sebastián, M. Á. (2014). Not a HOT dream. In R. Brown (Ed.), *Consciousness inside and out: Phenomenology, neuroscience, and the nature of experience* (Vol. 6, pp. 415–432). Springer Netherlands.

Timmermans, B., & Cleeremans, A. (2015). How can we measure awareness? An overview of current methods. In M. Overgaard (Ed.), *Behavioral methods in consciousness research* (pp. 21–46). Oxford University Press.

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5. https://doi.org/10.33735/phimisci.2021.37

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450–461. https://doi.org/10.1038/nrn.2016.44

Weisberg, J. (2014). Sweet dreams are made of this? A HOT response to Sebastián. In R. Brown (Ed.), *Consciousness inside and out: Phenomenology, neuroscience, and the nature of experience* (Vol. 6, pp. 433–443). Springer Netherlands.

Overgaard, M. S., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences*, *2*, 5.
https://doi.org/10.33735/phimisci.2021.37