Interdisciplinary Consciousness Studies needs Philosophers of Science

Asger Kirkeby-Hinrup

Abstract
Significant progress has been made over the last couple of decades with respect to empirical investigations of consciousness. The field of interdisciplinary consciousness studies is sprawling and growing at a rapid pace. There is, however, still much to figure out, and the philosophy of science has an important role to play in untangling the complex relationships between theory and empirical data. I start by providing a rough overview of the historical developments over the last couple of decades with an eye to how the research focus has shifted in that time. Next, I describe the current state of the field and highlight some of the core problems in interdisciplinary consciousness studies that would benefit from the involvement of philosophers of science. Finally, I summarize three contemporary endeavors in the attempt to assess and compare theories of consciousness. The goal is to convey that interdisciplinary consciousness studies is a field ripe for the application of philosophy of science, and hopefully inspire philosophers of science to come help us out.

Keywords: Consciousness; Neuroscience; Cognitive Science; Empirical Evidence; Falsification; Inference to the best explanation

1. Introduction

The field of interdisciplinary consciousness studies (i.e. work at the intersection between philosophy of mind, psychology, cognitive science, and neuroscience) has been blossoming over the last two and a half decades. The explosion of communication technology, focus on cross disciplinary interactions, new clinical technologies (and their reduced cost leading to increased access to these), have combined to usher in what is sometimes called *the empirical turn* in consciousness studies. To set the stage — and provide context — for the problems I will lay out below, it is useful with a brief — and consequently rough — historical overview. Obviously, the empirical turn did not emerge from nothing. There are many examples of philosophers of mind considering empirical phenomena significantly earlier

than the last couple of decades (e.g. LeDoux, Michel, & Lau, 2020; Miller, 2003; Nagel, 1971). However, two things characterize the start of the empirical turn in relation to this earlier period. The first concerns the scope, meaning that the empirical turn relates specifically to *consciousness* studies, and not to the broader domain of philosophy of mind. The second is a radical increase in the amount of empirical phenomena invoked to support theories of consciousness, and in the number of publications discussing empirical phenomena (see e.g. figure 2b in Yaron, Melloni, Pitts, & Mudrik, 2021). In the early stages of the empirical turn, it mainly was characterized by work drawing connections between a specific theory of consciousness (often a theory from the philosophical domain heralding back to before the empirical turn) and one or more empirical phenomena (e.g., Dretske, 2004; Lau, 2007). Crudely, one might say that at the onset of the empirical turn most researchers involved were concerned primarily with establishing empirical support for their preferred theory, i.e. delivering plausible interpretations of empirical phenomena in light of said theory. Likely spurred by prominent publications (Block, 2007 comes to mind) the focus in the empirical turn broadens in the years around 2010 and lists of empirical support (e.g., Lau & Rosenthal, 2011) and considerations of how to compare them (e.g. Block, 2009) receive increased attention. At that stage arguing for theories of consciousness — or comparing them — from an empirical perspective is becoming the standard. This led to the practice of contesting claims of empirical support (e.g., Kirkeby-Hinrup, 2014, 2016; Malach, 2011)(Brinck & Kirkeby-Hinrup, 2017; Kirkeby-Hinrup, 2014, 2016), and arguing against philosophical claims on empirical grounds (e.g., Sebastián, 2014) becoming more popular.

Only in the last couple of years, the focus has increasingly turned to where we go from here. With proponents of every theory having spent significant time and effort identifying a range of empirical phenomena supporting their respective theories, only now is significant attention allocated to questions about the best use of the purported empirical support, the prospects of resolving central questions in the debates, and the work empirical evidence can do for us in these regards. However, there is a range of not insignificant problems in this regard. Therefore, now is a very appropriate occasion to involve philosophers of science.

In the rest of this text, I highlight key aspects of the debates concerning empirical evidence and its application within interdisciplinary consciousness studies. I start by sketching a basic premise and a closely related background assumption shared by proponents of most theories. Then, I turn to highlighting some central problems regarding the application of empirical

data in the domain of consciousness studies. Finally, I present notable cases of new projects or approaches that independently of each other have emerged in parallel over the last couple of years with the shared objective to figure out where we go from here. To be clear, I will introduce these issues as neutrally as possible since the aim here is not to advance one theory or approach over the others. Rather, the aim of this paper is to bring to the attention of philosophers of science the field of interdisciplinary consciousness studies, to pique your interest, illustrate to you some problems where your expertise is needed, and entice you to explore further.

2. Theories of consciousness and empirical data

Before The most fundamental premise shared by most philosophers, cognitive neuroscientists, psychologists and most everyone else working in the field of interdisciplinary consciousness studies during the empirical turn, is the belief that consciousness can be naturalized. What this entails is formulated neatly by Jean Petitot *et al*. when they write: "By 'naturalized' we mean integrated into an explanatory framework where every acceptable property is made continuous with the properties admitted by the natural sciences" (Petitot, Varela, Pachoud, & Roy, 1999)(Petitot, Varela, Pachoud, & Roy, 1999, pp. 1–2). Normally, what this (roughly) is taken to entail is that the mind depends — in some substantive way — on the brain. The upshot of this, and what fuses these researchers together into a single field of study is (approximately) the belief that understanding the brain's role in relation to consciousness is central to understanding consciousness *per se* and its associated concepts (e.g. experience, cognition, meta-cognition, emotion, action, and perception). As Josh Weisberg (2013, p. 433) writes:

> […] A challenge rooted in empirical data. This is the proper way to approach consciousness […]

This, in turn, leads to the assumption that lies at the base of the empirical turn which is that the more empirical data a theory can explain or predict, the stronger it stands in comparison to other theories. This assumption can be said to be a kind of practical consequence of the central premise, since there seems to be a tension between the claim that understanding the brain is central to understanding consciousness and the claim that empirical data has no weight in our work (theories) to understand consciousness. Additionally, this assumption insinuates some process of inference to the best explanation. While this is merely implied in much of the literature, occasionally it is stated explicitly. For instance by Ned Block in his seminal paper (2007, p. 486):

> I have in mind […] the familiar default 'method' of inference to the best explanation, that is, the approach of looking for the framework that makes the most sense of all the data […]

This might seem straightforward and fairly uncontroversial. However, in practice a range of interconnected systemic problems with this project have manifested themselves through the debates. In the rest of this section, I briefly diagnose the most central of these problems and highlight some of their connections. In section 3, I will go into a little more depth with the possibility of inferring to the best explanation in the domain of interdisciplinary consciousness studies. Importantly, however, an inference to the best explanation (IBE) process is only one of a few endeavors currently underway to assess and compare theories. Two other prominent endeavors are also introduced in section 3.

2.1 Theories are conceptually sound

The vast majority of the theories available (and certainly each of the most prominent ones) are internally consistent conceptual frameworks and propose reasonably well-defined mechanisms underpinning phenomenal consciousness[1]. Examples of such proposed mechanisms include higher-order representation (e.g., Rosenthal, 1997), broadcasting in a global workspace (e.g., Baars, 1996) or information integration (Tononi, 2005). It might seem wrong to call this a problem and certainly there is a sense in which this is both very desirable and to be expected. It is desirable because certainly we want conceptually coherent and internally consistent theories. It is expected because the proponents of each theory are competent and serious researchers. Now, having multiple good (i.e. internally consistent, and conceptually coherent) candidate theories would not be a problem if there were uncontentious and reliable empirical means of distinguishing between them. The issue with theories of consciousness is that — at the present time — there appears to be neither (c.f. Chalmers, 1995; Overgaard & Kirkeby-Hinrup, 2021).

[1] Mainly, what is at stake in the debates between competing theories of consciousness concerns *phenomenal* consciousness. Covered by various names: qualia, phenomenality, experiential properties, subjective feeling or the *what it is like-ness* of being conscious or being in a given conscious state. For simplicity, I will use 'consciousness' to refer to phenomenal consciousness throughout unless otherwise noted.

2.2 Disagreements about (shared) central concepts

One major fault line roughly divides the field into two camps[2]: the nature and importance of phenomenality. On this question, the field can be divided into those who hold *deflationary* accounts (e.g., Rosenthal, 2008) and those who advance *inflationary* accounts (e.g., Block, 2011b)[3]. The questions in this context are "how much" phenomenality is there? What, if anything, can we say about the nature of phenomenality? And does it play any special role in consciousness and/or cognition? Importantly, resolving the debates about "how much" phenomenality there is does not piggyback on resolution of the issue about the neural machinery (mechanisms) underpinning consciousness. In other words, it is possible that there is a lot of phenomenality, even if a theory of consciousness traditionally associated with a deflationary view of phenomenality turns out to correctly reflect how the world is. The problem at the root of these issues is that we have no adequate empirical direct measure of phenomenality. Currently, the best proxy measure for phenomenality is considered to be subjective measures (such as verbal reports), but even the adequacy of these have routinely been called into question on different grounds. For instance, it has been argued that introspective judgements (that form the basis for subjective measures) are unreliable (e.g., Schwitzgebel, 2008), and it has even been argued that there may be phenomenality that is inaccessible for report (Block, 2008), i.e. that we have experiences that we do not and cannot know about. To many, this latter conjecture seems to be a conceptual falsehood. Phenomenal consciousness, they argue, conceptually entails that the experiencer is aware of the conscious experience (see e.g. Weisberg, 2011). For instance, many think it true that a mental state, one is in no way aware of being in, is not a conscious state in any meaningful sense of the word. Proponents of higher-order thought theory even deploy this idea as the foundation of their definition of consciousness; the so-called *Transitivity Principle* (TP), which is the idea that a conscious state is a state one is aware of oneself as being in (see e.g. Matey, 2006; Rosenthal, 1997; Weisberg, 2010). Conversely, it is often argued against proponents of higher-order theories that their theories have similar counterintuitive or conceptually false consequences. No place is this more on display than with respect to the possibility of misrepresentation in consciousness. Briefly, most variants of higher-order theory allow for individuals to consciously experience being in

[2] Normally, there will be a strong correlation between which theory one espouses and one's view on the nature of phenomenality. But this connection between certain theories and one or the other view of phenomenality is not the concern here.

[3] I take for granted that readers understand what deflationary and inflationary roughly entails in this context.

states that they are not really in, something that most opponents of higher-order theories find to be a decisive flaw with the theory (Block, 2011a; Wilberg, 2010), and possible even conceptually false (e.g. Kriegel, 2003; Wilberg, 2010 but see also Berger, 2014). The reply from the higher-order camp, is that the possibility of misrepresentation follows naturally from the understanding of the relevant higher-order states as representational states, because the concept of representation does not entail the existence of what is being represented (e.g. representations of unicorns). I will not go into further detail with either of these long running and complex debates. The aim here is merely to put on display the deep conceptual divides in the debates, where proponents of most theories are accused of deploying central concepts in ways that (their opponents take to) violate their essential meaning. This has the unfortunate side effect that much of the criticism leveraged against theories consists reiterations of the view that some concept should be understood the way the critic understands it, and that a theory is wrong because it does not. Indeed, accusations that theories are counterintuitive or controversial are not uncommon (e.g., Prettyman, 2020). And even worse, criticism based in question begging arguments, where the critic takes for granted his own understanding of a central concept and proceeds to show that the targeted theory is inconsistent when this understanding is deployed instead of the targeted theory's own understanding are not uncommon (Block, 2011a; Lane & Liang, 2008; Wilberg, 2010). It is not that the participants in the debates are blind to this issue and it is pointed out occasionally, for instance when Rosenthal in reply to criticism says "The phrase 'what it's like' is not reliable common currency." (Rosenthal, 2011, p. 434). This lack of common conceptual ground is exactly the issue I aim to highlight here. When the competing theories each are internally consistent, describe the target phenomenon using many of the same concepts — yet disagree about what those concepts actually mean — there is little avenue on conceptual grounds to determine which theory is correct, or even preferable. An unspoken acceptance in the field is emerging that the debate on conceptual grounds may have run its course and reached a stalemate. At least this could be seen as a plausible driver of the intensifying interest in empirical evidence. Given the shared fundamental premise — the belief the mind can be naturalized — mentioned in the introduction, the hope seems to be that empirical evidence may arbitrate in the debate and resolve the conceptual stalemate.

2.3 Conceptual bleed

As elaborated on above, there are central differences between the conceptual commitments of proponents of competing theories of consciousness. Furthermore, each of the competing theories of consciousness is sound and internally consistent. To a large extent this has led the debate into a stalemate, in which it is difficult to criticize a theory without begging the question against its underlying conceptual framework. However, because most people involved in the debates share the assumption that consciousness can be naturalized, the hope is that empirical evidence may arbitrate in these debates through determination of which theory is more empirically plausible. I.e., empirical evidence — the hope is — will resolve the disagreements in the conceptual domain.

The problem with conceptual bleed is that before it is feasible to draw inferences for or against a theory from empirical data one needs, as a minimum, an interpretation that maps the relevant concepts of the empirical data and the theory to each other. Plausibly, the preliminary conceptual mapping influences the kinds of inference available to be made from the empirical data. Furthermore, the philosophical predilections of the interpreter are bound to influence the mapping of concepts between theory and empirical data. If, for nothing else, simply because philosophical predilections (regardless of what they are) influence how one prefers to conceptualize and describe phenomena (e.g. consciousness). Thus, it seems that the conceptual commitments one has in the philosophical domain are likely to influence the way in which one interprets the empirical data. Importantly, this is not a question of bias in the interpretation. Rather, it is a natural consequence of the conceptual and theoretical commitments of the interpreter. It would be unfair to expect researchers, when interpreting relevant empirical data, not to make use of the concepts they think best describe and categorize the phenomenon under investigation. In sum, the conceptual commitments a given researcher has in the conceptual domain bleeds into, as it were, the empirical domain. Why is this a problem? The issue at the center here is that because of conceptual bleed, the differences in conceptual commitments between the theories will show up again in the interpretations of empirical data. This in turn means that comparing interpretations of empirical data becomes difficult if not impossible, because criticizing theory-laden interpretations runs into the same issues that we had in the conceptual domain with respect to begging the question against a theory. The upshot is that empirical evidence may not be able to help us distinguish between theories that are on equal standing in the conceptual domain. Unfortunately, this was exactly the work empirical evidence was introduced to do for us in these

debates. Importantly, the claim here is not that empirical evidence can never achieve this. Rather, the claim is that at the present time, what we see is many theories each giving plausible (given their respective conceptual commitments) interpretations of many of the same empirical phenomena, and there currently is no consensus on how to arbitrate between mutually exclusive interpretations. This gordian knot is a primary reason why the field of interdisciplinary consciousness studies needs the expertise of philosophers of science, especially so because many of the other problems regarding comparing theories of consciousness piggyback on conceptual bleed.

2.4 Assessing Interpretations of evidence

One very important aspect of the interdisciplinary debate in consciousness studies concerns how to assess the proposed empirical evidence for a given theory of consciousness. When it comes to assessment, what we are concerned with is validating, correcting, or rejecting evidence on a case-by-case basis. There are several issues with the practice of validating evidence, but before we turn to those, it is useful to motivate briefly why validation is important in the first place. First of all, it is instructive to note that validation has counterparts in other fields that are accepted broadly as both relevant and important. The most pertinent example of this can be found in the notion of replication in empirical sciences. Another example (albeit from outside of academics) can be found in the practice of fact checking in public discourse. What is shared by validation of empirical evidence in consciousness studies and its extra-disciplinary counterparts is the practice of determining — through independent assessment — whether a given finding or claim is correct (or plausible). There are many ways in which a proposed connection between an empirical finding and a philosophical claim may turn out to be faulty. Examples include straightforward errors in argumentative structure (e.g. Kirkeby-Hinrup, 2014), implausible interpretations, or ignoring equally reasonable interpretations (e.g. Brinck & Kirkeby-Hinrup, 2017; Kirkeby-Hinrup, 2016), and errors that arise from cognitive bias. Assessing proposed empirical support for theories is crucial to interdisciplinary consciousness studies, since we should not count an empirical finding among the phenomena that supports a theory if it does not in fact support the theory. Given that we think considering the empirical support of theories can move forward the debate between competing theories of consciousness it is imperative that we double check on a case-by-case basis (i.e. validate) the support of each theory. Elsewhere (Kirkeby-Hinrup & Fazekas, 2021), I have elaborated on – and given examples of – validation of empirical support proposed for theories of consciousness, so I will only touch on this briefly

here. Generally, there appears to be at a minimum two steps involved. The first step concerns the interpretation of the empirical evidence, as was discussed above in the section on conceptual bleed. In this step, what is of concern is assessing whether the interpretation of the empirical results is plausible in light of the empirical paradigm and the vocabulary and explanation given by the empirical scientists who originally reported the finding. Evaluating these interpretations is of special importance when two or more empirical findings or phenomena are combined to make an argument. This is because there may be discrepancies between the paradigms used, or important dissimilarities between the investigated empirical phenomena, which may require significant conceptual acrobatics in order to make the multiple findings fit into the same framework. Relevant questions in this regard include whether the same concepts are applied to — or operationalized in — the interpretation in a uniform way. Whether there are equivocations or vagueness in the application of terms from the conceptual framework. In the second step of the assessment what is of interest is the posited connections between the interpretation of the empirical data and a theoretical claim. The kinds of questions that are of interest in this context concerns how the proposed interpretations map onto the theoretical conceptual framework, and illuminating (by extrapolation, if it is not explicitly specified, as it rarely is) the argumentative structure leading from the interpretation to the theoretical claim (e.g. Kirkeby-Hinrup, 2014). The main problem with assessing empirical support again concerns conceptual bleed. Given that much of the vocabulary deployed in the interpretation of — and subsequent argument based on — an empirical finding will be imported from a theory of consciousness, much care needs to be taken in the evaluation. Indeed, in some cases the interpretations will not admit of much criticism since any criticism leveraged would be begging the question against the conceptual framework of the theory. This issue looms large in the work on assessing the deployment of empirical evidence and can make this work feel unrewarding at times because of the many inconclusive results. Nevertheless, for the reasons stated above, this work has an important role to play (even if the result ends up discarding many pieces of proposed empirical support). In relation to the philosophy of science, it is worth noting that traditional processes such as inference to the best explanation and abduction[4] do not seem to be straightforwardly applicable to this process, if for nothing else

[4] To the extent that one thinks of these as different. There is a case to be made that abduction is best applied to connect a single piece of empirical evidence and a single theory, whereas IBE is most suitably applied in the comparison of theories. However, I will not delve further into this issue here.

because usually the interpretations offered rarely involve more than one theory, which is implied by both IBE and abduction (Harman, 1965; Minnameier, 2004, 2010). In contrast to this, most cases of empirical support proposed in favor of a theory of consciousness considers only the one theory.

3. Assessing and comparing empirical support

The previous two sections covered where we were, and where we are. In this section, the focus is where we are trying to go. As mentioned in the introduction, the hope is that empirical evidence may help us arbitrate between competing theories of consciousness by allowing us to determine which theory is most plausible from an empirical perspective. Pertaining to this, there are several parallel efforts ongoing already in the field. These ongoing efforts can roughly be divided into three different categories: 1) falsification-type work, 2) IBE-type work, 3) criteria for theories of consciousness. Each of these three approaches has strengths and weaknesses, and it is — as of yet — an open question, which of these approaches is the most promising. Furthermore, it is worth noting that the assistance that philosophers of science can provide to each may also differ. In the rest of this section, I will briefly introduce each of these. These introductions are intended only as illustrations of the kind of work happening in the field and the associated problems to which philosophy of science may find propitious application.

3.1 Falsification-type work

The most prominent and promising endeavor working in the vein of falsification is called "Accelerating Research on Consciousness"[5] and includes five separate projects deploying the principle of adversarial collaboration. The core idea in this adversarial collaboration project is to engage with the proponents of competing theories of consciousness to reach an agreement on empirical paradigms on which the theories deliver different predictions. The differing predictions is what makes this approach akin to the process falsification. The crux of course is to then carry out the empirical work and test the competing predictions. Assuming no unforeseen derailing of the projects, the upshot of the adversarial collaboration will be that the prediction of one theory will not be confirmed. However, it is unclear if the theory whose prediction is not confirmed will be considered falsified in any substantive sense where further work on the theory is completely abandoned.

[5] Funded by the Templeton foundation to the tune of 20 million dollars. See https://www.templetonworldcharity.org/accelerating-research-consciousness-our-structured-adversarial-collaboration-projects for further info.

More likely, proponents of the losing (as it were) theory will presumably take this as an incentive to further develop the theory to account for the failed prediction. While, to some, this may sound eerily similar to explaining away the negative empirical result, it is not entirely unreasonable, given that many theories are still very early in their development, and have other empirical data to lean on. Nevertheless, this happening will certainly put pressure on the core tenet of the adversarial collaboration, which is to get proponents of each theory to step up to the plate in a substantial way, where this arguably should entail some commitment to the results even if those results turn out against them. To be clear, there is no indication (as of yet) that this will happen, but it is a possible scenario.

3.2 Inference to the best explanation-type work

The notion of inference to the best explanation is tacitly present in much of the work in the empirical turn and is occasionally even stated explicitly (cf. Block quote in section 2 above). Nevertheless, until now, there has not been any systematic effort to attempt this in practice. One reason for this likely is the lack of comprehensive datasets covering the empirical evidence proposed for each of the extant theories. In recent years there has been some effort to ameliorate this issue. For instance, Peter Fazekas and I (2021) recently published a complete dataset relating to the framework proposed by Ned Block. The compilation of empirical evidence we offer is accompanied by a proposal for a process of inference to the best explanation in the domain of interdisciplinary consciousness studies. However, this proposal is lacking significantly when it comes to the actual comparison of theories on the bases of their empirical support. In the paper, we acknowledge this shortcoming, and punt the issue to future work (we also suggest involving philosophers of science to assist with this). One suggestion worth highlighting in this context, is the idea to take a Bayesian approach to the dataset to quantify the amount of empirical support each theory enjoys. There are however several outstanding questions and issues when it comes to the feasibility of this. First of all, it will be impossible to carry out any comparison until datasets from more than one theory is available. Thus, a necessary first step would be the compilation of all the empirical support proposed in favor of the other theories. This, however, is merely a practical problem. A more worrying problem concerns how to carry out Bayesian updating in practice. The most significant issue here concerns how to set the priors necessary for a Bayesian updating process for the likelihood of each theory given the evidence. To compare theories on theory-neutral parameters, the quantification of each particular piece of proposed empirical support must be well motivated and

not arbitrarily bias a comparison. There are several paths one can take that can reasonably be said to remain neutral in this regard. For instance, one may argue that data from neuroimaging should be given more weight that introspective judgments, given that the former is seen as more objective, whereas the latter is inherently subjective. Alternatively, one may argue that widely replicated phenomena in neurotypical individuals should weigh heavier in the comparing than e.g. rare phenomena only reported in a few studies. It may also be possible to score empirical support based on the result of validation process we suggested (2021, section 2.3). Hence, it seems that there may at least be some ways to argue how we should set the priors of one phenomenon in relation to the priors of other phenomena. Nevertheless, while such practices avoid the accusation that priors are set entirely arbitrarily, they still allow significant wiggle room, which may be source of contention. One possible way to further strengthen this process would be if it was possible to show, that given the relative constraints envisioned above, some theories still performed better than others on a wide range of actual implementations of these relative constraints (e.g. if a theory mostly came out superior given a wide variety of settings of the priors). I shall not speculate more about the concrete implementation of inference to the best explanation based in Bayesianism, but trust that the above is sufficient to highlight the kind of work being done, and the kind of remaining work necessary.

3.3 Criteria for theories of consciousness

Evaluating theories of consciousness by deploying specific criteria, one posits a theory should be able to satisfy, is not new. In fact, often criticism is leveraged by suggesting that one theory or another does not satisfy a criterion (based on some feature that the critic take as an obvious truth about consciousness). Unfortunately, this approach has yielded little progress (see e.g. section 2.1 above). However, in a recent paper, Doerig and colleagues (2020) have proposed a set of criteria supposed to be neutral among theories, which they propose to deploy to evaluate and compare theories. Doerig et al. propose two categories of criteria for assessment (e.g. table in Doerig et al., 2020, p. 48). The first category, they dub *criteria*. This category they divide into four challenges a theory of consciousness may face. Depending on the hypothesized mechanisms underpinning consciousness, these challenges may be more or less problematic for a theory. The second proposed category Doerig et al. call *scope*. In this category, they list five classical distinctions about consciousness to assess which aspects of the phenomenon of consciousness is covered by a given theory. They score each theory on each distinction based on whether the answer is explicit, implied, or under-

determined by the theory. Prima facie, assessing the scope of theories makes sense because having good descriptions of the phenomena a theory seeks to provide an explanation of is crucial to assessing the validity and adequacy of the explanation. Furthermore, when we try to determine which candidate explanation of consciousness is preferable, what we mean by "preferable" reasonably is determined at least partly by how encompassing a candidate explanation is, i.e. how much — or how many aspects — of the phenomenon it explains. While the criteria proposed by Doerig et al. are based on empirical considerations, it does little to tell us how we should consider empirical evidence proposed in favor of theories. Furthermore, while their criteria — and the mapping of how different theories cope with these — serve to illustrate where each theory has challenges, it does not tell us how to handle cases where two or more theories can satisfy the same number of criteria, i.e. it does not tell us if any theory is preferable over others in such cases. The criteria approach has also received criticism (See Fahrenfort & van Gaal, 2021 for just one example), and has led to substantial debate (Doerig, Schurger, & Herzog, 2021).

4. Concluding remarks

In the above, I have done three things. First, I provided a brief history of the field of interdisciplinary consciousness studies. Secondly, I sketched some problems this field is facing with respect to determining which of the available competing theories is preferable. Finally, I sketched the three major projects currently being pursued to assess and compare theories. My aim has been to offer an introduction to the field of interdisciplinary consciousness studies, the problems we face in our work in this field, and the historical background against which these problems are framed. This was done to carve out some lacuna where philosophers of science may be of help. Significant progress has been made over the last couple of decades with respect to empirical investigations of consciousness, but there is still much to figure out, and the philosophy of science has an important role to play in untangling the complex relationships between theory and empirical data.

References

Baars, B. J. (1996). Understanding subjectivity: Global workspace theory and the resurrection of the observing self. *Journal of Consciousness Studies, 3*(3), 211–216.

Berger, J. (2014). Consciousness is not a property of states: A reply to Wilberg. *Philosophical Psychology, 27*(6), 829–842.

Block, N. (2007). Consciousness, accessibility, and the mesh between

psychology and neuroscience. *Behavioral and brain sciences, 30*(5–6), 481–499.

Block, N. (2008). Consciousness and Cognitive Access. *Proceedings of the Aristotelian Society, 108*, 289–317.

Block, N. (2009). Comparing the major theories of consciousness. In M. S. Gazzaniga, E. Bizzi, L. M. Chalupa, S. T. Grafton, T. F. Heatherton, C. Koch, & B. A. Wandell (Eds.), *The cognitive neurosciences* (pp. 1111–1122). Cambridge, MA, US: Massachusetts Institute of Technology.

Block, N. (2011a). The higher order approach to consciousness is defunct. *Analysis, 71*(3), 419–431.

Block, N. (2011b). Perceptual consciousness overflows cognitive access. *Trends in cognitive sciences, 15*(12), 567–575.

Brinck, I., & Kirkeby-Hinrup, A. (2017). Change blindness in higher-order thought: Misrepresentation or good enough? *Journal of Consciousness Studies, 24*(5–6), 50–73.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of consciousness studies,* 3, 200–219

Doerig, A., Schurger, A., & Herzog, M. H. (2020). Hard criteria for empirical theories of consciousness. *Cognitive neuroscience, 12*(2), 41–62.

Doerig, A., Schurger, A., & Herzog, M. H. (2021). Response to commentaries on 'hard criteria for empirical theories of consciousness'. *Cognitive neuroscience, 12*(2), 99–101.

Dretske, F. (2004). Change blindness. *Philosophical Studies, 120*(1–3), 1–18.

Fahrenfort, J. J., & van Gaal, S. (2021). Criteria for empirical theories of consciousness should focus on the explanatory power of mechanisms, not on functional equivalence. *Cognitive neuroscience, 12*(2), 93–94.

Harman, G. H. (1965). The inference to the best explanation. *Philosophical Review, 74*(1), 88–95.

Kirkeby-Hinrup, A. (2014). Why the rare Charles Bonnet cases are not evidence of misrepresentation. *Journal of Philosophical Research, 39*, 301–308.

Kirkeby-Hinrup, A. (2016). Change Blindness and Misrepresentation. *Disputatio, 8*(42), 37–56.

Kirkeby-Hinrup, A., & Fazekas, P. (2021). Consciousness and inference to the best explanation: Compiling empirical evidence supporting the access-phenomenal distinction and the overflow hypothesis. *Consciousness and cognition, 94*, 103173.

Kriegel, U. (2003). Consciousness as intransitive self-consciousness: Two

views and an argument. *Canadian Journal of Philosophy, 33*(1), 103–132.

Lane, T., & Liang, C. (2008). Higher-Order Thought and the Problem of Radical Confabulation. *The Southern journal of philosophy, 46*(1), 69–98.

Lau, H. (2007). A higher order Bayesian decision theory of consciousness. *Progress in brain research, 168*, 35–48.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in cognitive sciences, 15*(8), 365–373.

LeDoux, J. E., Michel, M., & Lau, H. (2020). A little history goes a long way toward understanding why we study consciousness the way we do today. *Proceedings of the National Academy of Sciences, 117*(13), 6976–6984.

Matey, J. (2006). Two HOTs to handle: The concept of state consciousness in the higher-order thought theory of consciousness. *Philosophical Psychology, 19*(2), 151–175.

Miller, G. A. (2003). The cognitive revolution: a historical perspective. *Trends in cognitive sciences, 7*(3), 141–144.

Minnameier, G. (2004). Peirce-suit of truth–why inference to the best explanation and abduction ought not to be confused. *Erkenntnis, 60*(1), 75–105.

Minnameier, G. (2010). Abduction, Induction, and Analogy. In L. Magnani, W. Carnielli, & C. Pizzi (Eds.), *Model-Based Reasoning in Science and Technology: Abduction, Logic, and Computational Discovery* (pp. 107–119). Berlin, Heidelberg: Springer Berlin Heidelberg.

Nagel, T. (1971). Brain bisection and the unity of consciousness. *Synthese, 22*.

Overgaard, M., & Kirkeby-Hinrup, A. (2021). Finding the neural correlates of consciousness will not solve all our problems. *Philosophy and the Mind Sciences, 2*.

Petitot, J., Varela, F., Pachoud, B., & Roy, J.-M. (1999). *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*: Stanford University Press.

Prettyman, A. (2020). The persistent problem of targetless thought. *Consciousness and cognition, 82*, 102918.

Rosenthal, D. M. (1997). A Theory of Consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The Nature of Consciousness: Philosophical Debates* (pp. 729–753): MIT Press.

Rosenthal, D. M. (2008). Consciousness and its function. *Neuropsychologia,*

46(3), 829–840.
Rosenthal, D. M. (2011). Exaggerated reports: reply to Block. *Analysis, 71*(3), 431–437.
Schwitzgebel, E. (2008). The Unreliability of Naive Introspection. *Philosophical Review, 117*(2), 245–273.
Tononi, G. (2005). Consciousness, information integration, and the brain. In L. Steven (Ed.), *Progress in brain research* (Vol. Volume 150, pp. 109–126): Elsevier.
Weisberg, J. (2010). Misrepresenting Consciousness. *Philosophical Studies, 154*(3), 409–433.
Weisberg, J. (2011). Abusing the notion of what-it's-like-ness: A response to Block. *Analysis, 71*(3), 438–443.
Wilberg, J. (2010). Consciousness and false HOTs. *Philosophical Psychology, 23*(5), 617–638.
Yaron, I., Melloni, L., Pitts, M., & Mudrik, L. (2021). The Consciousness Theories Studies (ConTraSt) database: analyzing and comparing empirical studies of consciousness theories. *bioRxiv*.

Asger Kirkeby-Hinrup
asger.kirkeby-hinrup@fil.lu.se