NEUROSCIENCE of Consciousness

Neuroscience of Consciousness, 2021, **7(2)**, 1–13

DOI: https://doi.org/10.1093/nc/niab021

Review Article

Special Issue: Consciousness science and its theories

Explanatory profiles of models of consciousness - towards a systematic classification

Camilo Miguel Signorelli[1,2,3,†,*], Joanna Szczotka[4,5,†] and Robert Prentner[6,7,†]

[1]Cognitive Neuroimaging Unit, INSERM U992, NeuroSpin, CEA, Gif sur Yvette F-91191, France; [2]Department of Computer Science, University of Oxford, 15 Parks Rd, Oxford OX1 3QD, UK; [3]Center for Brain and Cognition, Universitat Pompeu Fabra, Edifici Merce Rodereda, Carrer de Ramon Trias Fargas, 25, Barcelona 08018, Spain; [4]Center for Sleep and Consciousness, University of Wisconsin-Madison, 6001 Research Park Blvd, Madison WI 53719, USA; [5]Consciousness Lab, Institute of Psychology, Jagiellonian University, 6 Ingardena, Kraków 30-060, Poland; [6]Department of Cognitive Sciences, University of California, 3151 Social Science Plaza, Irvine CA 92697-5100, USA;[7]Center for the Future Mind, Florida Atlantic University, 777 Glades Road - SO 283, Boca Raton FL 33431-0991, USA

[†]All authors contributed equally to this work

*Correspondence address. Cognitive Neuroimaging Unit, INSERM U992, NeuroSpin, CEA, Gif sur Yvette F-91191, France. E-mail: camiguel@uc.cl

Abstract

Models of consciousness aim to inspire new experimental protocols and aid interpretation of empirical evidence to reveal the structure of conscious experience. Nevertheless, no current model is univocally accepted on either theoretical or empirical grounds. Moreover, a straightforward comparison is difficult for conceptual reasons. In particular, we argue that different models explicitly or implicitly subscribe to different notions of what constitutes a satisfactory explanation, use different tools in their explanatory endeavours and even aim to explain very different phenomena. We thus present a framework to compare existing models in the field with respect to what we call their 'explanatory profiles'. We focus on the following minimal dimensions: mode of explanation, mechanisms of explanation and target of explanation. We also discuss the empirical consequences of the discussed discrepancies among models. This approach may eventually lead to identifying driving assumptions, theoretical commitments, experimental predictions and a better design of future testing experiments. Finally, our conclusion points to more integrative theoretical research, where axiomatic models may play a critical role in solving current theoretical and experimental contradictions.

Keywords: models of consciousness; explanation; unification; phenomenology; integrated information; global workspace; NCCs; quality; quantity

Introduction

Models of consciousness set out to provide a principled description of how the physical domain relates to conscious experience (Seth 2007; Seth 2009; Durham *et al.* 2020). In the last decades, consciousness researchers put forward an abundance of conceptual and formal proposals, drawing from neuroscience, physics, mathematics, philosophy and experimental psychology. In an early scientific phase, it is natural to expect many competing models to develop in parallel to each other. A more mature stage should entail a substantial cross-talk between them, aiming at distilling critical similarities and differences between them, extracting precise empirical predictions (Boly *et al.* 2017) and lastly, eliminating falsified frameworks through a set of crucial empirical experiments, as presently envisioned (Reardon 2019; Melloni *et al.* 2021).

An alternative would be to demand that competing theories need to 'converge' to a unified, synthesized account in order to make progress (Northoff and Lamme 2020; Wiese 2020). In the following paragraphs, we posit that there are currently several, serious impediments to both crucial experiments but also convergence approaches. Arguably, more conspicuous and empirically tangible differences in the theories (such as their postulated neural correlates: prefrontal cortex or posterior hot zone) derive from much deeper, implicitly held deviations in theoretical and philosophical assumptions. In particular, proponents of different theories seem to substantially disagree on what would constitute a 'satisfactory explanation' of consciousness in the first place. Therefore, the aims of these theories are sometimes different. Once all these discrepancies are fully made explicit, the major models can start to enrich each other in a meaningful way. At least in some cases, theoretical misalignment between the models might boil down to different angles of looking at the same problem.

Although navigating through such a highly diversified theoretical landscape remains challenging, there is hope that one

Received: 18 January 2021; **Revised:** 27 May 2021; **Accepted:** 18 August 2021

Highlights

- Models of consciousness constitute hypotheses on how consciousness relates to the physical domain.
- Numerous competing frameworks build on different philosophical grounds and favour particular scientific methodologies.
- We compare models based on three analytical dimensions: mode of explanation, mechanisms of explanation and target of explanation.
- Differences in the conception of explanation have empirical consequences: the futility of localization approaches, the potential of second-person methodologies and different senses of 'content' of consciousness.
- Future research needs integrative approaches, where mathematical and axiomatic models may play a critical role.

could account for the large variation in the field by using only relatively few axes of comparison. To this date, there have been only a couple of systematic attempts to thoroughly compare the contemporary models of consciousness (Klink *et al.* 2015; Block 2009; Northoff and Lamme 2020). Arguably, however, all these endeavours have not gone beyond simply collating the theories' different explanatory targets and their main employed paradigms. In addition, the way the theories have been classified so far [such as targeting either 'phenomenal' or 'access' consciousness (Northoff and Lamme 2020) or whether they are related to the 'pre-' or 'post-'stimulus neural activity] might have created more confusion than elucidation (Rosenthal 2002a; Block 2007; Rosenthal 2020).

In order to identify the most critical points of contention in the field, we first introduce the philosophical landscape (see The problem of consciousness Section), then we explicate three crucial directions in which frameworks diverge most noticeably with respect to their explanatory pretence (see Method of classification Section). Among these directions, we distinguish mode of explanation (mechanistic vs. unificationist), mechanism of explanation (functional vs. causal) and target of explanation (quality vs. quantity of consciousness). These dimensions are discussed along contemporary models of consciousness, which creates a multi-dimensional explanatory profile for each. We also discuss three empirical consequences of this classification (see Classifying models of consciousness Section). We conclude by advocating for a more integrative approach, hinting at already existing empirical and theoretical (mathematical) tools (see Conclusions: integrative methods for the future Section).

We narrow our scope to 14 influential models. While being aware that our selection of models is not exhaustive, we do hope that our work would still spark meaningful discussions and inspire orderly and structured comparisons in the field.

The problem of consciousness

Philosophy does not solve problems; it rather helps to emphasize and re-conceptualize them, making problems amenable to scientific investigation. Therefore, we first look at some of the main concepts relevant to the question of consciousness, which are often implicitly held by current scientific theories.

'What-it-is-like' to have an experience

The idea that consciousness comprises an inner subjective feeling poses perhaps the biggest challenge to any model of consciousness. Explanatory projects in the contemporary neuroscience of consciousness thus often seem to target the notion of 'phenomenal consciousness' that goes back to work by Thomas Nagel from the 1970s, emphasizing the 'what-it-is-likeness' of conscious experience. Nagel argued that the purely objective study of an entity, such as the one science provides, does not allow any inference about the subjective character of being such an entity. This has sometimes been misconstrued as the claim that a purely subjective phenomenon such as consciousness cannot be studied at all within the objective framework of science and should best be left alone, an idea which has been forcefully disputed by John Searle (Searle 1998; Searle 2000).

However, the position that phenomenal consciousness should be the end-all of any scientific theory of consciousness has been contested. Most prominent here is the illusionism proposed early on by Daniel Dennett (Dennett 1988; Dennett 1991), which got traction recently but still comes with its own issues (Frankish 2017; Chalmers 2018; Dennett 2018). A frequent criticism of such illusionist approaches is the charge of being anti-realist about consciousness. Yet, this charge can be finessed (Fallon 2020). What is typically denied by illusionists is the existence of some non-physical essence of conscious experience – the seemingly qualitative nature of experience – but not the reality of experience itself. This can, e.g., be illustrated with the 'multiple drafts model' of consciousness put forward by Dennett (Dennett 1991). If the illusionist claim is understood in this weaker way, the majority of neuroscientists would likely agree. However, it is then not clear how this position would contribute to a scientific understanding of consciousness, effectively arguing against a philosophical straw man.

Another related distinction is the one between the notions of 'phenomenal' and 'access' consciousness (Block 2005; Block 1995). The former refers explicitly to its what-it-is-likeness, sometimes assumed to correspond to a 'minimal sense' of conscious experience without necessarily requiring reportability (Metzinger 2020; Changeux 2006). In particular, some theories of consciousness claim to be about exactly this phenomenal aspect and thus carry a distinct explanatory pretence. The latter notion, 'access' consciousness, corresponds to centralized availability for processing of information and the reportability of a conscious experience. It also refers to phenomena that are closely related to consciousness in other aspects (e.g. attention or meta-cognition). It is an open debate in the scientific study of consciousness whether an explanation purely in terms of access consciousness is truly satisfactory or whether it is not in fact the 'only' scientifically rigorous approach (Naccache 2018; Fazekas and Overgaard 2018).

Much of the prominence that phenomenal consciousness received as a potential scientific topic can be traced back to the work of David Chalmers who introduced the notion of the 'hard problem of consciousness' (Chalmers 1995a), the difficulty to explain why certain forms of physical information processing should feel like anything at all or how physical and phenomenal facts are related to each other. Chalmers argued against the reducibility of consciousness and initially advocated a 'natural dualism' (Chalmers 1997). Natural dualists assume that reality is composed of two sets of (irreducible) properties. Chalmers, moreover, assumed that those refer to two properties of 'information', and that the functional (psychological) structure of information

processing in the brain is mirrored in the phenomenal structure of consciousness. Chalmers later made the case for 'panpsychism' (Strawson 2006; Chalmers 2013a; Goff 2019), the view that consciousness is ubiquitous in nature and serves as the irreducible 'intrinsic nature' that grounds physical properties. Panpsychism is related to 'dual-aspect monism' that considers consciousness and physics merely as two aspects of a single underlying reality (Atmaspacher 2014). Some dual-aspect monists are panpsychists, although not necessarily so [other dual-aspect monists assume the single underlying reality to be neither of mental nor material but of 'neutral' nature (Stubenberg 2018)].

An alternative way to approach the problem of consciousness is the 'biological naturalism' of John Searle (Searle 2000). While Searle acknowledges the phenomenal character of consciousness, he finds it 'obvious' that it emerged from the brain similar to the way bile is produced by the liver. However, to date, no viable mechanism has been identified for this process. The problems for Searle's approach are representative of all materialistic or 'physicalist' approaches to consciousness that are discussed later in the paper, at least where they pretend to shed light on phenomenal consciousness. The majority of approaches in the scientific study of consciousness endorses the position of physicalism.

Another monistic response is to invert the hard problem and argue that the physical world is a product of consciousness. Many idealists, roughly, state that matter exists only insofar as it is represented in consciousness – although idealism in fact comprises a set of many different (heterogeneous but related) views. The idealistic position was once the dominant world view in much of Western and Eastern culture, but received a massive blow in the 20th century. Most philosophers and scientists do not take this option seriously anymore, although the climate seems to change with more and more scholars advocating such a view (Marshall *et al.* 2001; Hoffman 2008; Hoffman and Prakash 2014; Kastrup 2017; Chalmers 2019).

Finally, one could try to bracket metaphysical issues in the study of consciousness and instead follow an approach advocated by Francisco Varela. Whereas Chalmers postulated that it seems as if there is a need for 'extra ingredients' to physical theory and whereas the various monisms express certain metaphysical assumptions, Varela suggested to regard consciousness and brain processes as 'mutually constraining phenomena' (Thompson 2007; Rodríguez 2008; Varela 1996) that ground an 'empirical' approach to consciousness. More generally, Varela's 'neurophenomenology' is an adaption of an earlier continental approach [known as 'phenomenology' (Gallagher and Zahavi 2008; Kaufer and Chemero 2015)] to cognitive neuroscience that seeks to uncover the necessary structures of all experiences (including the ones that give rise to scientific knowledge). Neurophenomenology thereby suspends any metaphysical judgement if possible, thus realizing a methodological desideratum of earlier phenomenologists, who intended to go 'back to the things themselves', i.e. back to experience. Whether or not consciousness is in fact an emergent phenomenon, this leaves unanswered the question how explanations in terms of brain dynamics and conscious experience mutually constrain each other, e.g., to what extent consciousness is able to influence its physical substrate (Thompson and Varela 2001) or how any supposed 'backreaction' would manifest itself in scientific data.

All these proposals have their own problems, even though they come in very different guises. Illusionists need to explain why the illusion of consciousness appears as something real and vivid [the 'illusion problem' (Frankish 2017)] and argue that there is in fact no problem of consciousness over and above the problems of 'access' consciousness (problems which still need to be solved though). Alternatively, one would need to specify the reductive relation between 'phenomenal' consciousness and matter (e.g. solve the 'hard problem' and give a model of how consciousness actually emerges from the brain) or explain the causal efficacy of conscious experience.

From philosophy to scientific models of consciousness

A framework for the study of consciousness refers to a group of premises and assumptions to guide experiments and interpret results. More specifically, a model of consciousness conveys concrete hypotheses, predictions, mechanisms, and explanations of the associated phenomena. A proper theory for consciousness consists, by contrast, of a set of explicit (and often formalized) systematic premises plus a concrete model to enable the testing of (empirical or theoretical) predictions and eventually its implementation and manipulation. The approaches discussed in this paper correspond to models that mostly operate on implicit assumptions. Even though they do not resemble proper theories, in the sense just outlined, we will use both the terms 'theory' and 'model' interchangeably in the remainder of this article to better conform to the literature.

A first step to better understand models of consciousness is to make explicit their underlying philosophical assumptions. These assumptions inform and influence models of consciousness. In Fig. 1, we summarize the main relationships between philosophy and early and modern models of consciousness.

In this article we consider early models such as the mechanistic model of Crick and Koch (CK) (Crick and Koch 1998; Crick and Koch 2003), Dynamical Core (DC) (Edelman 2003), Multiple Drafts Model (MDM) (Dennett 1991), Orchestrated Objective Reduction (OrchOR) (Hameroff and Penrose 2014), Global workspace (GW) (Baars 1988; Baars 2005), Thalamo-Cortical loops and Sensorimotor Couplings (TCL) (Llinas *et al.* 1998; Llinás 2003), the dualist proposal by Beck and Eccles (BE) (Beck and Eccles 1992) and Enactive and Radical Embodiment (ERE) (Varela 1996; Thompson and Varela 2001; Varela *et al.* 2001; Lutz *et al.* 2002). Modern models are Global Neuronal Workspace (GNW) (Dehaene and Naccache 2001; Dehaene and Changeux 2011), Higher-Order Thought Theory (HOT) (Rosenthal 2002b; Rosenthal 2008), Recurrent Processing Theory (RP) (Lamme 2003; Lamme 2010; Fahrenfort *et al.* 2017), Predictive Processing and Interoception (PP&I) (Wiese and Friston 2021; Friston *et al.* 2020), Integrated Information Theory (IIT) (Oizumi *et al.* 2014; Tononi *et al.* 2016; Haun and Tononi 2019), Attention Schema Theory (AST) (Graziano and Kastner 2011; Graziano *et al.* 2020; Webb and Graziano 2015), Conscious Agent Networks (CAN) (Hoffman and Prakash 2014; Fields *et al.* 2018) and Temporo-spatial Theory of Consciousness (TTC) (Northoff 2013; Northoff and Huang 2017).

Early models typically inform later ones, sometimes via direct succession (e.g. GW/GNW), or via integration of concepts developed in previous models (e.g. TTC). Models may also initiate dialogues and remain under dynamic influences with each other. For example, both postulates of GNW and IIT do remain consistent to some extent with mechanisms put forward by the predictive coding approach (PP&I). Sometimes, the relation between different models or theories has more of an implicit nature. For example, IIT is sometimes thought to express some form of panpsychism,

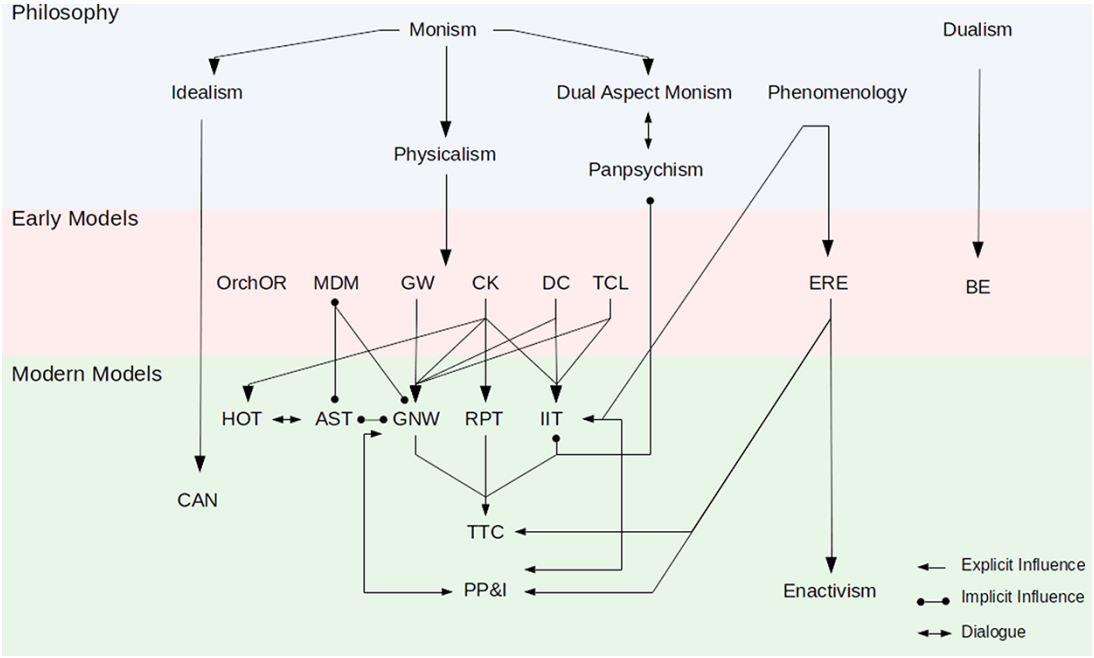

Figure 1. Philosophy and models of consciousness. Relationships and influences between philosophy and early and contemporary models of consciousness.
Acronyms stand for: Crick and Koch (CK), Dynamical Core (DC), Multiple Draft Model (MDM), Orchestrated Objective Reduction (OrchOR), Global workspace (GW), Thalamo-Cortical loops and Sensorimotor Couplings (TCL), Beck and Eccles (BE), Enactive and Radical Embodiment (ERE), Global Neuronal Workspace (GNW), Higher-Order Thought Theory (HOT), Recurrent Processing Theory (RP), Predictive Processing and Interoception (PPI), Integrated Information Theory (IIT), Attention Schema Theory (AST), Conscious Agent Networks (CAN), and Temporo-spatial Theory of Consciousness (TTC).

Table 1. Axes for a three-dimensional classification map with two opposite directions per dimension (represented by positive or negative values). These axes were later used to score models of consciousness, see details in section Supplementary data: Scoring Method.

Axes	First direction (+)	Second direction (–)
Mode of explanation	Mechanistic	Unificationist
Mechanism of explanation	Causal	Functional
Target of explanation	Quality	Quantity

although this is done rather implicitly, inferred via analysing IIT's basic premises.

Method of classification

Our initial examination of the current models of consciousness involves placing every theory within three orthogonal dimensions corresponding to the positions they espouse with regard to explanation. The first axis stands for the 'mode of explanation' assumed by the model (mechanistic vs. unificationist), the second for the 'mechanism of explanation' (functional vs. causal) and the third for the 'target of explanation' (quantity vs. quality of consciousness). This classification results in 'explanatory profiles' for each theory, which accounts for a substantial amount of variance in the theoretical landscape.

Importantly, the mode of explanation is not to be conflated with a mechanism of explanation, i.e. while the mode informs about the ultimate 'aim' of the explanation (what constitutes a satisfactory answer to the 'why' question), the mechanism demonstrates which particular 'tool' can get us towards that predefined aim (i.e. the 'how' question).

Mode of explanation

One of the most important hallmarks of a successful theory is its explanatory power. Nonetheless, the very notion of explanation (what it means to successfully 'explain' something), despite its deceptive simplicity, can be unpacked as highly heterogeneous (Nagel 1961; Salmon 1990; Woodward 2019; Woodward 2013; Woodward 2004; Strevens 2004; Colombo 2017; de Regt and Baumberger 2019). In particular, the very notion of 'explaining consciousness', although ubiquitous in the literature, constitutes a deceptive 'umbrella term', under which one can identify strikingly different theoretical goals. That 'consciousness' is frequently used as an umbrella term for various mental phenomena has been often stated in the literature [e.g. (Van Gulick 2018)]. Here we wish to emphasize that the same is true for the seemingly innocent notion of its 'explanation'. The field of consciousness science has been flooded with numerous works contemplating whether the full explanation of this phenomenon is even feasible (McGinn 1988; Chalmers 1995b; Hohwy and Frith 2004). Therefore, we focus on two alternatives that seem particularly relevant in the domain of consciousness studies: 'mechanistic approaches' and 'unification' (Salmon 1990; Strevens 2004).

A mechanistic approach posits that a particular occurrence is explained once we demonstrate how it 'fits into' the presumed spatio-temporal causal structure of the world, i.e. to explain an event is to identify its cause (Reutlinger 2017; Salmon 1990). This is typically achieved via referring to the transmission of a counterfactual and local modification (a 'mark') of the associated physical, typically spatio-temporally continuous, process (Woodward 2004). Mechanistic accounts typically subscribe to a realist notion of causation that grounds the explanation, prediction and other inferences (Nathan 2020). A theory of consciousness that holds the mechanistic view would accordingly imply that the

satisfactory explanation of subjective experience requires meticulous unfolding of the chain of causes and effects occurring in the nervous system that leads to a conscious experience of some kind. Such a position imposes a strictly empirical agenda, investigating which neurophysiological events precede and 'give rise' to a conscious experience (Craver 2007). Importantly, the driving force behind such accounts is a philosophical 'assumption' that such a chain of causes is 'independently given' at some objective level of description at some spatio-temporal scale (e.g. of the brain). Naturally, these frameworks tend to cluster together under branches of ontological or methodological reductionism.

On the other hand, the unificationist stance seeks to provide a 'unified account' of a range of different phenomena or laws (Glymour 1980; Kitcher 1981), given causal or non-causal explanations (Reutlinger 2017; Salmon 1990; Batterman and Rice 2014), which were previously thought to be unrelated – or related in a mysterious or seemingly arbitrary way (core historical examples are Maxwell's unification of electricity and magnetism or Newton's unification of terrestrial and celestial motion). Explaining consciousness under the unificationist framework would give priority to demonstrating how the phenomenon of consciousness is embedded into a parsimonious, coherent framework. As a result, the unificationist seeking a satisfactory explanation would be inclined to associate less with empiricism, and more with formal tools, mathematics and non-reductive philosophy. On the other hand, unificationist explanations have been argued to be most relevant for physics but at odds with biology and neuroscience (Anderson and Chemero 2013; Bayne 2018).

While mechanistic accounts, implicitly or explicitly, always assume the realist notion of causation in space and time, unificationist accounts postulate that explaining consciousness cannot be exhausted by studying the spatio-temporal chain of causes and effects localized in the brain. In particular, assumptions about the causal order and spatio-temporal descriptions are often being put on hold and treated as an 'explanandum' (a phenomenon to be explained) rather than as 'explanans' (the grounding of explanation) (Barnes 1992). Causes, at this point, can be regarded as explanatory postulates or theoretical hypotheses which do not exist independently of the explanations that describe them (Nathan 2020; Barnes 1992). Specifically for the science of 'consciousness', it should also be made clear how or why consciousness is related to the explanation of such causal chains (e.g., by arguing that causal orders exist only 'within' consciousness).

Numerous examples show that both types of explanation have contributed to significant progress in science (Barnes 1992; Woodward 2004). Our aim is not to assess which approach produces better explanations, but to simply recognize that different frameworks of consciousness will be inclined to differ already on this very basic assumption.

Mechanism of explanation

We define the term 'mechanism' after Illari and Williamson as 'entities and activities organized in such a way that they are "responsible" for the phenomenon' (Illari and Williamson 2012). This way of defining the term is independent from any specific theory of causation. Numerous examples exist where explanations can be purely mathematical, the role of causation being denied or simply ignored (Batterman and Rice 2014; Reutlinger 2017).

In the context of consciousness science, an increasingly popular division, introduced by Doerig *et al.* (Doerig *et al.* 2019) and specifically addressing different kinds of mechanisms, distinguishes explanations as either 'functional' or 'causal'. The stance of functionalism primarily states that consciousness can be generated as long as a particular function is realized (Block 1996), without any specific constraint on the exact causal machinery behind it. In principle, any system may become conscious as long as it executes the functions associated with conscious experience. On the opposite side, causal theories do not necessarily unsubscribe from the view that a particular function might be typically associated with consciousness, but the burden of explanation is placed on 'how' such a function is implemented: establishing what and how elements of a system interact. These theories seek relevant causal relationships within the model's target system(s) (Batterman and Rice 2014). In those terms, causal models support the idea that only a system with the right causal relationships will lead to conscious experience.

One could further illustrate this distinction by introducing the idea of a structure-preserving map (M) between two objects. Causal models insist that the causal system's structure (S) explains consciousness, $M : S \rightarrow C$, where the arrow refers to a mapping that preserves system structure (whether or not a particular and objective spatio-temporal structure is the most relevant for analysis). Functional models, however, would argue that the functional structure of a system (F) explains consciousness, $M : F \rightarrow C$, and thus it is the function that is preserved by the arrow. These two types of models appear exclusive, in the sense that S and F are different objects. However, if we focus on the nature of the map, independently of their objects, the difference becomes a question of 'degree'. Causal theories usually assume there is only one way to preserve the causal structure and the phenomenology of subjective reports. This is tantamount to postulating an isomorphism between domains (there exists only one such arrow) (Tsuchiya *et al.* 2016), whereas functional theories would claim/assume that there are multiple ways (arrows) to preserve the function and therefore giving rise to consciousness, i.e. M', M'', M''', etc.

Superficially, mechanistic accounts and causal models seem to overlap, as both axes refer to the vocabulary of causation. To avoid any potential confusion or apparent contradiction, it is critical to recall our distinction between the 'aim' and 'tool' of explanation (which could be referred to as 'metaphysical' and 'pragmatical' commitments of the theory, respectively). While mechanistic explanations subscribe to the realist notion of causation, causal models do not necessarily do so and could merely 'use' causes (rather than functions) to unfold the relevant interactions within the target system. In principle, they could still maintain that causes are not 'the only' explanandum, in the sense that explaining them still would not suffice to explain consciousness. Thus, a model can be non-mechanistic and causal at the same time without contradiction.

Target of explanation

Lastly, the target of an explanation corresponds to the aspects of a particular phenomenon that scientists intend to explain. The most basic distinction that emerged over the last decades is the one between 'quality' and 'quantity'. The quality of consciousness is what makes consciousness feel 'the way' it does (cf. also 'What-it-is-like' to have an experience Section), and the quantity corresponds to what makes the system conscious rather than unconscious. A model targeting quality should therefore account for why any stimulus should feel a 'particular way' and what makes an experience spatial, visual, auditory, painful or temporal, while a model targeting quantity intends to account for global markers differentiating conscious vs. unconscious systems.

A full-fledged theory of consciousness needs to explain both the quantity and quality aspects of subjective experience. Several theories tend to focus exclusively on global markers differentiating conscious vs. unconscious systems, and the problem of quality in such cases is delegated to the external world (i.e. sensory cues feel the way they do solely because they 'carry' their quality from the environment, or stimulate 'correct', labelled receptors). However, there are many, conceptual discussions, each pointing to a specific problem of such a 'delegation'; see arguments on brains in vats (Horgan *et al.* 2004), inverted spectrum (Shoemaker 2000), actual cases of perceptual variation (Block 1999), the fact that sensory characters correlate much more with neural patterns than anything else in the external world (Pautz 2014), dramatic perceptual alterations in psychedelic experiences (Bayne and Carter 2018), perceptual illusions, the generic non-preservation of phenomenal and environmental structure (Prakash *et al.* 2020), etc. A much more challenging project would strive to explain why an experience feels the way it does, e.g. based on the internal architecture and dynamics of the brain, without referral to the external world as something that 'stores' or 'produces' any qualities [but see (Graziano *et al.* 2020)].

Classifying models of consciousness

According to the analytical dimensions defined in the previous section, we now classify and discuss selected models of consciousness in a three-dimensional map (Fig. 2). This classification intends to illustrate our framework. It is a provisional attempt based on quotations and a rough score, which provide support to our discussion. We have first focused on reviewing the relevant articles that describe the selected theories. Consequently, we have scoured representative works for statements pinpointing what a satisfactory explanation would look like according to each theory. We have then determined numerical values along the dimensions specified in the previous section (see sections Supplementary: Relevant Citations, and Supplementary: Scoring Method).

Mode of explanation

Based on our classification, we unpack the explanatory profile of each theory. We have found individual discrepancies, with most of the theories leaning clearly towards one or another mode. Some frameworks, including GNW and HOT, gave us consistent clues to classify them under the mechanistic cluster. A statement that would drive us to the mechanistic classification could be, e.g., the following statement by Rosenthal, a proponent of HOT: 'we understand something only when we can explain it, and explaining a natural phenomenon typically if not always means locating it in its distinctive causal nexus' (Rosenthal 2008), or by Dehaene, a proponent of GNW: 'tools of cognitive psychology and neuroscience may suffice to analyze consciousness' (Dehaene and Naccache 2001). On the other hand, unificationist propensities are more apparent in frameworks such as IIT: 'IIT provides a "principled explanation for several seemingly disparate facts" about the PSC [Physical Substrate of Consciousness]'; or PP: '… it could unify existing approaches under a single overarching principle (i.e., the FEP)' (Wiese and Friston 2021). Such unifying tendencies are also expressed by proponents of AST, who aim for a 'standard model of consciousness' (Graziano *et al.* 2020) by taking a deflationary stance on the reality of consciousness or, on the other extreme, by proponents of CAN: 'if we want to go beyond this "applied science" and understand the true nature of the mind and the reality beyond it, we can't look to neurons or brains' (Fields *et al.* 2018). The full list of relevant quotes, justifying each model's classification, can be found in Supplement Table S2.

This first dimension clearly constitutes a parsimonious dividing force, introducing tensions between the models' individual goals. Arguably, identifying which mode of explanation a theory supports can also predict how a certain theory would pragmatically proceed in its investigations. Although both groups of theories would not deny the relevance of empirical research, the mechanistic models would take on a cautious 'bottom-up' approach, withholding the drawing of any firm conclusions without substantial accumulation of incoming data (Michel *et al.* 2019). Unification accounts, on the other hand, would be much more prone towards a 'top-down' approach, giving priority to those empirical predictions which can make sense of seemingly disparate phenomena from the perspective of the framework itself.

Mechanism of explanation

There is a marked difference in which explanations can provide explanatory power, by either concentrating on overall functional roles or by describing network parts and the interactions between them. Pertinent examples for functional theories are HOT and GNW, according to which, consciousness can arise in a physical system as long as it realizes 'meta-representation' or 'global broadcasting', respectively. On the other hand, models such as IIT and RPT lean towards elucidating the structure of causal interactions at the level of network analysis. For a full list of relevant quotes disclosing either functional or causal inclinations in explaining consciousness, see Supplement Table S3.

Although we believe that most theories do explicitly or implicitly differ in these explanatory assumptions, Fahrenfort and van Gaal point out that 'most' empirical theories would eventually aim at explanations involving causal implementations, rather than functions (Fahrenfort and van Gaal 2021). In line with that view, we have indeed come across many examples in which models of consciousness, even those labelled as functional, in actuality resort frequently to the language of 'causal interactions'. For example, as stated by Rosenthal: 'On the HOT hypothesis, a conscious state is a compound state, consisting of the state one is conscious of together with a HOT. So the causal role a conscious state plays is actually the interaction of two causal roles: that played by the state itself and that played by the HOT' (Rosenthal 2008). Another relevant instance could be predictive processing theory (PP). Despite having been classified as functional in (Doerig *et al.* 2019), one variant of PP clearly states that implementing the adequate computational/functional principles is only a necessary, but not sufficient condition for consciousness. As explicitly mentioned by Wiese and Friston (Wiese and Friston 2021), computations need to be physically instantiated in the right architecture and not all virtual machines that realize approximate Bayesian inference should be considered conscious. The reason they give is that a Markov blanket of the physical system must be based on the system's dynamics, and the dynamics strictly depends on the system's structure. That drives us to conclude that a popular framework has recently moved significantly away from functionalism and can now be classified as a causal theory, although different interpretations might exist [e.g. (Clark 2016)].

Additionally, some models seem difficult to classify as either functional or causal in the above sense. For example, TTC aims to accommodate both functional and causal types of explanations, making it difficult to identify its commitments (although

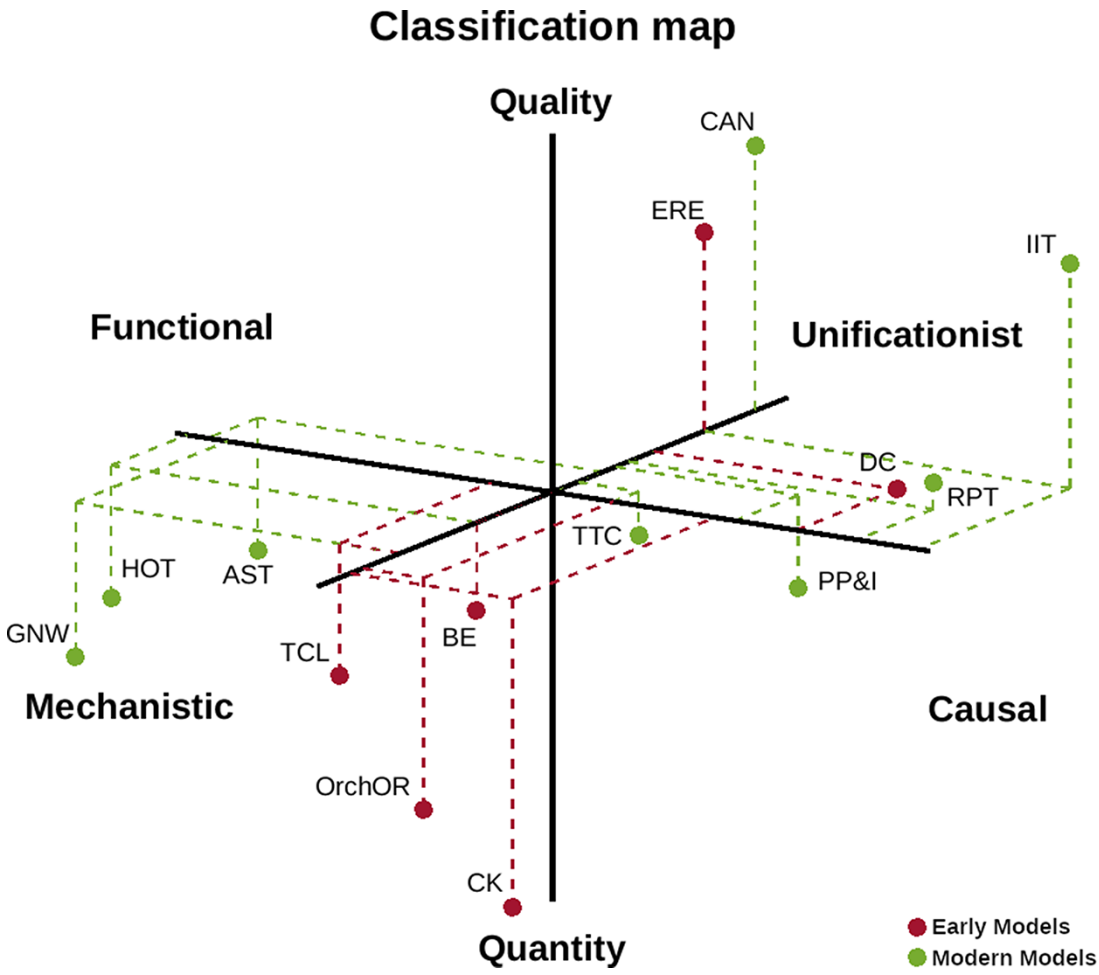

Figure 2. Models Classification. The selected models inhabit different explanatory spaces according to our three analytical dimensions. After scanning the literature, summarizing and discussing main tendencies in light of three-axis classification in Table 1, each author independently scored models between 1 and 3: 1 corresponds to a slightly commitment to the direction of the axis and 3 a strong commitment. 0 value was used any time that the model does not qualify for the + or – direction. This quantification is provisional; it only intents to illustrate our framework and it is not based on statistical data (see details in section Supplementary data: Scoring Method).

according to some key quotes it seems to lean towards causal approaches). Other models such as ERE, CAN and BE are neither functional nor causal in the sense of Doerig *et al.* (Doerig *et al.* 2019). This is because they do not share the underlying philosophical assumptions typically made by neuroscientific theories of consciousness (cf. 'What-it-is-like' to have an experience Section). ERE assumes a form of dynamical co-emergence, i.e. consciousness co-arises with the system in such a way that there is a contextual constraint between the biological system (living body) and the experience (lived body), making both interdependent (Thompson 2007). In other words, according to ERE consciousness does not only emerge from the brain and the body, but it also actively constrains them, creating a causally reciprocal relationship between neural events and experience (Thompson and Varela 2001). Lastly, other non-trivial cases include CAN (consciousness in this case is fundamental and the physical realm is what emerges from it) and BE (which is premised on dualist).

Target of explanation

The final dimension corresponds to the target of explanation, i.e. quality vs. quantity of consciousness. We analysed relevant papers looking for statements revealing 'what' the theory aims to explain. Theories that focus mainly on the contrast between wakefulness and other impaired conditions (e.g. TCL) or on what neural activity might underlie the transition from the stimulus being 'unseen' to 'seen' (e.g. GNW, HOT and CK) can be classified as theories of quantity. We encourage the reader to notice that the 'why' question posed in case of such theories is not 'why a stimulus feels this way', but rather 'what constitutes the switch rendering content visible'. The stimulus here can be replaced with any other stimulus, as the quality (although not necessarily the structure) of content remains irrelevant and out of the explanatory target. Importantly, answering the first question leads us to 'phenomenal content', while the second question corresponds to 'access content'.

An attempt to explain the quality rather than sole quantity is exemplified by IIT. IIT tries to identify the mechanisms behind the phenomenal character of experience, i.e. the quality of its content ('what makes an experience visual, auditory, colourful, painful?'). In that vein, it poses that the quality of consciousness is in one-to-one correspondence with the geometry, concepts and relations encapsulated by the Maximally Irreducible Conceptual Structures (Haun and Tononi 2019). IIT predicts that the contents of consciousness are entirely specified by the internal workings of elementary mechanisms of the main complex. Notably though, IIT's agenda to target quality is still in its nascent stages and has

not gone beyond trying to explain the spatiality of experience yet (Haun and Tononi 2019).

Other models, to a greater or lesser extent, also strive to address quality. One example is ERE. In this theory, first-person reports inform the interpretation of neural signatures, thus co-determining the actual explanandum of the theory (Lutz et al. 2002; Lutz and Thompson 2003; Lutz et al. 2015; Windt et al. 2016; Petitmengin et al. 2019; Poletti et al. 2021). In one particular example, experienced meditators were able to wilfully influence neural activity (both short and long term) (Lutz et al. 2004). This reveals a more complex relationship between experiential content and its neural substrate: quantitative measures constrain experienced qualities, but quality might in turn constrain quantity (Varela 1996). Another example, RPT, largely focuses on questions related to phenomenal awareness. There are appreciable ways in which it can explain some qualitative aspects of consciousness, such as the figure-ground segmentation, i.e. how one might sometimes perceive a texture to be a surface that lies on top of a background (Roelfsema et al. 2002; Scholte et al. 2008; Fahrenfort et al. 2017). Functional models also claim to target quality, however in a different way, typically using recognition and detection paradigms. For example, HOT, although it tends to delegate 'qualities' to largely unspecified mechanisms of low-level representation or the external world, puts forward explanations why we might sometimes 'feel confident' of the perceptual experience (Brown et al. 2019)).

It is, therefore, evident that different theories adopt different strategies in addressing the problem of quality, depending on their underlying philosophical assumptions. Arguably, these discussions might pave the way for other theories' proponents to appreciate different kinds of explanatory aims and recognize that a mature theory of consciousness should address both the quantity and quality of experience. Novel questions inspired by the focus on quality for theories of quantity could be, *inter alia*: How exactly does the 'pattern' of global ignition inform phenomenal content? What distinguishes visual or auditory re-entrant processing, or in other words, what mechanism makes these two modalities 'feel' differently? Notice that circumventing the problem by referring to the function of the primary sensory cortices (e.g. pre-labelled as visual or auditory) is not tenable, because it does not inform 'what precisely makes' a primary cortex visual or auditory in the first place. Alternatively, for models such as AST, what, on the level of modelling or re-representing, 'specifically' correlates with the (illusory) attribution of qualia?

Empirical consequences

The fact that different theories subscribe to different modes, mechanisms and targets of explanation has several empirical consequences. One of them pertains to the localization of neural correlates of consciousness (NCCs) (Boly et al. 2017; Klink et al. 2015). First off, there is a substantial disagreement between the so-called first order (e.g. IIT and RPT) and higher-order theories (e.g. HOT) insofar that the former typically assign NCCs to the early sensory cortices while the latter to the frontal-parietal network. It is worth pointing out that the 'frontal theories' tend to systematically cluster under the mechanistic mode of explanation, often interested in questions of quantity rather than quality (i.e. under what conditions a particular stimulus would be classified as consciously perceived or not) (see Fig. 2). Notwithstanding these distinctions, most of the models reviewed above are dynamical global network approaches and therefore non-localist in principle (Dehaene and Naccache 2001). The network is relevant; the nodes alone are not. For example, GNW is often misunderstood as a fixed architecture encompassing the fronto-parietal cortices. In actuality, it comprises dynamic neural contributions that define this hypothetical global network (Dehaene and Changeux 2005). The only anatomical constraint is that relevant regions should be connected by long axons of pyramidal neurons. During decades of research, GNW identified pyramidal neurons in Layers II and III as candidate mechanisms. In light of recent relevant evidence (Suzuki and Larkum 2020), they have put forward Layer V as a more likely GW substrate (Mashour et al. 2020). On the contrary, IIT postulates that the true NCC lies in the posterior hot zone; one of the main points of contention with GNW being the counterintuitive role of inactive units, contributing to the cause–effect information just as the active ones do (Oizumi et al. 2014; Siclari et al. 2017). HOT seems not to specify network mechanisms in sufficiently rigorous details. Even though the original framework has been worked out empirically (Lau and Passingham 2006; Brown et al. 2019), this family of models leave a lot of space to accommodate ambiguous experimental data.

Other inconclusive findings include the study of posterior hot zone involvement in dreams (Siclari et al. 2017; Mashour et al. 2020) and Perturbational Complexity Index (PCI) (Casarotto et al. 2016). In the first case the reduction of low-frequency activity in posterior zones of the brain correlates with dreams during rapid-eye-movement (REM) sleep and non-REM sleep (Siclari et al. 2017). At the same time, content-specific dreams involve high-frequency activity in the frontal and prefrontal cortices. Some authors interpret these results as evidence of posterior zones for phenomenal consciousness and support of IIT, while others view them as a clear evidence of the role of prefrontal cortex and GNW (Mashour et al. 2020). In the second case, the PCI index inspired by the IIT framework seems compatible with the global ignitory activity of GNW (Mashour et al. 2020), as well as with other models such as ERE. Another recent study inspired by GNW found that the thalamic nucleus of monkey brains under deep brain stimulation restores signatures of consciousness and reactivate nodes of the GNW that remain inactive under anaesthesia (Tasserie 2020). However, as the authors also pointed out, these results are also compatible with thalamo-cortical loops theories and IIT.

Therefore, the postulation of either prefrontal regions and posterior regions to be the primary locus of the NCC does not by itself speak to different explanatory approaches and probably merely reflects the availability of experimental techniques and methodological choices (Yaron et al. 2021). Moreover, the question of localization makes sense only if one endorses a mechanistic and causal explanatory framework simultaneously, which is only the case for CK and OrchOR models. If one, by contrast, looks for unification or function, then the fact that, say, frontal or posterior regions are most relevant for the NCC, is contingent. This is particularly true for those views which do not adhere to the explanatory primacy of causal chains within space-time (e.g. CAN). As such, experiments that try to identify the NCC do not provide evidence for or against such a theory. But even bracketing questions of explanatory mode, localization approaches are problematic: evidence for some mechanisms at any particular level does not falsify the relevance of other mechanisms at other scales. An interesting example is a controversial recent comparison between IIT and GNW at the single-unit level. At first glance, results suggest that GNW is supported by the evidence, while IIT is not (Noel et al. 2019). However, considering the active single-unit level as the optimal spatio-temporal scale for testing IIT remains problematic. This would force an assumption of GNW (i.e. that

consciousness is correlated to only active neurons) onto IIT, which is *prima facie* not warranted. But this is a conceptual issue having to do with explanatory commitments – and not an empirical one.

Another empirical consequence of the way a theory emphasizes either quality or quantity has implications for the methodology used. In general, it has been long recognized that the scientific study of consciousness utilizes first-person and third-person approaches (Olivares *et al.* 2015). The former includes subjective reports and phenomenological interviews (Chalmers 2013b) and the latter refers to objective measures of physical states, using different techniques such as electroencephalography, functional magnetic resonance imaging and magnetoencephalography, among others. Most models of consciousness claim to employ both subjective and objective accounts. However, their assigned importance varies across the models. In particular, those theories which seek to explain quality will wish to utilize a method that is specifically suited to make the qualitative aspect of experience precise. Particularly promising accounts are called 'second-person' methods, referring to interview techniques that incorporate verbal and non-verbal reports in order to obtain a well-informed subjective report (Olivares *et al.* 2015; Lutz *et al.* 2015; Petitmengin *et al.* 2019). This approach is motivated by earlier research in neurophenomenology (Thompson 2004; Varela 1996; Lutz *et al.* 2002). The second-person method is different from the first-person method in that the former is guided by an interviewer who reads and interprets various indicators from the first-person subjective report. Given these indicators, the interviewer is able to ask more refined questions that force a subject to closely specify her reports.

Related to the quality versus quantity distinction is the one between levels and contents of consciousness (Bachmann and Hudetz 2014; Bayne *et al.* 2016; Storm *et al.* 2017). Levels of consciousness convey global signatures of consciousness, from which different paradigms contrast awake neural activity against non-awake or disrupted conditions such as sleep, chronic disorders of consciousness and anaesthesia, among others (Signorelli *et al.* 2021a; Signorelli 2021). By contrast, paradigms looking for contents of consciousness survey conscious experiences through contrasting perceptual analysis (perceived vs. unperceived) and multiple psychophysical reporting paradigms. Examples include masked stimuli, high-contrast figures, binocular rivalry, flash suppression, motion-induced blindness and attentional paradigms, among others (Klink *et al.* 2015). However, one must not conflate the study of contents with the study of the specific phenomenology of such contents. Yet, approaches that study contents quantitatively might still be understood as studying the structure 'between' instances of (qualitative) consciousness. For example, it has recently been argued that postdictive studies in the auditory domain could uncover the temporal structure of perception (Herzog *et al.* 2020) – something that has previously been discussed extensively in the phenomenological literature. Thus, in general, it is more appropriate to distinguish between an 'access content' and a 'phenomenological content', the former allowing an indirect (often dynamical) inference of the structure of consciousness and the latter specifically targeting the 'what-it-is-likeness' of experience.

Conclusions: integrative methods for the future

In light of our previous sections, the empirical testing of models of consciousness is far from trivial. Comparing models of consciousness is not only difficult due to experimental limitations, but also due to the fact that different models operate on very different and mostly implicit assumptions about modes, mechanisms and targets of explanation. We thus introduced a classification scheme to make the different explanatory profiles of leading models explicit. To our knowledge, this is the first time that all these models and perspectives are reviewed systematically, focusing on different explanatory aspects and analytical dimensions, and organized in one single and comprehensible classification.

Our provisional classification framework serves as an invitation for theorists to weigh in on how their own models might be classified. In future attempts, a representative number of experts, performing statistics, principal component analyses and test–retest reliability should be added, among others. One might also speculate that new models will occupy empty spaces in our taxonomy and current models of consciousness will change their location as new versions of them emerge, either in light of new evidence or more refined theoretical discussions.

Having perused the relevant literature around 14 popular frameworks, we suggest that a number of disputes in the field of consciousness studies might stem from differently set explanatory goals and targets. Although both mechanistic and unificationist accounts have their advantages, there are certain areas of research where one of them might turn out to be more suitable than the other. Some types of explanations might require to step back from investigating the specific empirical details and look more at the overall mathematical structure: the reason 'why' a person cannot untie a particular knot may stem from a topological fact about the knot, rather than from a detailed causal trajectory illustrating the attempt of its disentanglement. Several examples show that a full-fledged explanation of a phenomenon might sometimes require more than a causal story (Reutlinger 2017). On the other hand, unificationist accounts face their own problems, such as the 'problem of asymmetry': if A explains B, B does not explain A. It is heavily contested whether unificationist approaches could accommodate this intuition (Barnes 1992).

It is yet to be established which type of explanation is most adequate for the science of consciousness or whether even different modes of explanation would be required [e.g. manipulationist accounts (Woodward 2004; Woodward 2019)]. Nonetheless, one of the messages of this paper is to emphasize that the first step towards a more mature science of consciousness is the recognition that the question of why brain activity is correlated to subjective experience can be understood in a variety of ways. We suggest that to fully explain 'what it means to be conscious', one needs to first be precise about what it would mean to 'explain' something.

Remaining aware of the disagreements within the field (and its early history), one might also try to extract what most models agree on. Some empirical approaches intend to follow this kind of pragmatism. For example, the concept of criticality in dynamical system theory shows to be compatible with both access and phenomenal consciousness signatures (Tagliazucchi 2017). The analytical method of connectome harmonics also aims to unify different signatures of consciousness, from a more general perspective of brain functioning and physical system theory (Atasoy *et al.* 2017; Atasoy *et al.* 2019; Luppi *et al.* 2020). Large-scale models using different anatomical, functional and molecular layers of description also present promising features to integrate different mechanisms at different scales (Kringelbach *et al.* 2020), as well as various signatures of consciousness (Signorelli *et al.* 2021a; Signorelli 2021). Recently, optogenetic experiments demonstrated that the

biophysics of pyramidal neurons in cortical layer V integrates two contentious mechanisms associated with consciousness, i.e. cortico-cortical loops and higher-order thalamo-cortical loops (Suzuki and Larkum 2020; Aru *et al.* 2020). More research aiming at synthesizing different findings is currently underway.

Recent mathematical works too have recognized the need for integration within a sound theoretical foundation. This new trend, 'mathematical consciousness science' (AMCS 2021), employs formal and rigorous methods to explore ways to distinguish various models and derive new empirical predictions. Some examples are the mathematical developments, based on IIT (Oizumi *et al.* 2016; Tsuchiya *et al.* 2016; Kleiner and Tull 2020) and mathematized phenomenology (Yoshimi 2007; Ehresmann and Gomez-Ramirez 2015; Prentner 2019; Signorelli *et al.* 2021b); other approaches are based on symmetry (Kleiner 2020), category theory (Northoff *et al.* 2019; Tsuchiya and Saigo 2020) or the compositionality of processes (Signorelli and Meling 2021; Signorelli *et al.* 2021c; Signorelli and Boils 2021; Signorelli *et al.* 2021b). Some models explicitly address different metaphysical starting points, such as idealism (Hoffman and Prakash 2014) or decompositional approaches of dual-aspect monism (Atmanspacher 2020). Common to all these approaches is that, inspired by the transparency of mathematics, they explicitly define their core assumptions. In the end, whether they are of any value will be determined by how much explanatory power they bring into the constantly accruing experimental evidence.

The science of consciousness needs integrative frameworks, and integrative frameworks by definition are multidisciplinary. In the future, and going beyond empirical methods and mathematics, a dialogue with artists, meditators and proponents of the humanities may help us to think outside the box and rediscover some aspects of conscious experience that have largely been unattended to.

Supplementary data

Supplementary data is available at *NCONSC Journal* online.

Acknowledgement

The authors thank Athena Demertzi, Johannes Fahrenfort, Chris Fields, Don Hoffman, Bechir Jarraya, Lucia Melloni, Anil Seth, Jacobo Sitt, Wanja Wiese and two anonymous referees for helpful comments and constructive feedback.

Data availability

There is no experimental data involved in this research.

Funding

This work received financial support by Comisión Nacional de Investigación Ciencia y Tecnología (CONICYT, currently ANID) through Programa Formacion de Capital Avanzado (PFCHA), Doctoral scholarship Becas Chile: CONICYT PFCHA/DOCTORADO BECAS CHILE/2016 - 72170507, Polish Ministry of Science and Higher Education's Diamond Grant DI2016 020046, and the project Comparative investigation of the cortical circuits in mouse, NHP and human (CORTICITY).

Conflict of interest statement

None declared.

References

AMCS, (2021). *Association for Mathematical Consciousness Science (AMCS)*. https://amcs-community.org.

Anderson ML, Chemero T. The problem with brain GUTs: Conflation of different senses of "prediction" threatens metaphysical disaster. *Behav Brain Sci* 2013;**36**:204–5.

Aru J, Suzuki M, Larkum ME. Cellular mechanisms of conscious processing. *Trends Cognit Sci* 2020;**24**:814–25.

Atasoy S, Deco G, Kringelbach ML, Pearson J. Harmonic brain modes: A unifying framework for linking space and time in brain dynamics. *Neuroscientist* 2017;**24**:277–293.

Atasoy S, Deco G, Kringelbach ML. *The Functional Role of Critical Dynamics in Neural Systems* (ed. Tomen N), Playing at the edge of criticality: expanded whole-brain repertoire of connectome-harmonics, Vol. 11, Springer International Publishing, 2019, 27–45.

Atmanspacher H. The Pauli-Jung conjecture and its relatives: a formally augmented outline. *Open Philos* 2020;**3**:527–49.

Atmaspacher H. 20th century variants of dual-aspect thinking. *Mind Matter* 2014;**12**:245–69.

Baars Bernad. *A Cognitive Theory of consciousness*. New York: Cambridge University Press, 1988.

Baars BJ. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Prog Brain Res* 2005;**150**:45–53.

Bachmann T, Hudetz AG. It is time to combine the two main traditions in the research on the neural correlates of consciousness: C=LxD. *Front Psychol* 2014;**5**:1–13.

Barnes E. Explanatory unification and the problem of asymmetry. *Philos Sci* 1992;**59**:558–71.

Batterman RW, Rice CC. Minimal model explanations. *Philos Sci* 2014;**81**:349–76.

Bayne T, Hohwy J, Owen AM. Are there levels of consciousness? *Trends Cognit Sci* 2016;**20**:405–13.

Bayne T. On the axiomatic foundations of the integrated information theory of consciousness. *Neurosci* 2018;**2018**:1–8.

Bayne T, Carter O. Dimensions of consciousness and the psychedelic state. *Neurosci* 2018;**4**:1–8.

Beck F, Eccles JC. Quantum aspects of brain activity and the role of consciousness. *Proc Natl Acad Sci* 1992;**89**:11357–61.

Block N. On a confusion about a function of consciousness. *Behav Brain Sci* 1995;**18**:227–47.

Block N. *The Encyclopedia of Philosophy Supplement* What is functionalism? 1996.

Block N. Sexism, racism, ageism, and the nature of consciousness. *Philos Top* 1999;**26**:39–70.

Block N. Two neural correlates of consciousness. *Trends Cognit Sci* 2005;**9**:46–52.

Block N. Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behav Brain Sci* 2007;**30**:481–548.

Block N. *The Cognitive Neurosciences*, Comparing the major theories of consciousness, Cambridge, MA, US: Massachusetts Institute of Technology, 2009, 4th edn.

Boly M, Massimini M, Tsuchiya N, Postle BR, Koch C, Tononi G. Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and neuroimaging evidence. *J Neurosci* 2017;**37**:9603–13.

Brown R, Lau H, LeDoux JE. Understanding the higher-order approach to consciousness. *Trends Cognit Sci* 2019;**23**:754–68.

Casarotto S, Comanducci A, Rosanova M, Sarasso S, Fecchio M, Napolitani M, Pigorini A, Casali AG, Trimarchi PD, Boly M,

Gosseries O, Bodart O, Curto F, Landi C, Mariotti M, Devalle G, Laureys S, Tononi G, Massimini M. Stratification of unresponsive patients by an independently validated index of brain complexity. *Ann Neurol* 2016;**80**:718–29.

Chalmers DJ. Facing up to the problem of consciousness. *J Conscious Stud* 1995a;**2**:200–19.

Chalmers DJ. The puzzle of conscious experience. *Scientific American* 1995b;**273**:80–6.

Chalmers DJ. *The Conscious Mind*. Oxford University Press, 1997.

Chalmers DJ. Panpsychism and panprotopsychism. *The Amherst Lecture in Philosophy*, 2013;**8**:1–35.

Chalmers DJ. How can we construct a science of consciousness? *Ann N Y Acad Sci* 2013b;**1303**:25–35.

Chalmers DJ. The meta-problem is the problem of consciousness. *J Conscious Stud* 2018;**25**:6–61.

Chalmers DJ. *The Routledge Handbook of Panpsychism* (ed. Seager William), Idealism and the mind-body problem, New York: Routledge, 2019, 353–73.

Changeux J-P. The ferrier lecture 1998: The molecular biology of consciousness investigated with genetically modified mice. *Philos T R Soc B* 2006;**361**:2239–59.

Christiaan Klink P, Self MW, Lamme VAF, Roelfsema PR. *The Constitution of Phenomenal Consciousness. Toward a Science and Theory*. (ed. Miller SM), Theories and methods in the scientific study of consciousness, John Benjamins Publishing Company, 2015, 17–47.

Clark A. *Surfing Uncertainity. Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press, 2016.

Colin M. Consciousness and content, *Proceedings of the British Academy*, Vol. 74 (1988), 219–39.

Colombo M. Why build a virtual brain? Large-scale neural simulations as jump start for cognitive computing. *J Exp Theor Artif In* 2017;**3079**:1–10.

Craver CF. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press: Oxford University Press, 2007.

Crick F, Koch C. Consciousness and neuroscience. *Cereb Cortex* 1998;**8**:97–1007.

Crick F, Koch C. A framework for consciousness. *Nat Neurosci* 2003;**6**:119–26.

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

Dehaene S, Changeux JP. Ongoing spontaneous activity controls access to consciousness: A neuronal model for inattentional blindness. *PLoS Biol* 2005;**3**:0910–27.

Dehaene S, Changeux J-P. Experimental and theoretical approaches to conscious processing. *Neuron* 2011;**70**:200–27.

Dennett DC. *Consciousness in Modern Science* (eds. Marcel A and Bisiach E), Quining Qualia, Oxford University Press, 1988.

Dennett DC. *Consciousness Explained*. London: Penguin Books Ltd, 1991.

Dennett DC. Facing up to the hard question of consciousness. *Philos T R Soc B* 2018;**373**:20170342.

de Regt HW, Baumberger C. What is scientific understanding and how can it be achieved? In: McKain K, Kampourakis K (eds.), *What Is Scientific Knowledge?* New York: Routledge, 2019:66–81.

Doerig A, Schurger A, Hess K, Herzog MH. The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Conscious Cogn* 2019;**72**:49–59.

Durham I, Kleiner J, Kremnitzer Y, Mason J, Prentner R (eds). *Models of Consciousness*, Vol. 22, 2020, Special edition.

Edelman GM. Naturalizing consciousness: A theoretical framework. *Proc Natl Acad Sci* 2003;**100**:5520–4.

Ehresmann AC, Gomez-Ramirez J. Conciliating neuroscience and phenomenology via category theory. *Prog Biophys Mol Biol* 2015;**119**:347–59.

Fahrenfort JJ, Van Leeuwen J, Olivers CNL, Hogendoorn H. Perceptual integration without conscious access. *Proc Natl Acad Sci USA* 2017;**114**:3744–9.

Fahrenfort JJ, Simon VG. Criteria for empirical theories of consciousness should focus on the explanatory power of mechanisms, not on functional equivalence. *Cognit Neurosci* 2021;**12**:93–4.

Fallon F. Dennett on consciousness: realism without the hysterics. *Topoi* 2020;**39**:35–44.

Fazekas P, Overgaard M. Perceptual consciousness and cognitive access: an introduction. *Philos T R Soc B* 2018;**373**:20170340.

Fields C, Hoffman DD, Prakash C, Singh M. Conscious agent networks: Formal analysis and application to cognition. *Cognit Syst Res* 2018;**47**:186–213.

Frankish K. *Illusionism*. Exeter: Imprint Academic, 2017.

Friston KJ, Wiese W, Allan Hobson J. Sentience and the origins of consciousness: From cartesian duality to Markovian monism. *Entropy* 2020;**22**:1–31.

Gallagher S, Zahavi D. *The Phenomenological Mind*. London: Routledge, 2008, first edn.

Glymour C. Explanations, tests, unity, and necessity. *Nous* 1980;**14**:31–50.

Goff P. *Galileo's Error: Foundations for a New Science of Consciousness*. New York: Phanteon Books, 2019.

Graziano MSA, Kastner S. Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cognit Neurosci* 2011;**2**:98–113.

Graziano MSA, Guterstam A, Bio BJ, Wilterson AI. Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. *Cogn Neuropsychol* 2020;**37**:155–72.

Hameroff SR, Penrose R. Consciousness in the universe: A review of the "Orch OR" theory. *Phys Life Rev* 2014;**11**:39–78.

Haun A, Tononi G. Why does space feel theway it does? Towards a principled account of spatial experience. *Entropy* 2019;**21**:1160.

Herzog MH, Drissi-Daoudi L, Doerig A. All in good time: long-lasting postdictive effects reveal discrete perception. *Trends Cognit Sci* 2020;**24**:826–37.

Hoffman DD. Conscious realism and the mind-body problem. *Mind Matter* 2008;**6**:87–121.

Hoffman DD, Prakash C. Objects of consciousness. *Front Psychol* 2014;**5**:1–22.

Hohwy J, Frith C. Can neuroscience explain consciousness? *J Conscious Stud* 2004;**11**:180–98.

Horgan TE, Tienson JL, Graham G. *The Externalist Challenge* (ed. Schantz Richard), Phenomenal intentionality and the brain in a vat, De Gruyter, 2004.

Illari PMK, Williamson J. What is a mechanism? Thinking about mechanisms across the sciences. *Eur J Philos Sci* 2012;**2**:119–35.

Jordy T , *Functional Neuro-Imaging Study of Deep Brain Stimulation Mechanisms for the Restoration of Consciousness using a Non-Human Primate Model*. PhD thesis, Université Paris-Saclay, 2020.

Kastrup B. An ontological solution to the mind-body problem. *Philosophies* 2017;**2**:10.

Kaufer S, Chemero A. *Phenomenology: An Introduction*. New York: Polity, 2015.

Kitcher P. Explanatory unification. *Philos Sci* 1981;**48**:507–31.

Kleiner J. Mathematical models of consciousness. *Entropy* 2020;**22**:609.

Kleiner J, Tull S, *The Mathematical Structure of Integrated Information Theory*, (2020), 1–22.

Kringelbach ML, Cruzat J, Cabral J, Knudsen GM, Carhart-Harris R, Whybrow PC, Logothetis NK, Deco G. Dynamic coupling of whole-brain neuronal and neurotransmitter systems. *Proc Natl Acad Sci USA* 2020;**117**:9566–76.

Lamme VAF, *Why Visual Attention and awareness are different*, (2003).

Lamme VAF. How neuroscience will change our view on consciousness. *Cognit Neurosci* 2010;**1**:204–20.

Lau HC, Passingham RE. Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proc Natl Acad Sci USA* 2006;**103**:18763–8.

Llinas R, Ribary U, Contreras D, Pedroarena C. The neuronal basis for consciousness. *Philos Trans R Soc Lond B Biol Sci* 1998;**353**:1841–9.

Llinás R. Consciousness and the thalamocortical loop. *Int Congr Ser* 2003;**1250**:409–16.

Luppi AI, Vohryzek J, Mediano PAM, Adapa R, Pappas I, Finoia P, Allanson J, Atasoy S, Stamatakis EA. Connectome harmonic decomposition of human brain dynamics reveals a landscape of consciousness. *bioRxiv* 2020. https://www.biorxiv.org/content/10.1101/2020.08.10.244459v1.

Lutz A, Lachaux J-P, Martinerie J, Varela FJ. Guiding the study of brain dynamics by using first-person data: Synchrony patterns correlate with ongoing conscious states during a simple visual task. *Proc Natl Acad Sci USA* 2002;**99**:1586–91.

Lutz A, Thompson E. Neurophenomenology:integrating subjective experience and brain dynamics in the neuroscience of consciousness. *J Conscious Stud* 2003;**10**:31–52.

Lutz A, Greischar LL, Rawlings NB, Ricard M, Davidson R. Long-term meditators self-induce high-amplitude gamma synchrony during mental practice. *Proc Natl Acad Sci* 2004;**101**:16369–16373.

Lutz A, Jha AP, Dunne JD, Saron CD. Investigating the phenomenological matrix of mindfulness-related practices from a neurocognitive perspective. *Am Psychol* 2015;**70**:632–58.

Marshall P, Physicalism B, Kelly EF, Crabtree A, Lanham PM. Transforming the world into experience. *J Conscious Stud* 2001;**8**:59–76.

Masafumi O, Naotsugu T, Shun Ichi A. Unified framework for information integration based on information geometry. *Proc Natl Acad Sci USA* 2016;113:14817–14822.

Mashour GA, Roelfsema P, Changeux JP, Dehaene S. Conscious processing and the global neuronal workspace hypothesis. *Neuron* 2020;**105**:776–98.

Melloni L, Mudrik L, Pitts M, Koch C. Making the hard problem of consciousness easier. *Science* 2021;**372**:911–12.

Metzinger T. Minimal phenomenal experience. *PhiMiSci* 2020;**1**:7.

Michel M, Beck D, Block N, Blumenfeld H, Brown R, Carmel D, Carrasco M, Chirimuuta M, Chun M, Cleeremans A, Dehaene S, Fleming SM, Frith C, Haggard P, He BJ, Heyes C, Goodale MA, Irvine L, Kawato M, Kentridge R, King JR, Knight RT, Kouider S, Lamme V, Lamy D, Lau H, Laureys S, LeDoux J, Lin YT, Liu K, Macknik SL, Martinez-Conde S, Mashour GA, Melloni L, Miracchi L, Mylopoulos M, Naccache L, Owen AM, Passingham RE, Pessoa L, Peters MAK, Rahnev D, Ro T, Rosenthal D, Sasaki Y, Sergent C, Solovey G, Schiff ND, Seth A, Tallon-Baudry C, Tamietto M, Tong F, van Gaal S, Vlassova A, Watanabe T, Weisberg J, Yan K, Yoshida M. Opportunities and challenges for a maturing science of consciousness. *Nat Hum Behav* 2019;**3**:104–7.

Naccache L. Why and how access consciousness can account for phenomenal consciousness. *Philos T R Soc B* 2018;**373**:20170357.

Nagel E. *The Structure of Science: Problems in the Logic of Scientific Explanation*. New York: Harcourt, 1961.

Nathan MJ. *Foundations of Science*, Causation vs. causal explanation: which is more fundamental? Vol. 0123456789, 2020.

Noel JP, Ishizawa Y, Patel SR, Eskandar EN, Wallace MT. Leveraging nonhuman primate multisensory neurons and circuits in assessing consciousness theory. *J Neurosci* 2019;**39**:7485–500.

Northoff G. What the brain's intrinsic activity can tell us about consciousness? A tri-dimensional view. *Neurosci Biobehav Rev* 2013;**37**:726–38.

Northoff G, Huang Z. How do the brain's time and space mediate consciousness and its different dimensions? Temporo-spatial theory of consciousness (TTC). *Neurosci Biobehav Rev* 2017;**80**:630–45.

Northoff G, Tsuchiya N, Saigo H. Mathematics and the brain: a category theoretical approach to go beyond the neural correlates of consciousness. *Entropy* 2019;**21**:1234.

Northoff G, Lamme V. Neural signs and mechanisms of consciousness: Is there a potential convergence of theories of consciousness in sight? *Neurosci Biobehav Rev* 2020;**118**:568–87.

Oizumi M, Albantakis L and Tononi G, Sporns O. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput Biol* 2014;**10**:e1003588.

Olivares FA, Vargas E, Fuentes C, Martínez-Pernía D, Canales-Johnson A. Neurophenomenology revisited: second-person methods for the study of human consciousness. *Front Psychol* 2015;**6**:1–12.

Pautz A. *Consciousness Inside and Out: Phenomenology, Neuroscience, and the Nature of Experience* (ed. Brown R), The real trouble with phenomenal externalism: new empirical evidence for a brain-based theory of consciousness, Dordrecht: Springer, 2014, 237–317.

Petitmengin C, Remillieux A, Valenzuela-Moguillansky C. Discovering the structures of lived experience. *Phenomenol Cogn Sci* 2019;**18**:691–730.

Poletti S, Abdoun O, Zorn J, Lutz A. Pain regulation during mindfulness meditation: phenomenological fingerprints in novices and experts practitioners. *Eur J Pain* 2021;**25**:1583–602.

Prakash C, Chris Fields DD Hoffman RP, Singh M. Fact, fiction, and fitness. *Entropy* 2020;**22**:1–23.

Prentner R. Consciousness and topologically structured phenomenal spaces. *Conscious Cogn* 2019;**70**:25–38.

Reardon S. Rival theories face off over brain's source of consciousness. *Science* 2019;**366**:293.

Reutlinger A. Explanation beyond causation? New directions in the philosophy of scientific explanation. *Philos Compass* 2017;**12**:1–11.

Roelfsema PR, Lamme VAF, Spekreijse H, Bosch H. *Figure - ground segregation in a recurrent network architecture*. *J Cognit Neurosci* 2002;**14**:525–37.

Rodríguez E. *Ciencias de la mente: aproximaciones desde latinoamerica* (eds. Kronmüller E and Cornejo C), Ideas para naturalizar el estudio de la conciencia, Juan Carlos Sáez Editor, Editorial Granica, 2008, 301–24.

Rosenthal DM. How many kinds of consciousness? *Conscious Cogn* 2002a;**11**:653–65.

Rosenthal DM. *Philosophy of Mind: Classical and Contemporary Readings* Explaining consciousness, 2002b, 406–21.

Rosenthal DM. Consciousness and its function. *Neuropsychologia* 2008;**46**:829–40.

Rosenthal D. Assessing criteria for theories. *Cognit Neurosci* 2020;**00**:1–2.

Salmon WC. Scientific explanation: causation and unification. *CríTica (MéXico D. F. En línea)* 1990;**22**:3–23.

Seth A. Models of consciousness. *Scholarpedia* 2007;**2**:1328.

Searle JR. How to study consciousness scientifically. *Philos Trans R Soc Lond B Biol Sci* 1998;**353**:1935–42.

Searle JR. Consciousness. *Annu Rev Neurosci* 2000;**23**:557–78.

Seth A. Explanatory correlates of consciousness: Theoretical and computational challenges. *Cognit Comput* 2009;**1**: 50–63.

Shoemaker S. Phenomenal character revisited. *Philos Phenomenol Res* 2000;**60**:465.

Siclari F, Baird B, Perogamvros L, Bernardi G, LaRocque JJ, Riedner B, Boly M, Postle BR, Tononi G. The neural correlates of dreaming. *Nat Neurosci* 2017;**20**:872–8.

Signorelli CM, *Theoretical Models and Measures of Conscious Brain Network Dynamics. An Integrative Approach*, PhD thesis, 2021.

Signorelli CM, Boils JD, *Multilayer Networks as Embodied Consciousness Interactions. A Formal Model Approach*, (2021), To be submitted.

Signorelli CM, Meling D. Towards new concepts for a biological neuroscience of consciousness. *Cogn Neurodynamics* 2021. 10.1007/s11571-020-09658-7.

Signorelli CM, Wang Q, Coecke B. Reasoning about conscious experience with axiomatic and graphical mathematics. *Conscious Cogn* 2021b. 10.1016/j.concog.2021.103168.

Signorelli CM, Wang Q, Khan I. A compositional model of consciousness based on consciousness-only. *Entropy* 2021c;**23**:308.

Steven Scholte H, Jolij J, Fahrenfort JJ, Lamme VAF. Feedforward and recurrent processing in scene segmentation: Electroencephalography and functional magnetic resonance imaging. *J Cognit Neurosci* 2008;**20**:2097–109.

Storm JF, Boly M, Casali AG, Massimini M, Olcese U, Pennartz CMA, Wilke M. Consciousness regained: disentangling mechanisms, brain systems, and behavioral responses. *J Neurosci* 2017;**37**:10882–93.

Strawson G. Realistic monism: why physicalism entails panpsychism. *J Conscious Stud* 2006;**13**:3–31.

Strevens M. The causal and unification approaches to explanation unified - causally. *Nous* 2004;**38**:154–76.

Stubenberg L. *The Stanford Encyclopedia of Philosophy* Neutral Monism, Metaphysics Research Lab, Stanford University, 2018.

Suzuki M, Larkum ME. General anesthesia decouples cortical pyramidal neurons. *Cell* 2020;**180**:666–676.e13.

Tagliazucchi E. The signatures of conscious access and its phenomenology are consistent with large-scale brain communication at criticality. *Conscious Cogn* 2017;**55**:136–47.

Thompson E, Varela FJ. Radical embodiment: neural dynamics and consciousness. *Trends Cognit Sci* 2001;**5**:418–25.

Thompson E. Life and mind: from autopoiesis to neurophenomenology. A tribute to francisco varela. *Phenomenol Cogn Sci* 2004;**3**:381–98.

Thompson E. *Mind in Life*. Cambridge, Massachusetts: Harvard University Press, 2007.

Tononi G, Boly M, Massimini M, Koch C. Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neurosci* 2016;**17**:450–61.

Tsuchiya N, Taguchi S, Saigo H. Using category theory to assess the relationship between consciousness and integrated information theory. *Neurosci Res* 2016;**107**:1–7.

Tsuchiya N, Saigo H, *Applying Yoneda's Lemma to Consciousness Research: Categories of Level and Contents of Consciousness*, (2020), Preprint.

Van Gulick R. *Consciousness*, *The Stanford Encyclopedia of Philosophy* (2018).

Varela FJ. Neurophenomenology: a methodological remedy for the hard problem. *J Conscious Stud* 1996;**3**:330–49.

Varela F, Lachaux J-P, Rodriguez E, Martinerie J. The brainweb: phase synchronization and large-scale integration. *Nat Rev Neurosci* 2001;**2**:229–39.

Webb TW, Graziano MSA. The attention schema theory: A mechanistic account of subjective awareness. *Front Psychol* 2015;**6**: 1–11.

Wiese W. The science of consciousness does not need another theory, it needs a minimal unifying model. *Neurosci* 2020;**2020**:1–7.

Wiese W, Friston KJ. The neural correlates of consciousness under the free energy principle: From computational correlates to computational explanation. *PhiMiSci* 2021;**22**:1–31: forthcomin.

Windt JM, Nielsen T, Thompson E. Does consciousness disappear in dreamless sleep? *Trends Cognit Sci* 2016;**20**:871–82.

Woodward J. *Making Things Happen*. Oxford: Oxford University Press, 2004.

Woodward J. II–James woodward: mechanistic explanation: its scope and limits. *Aristot Soc Suppl Vol* 2013;**87**:39–65.

Woodward J. *The Stanford Encyclopedia of Philosophy* Scientific explanation, Metaphysics Research Lab, Stanford University, 2019.

Yaron I, Melloni L, Pitts M, Mudrik L. The consciousness theories studies (ConTraSt) database: analyzing and comparing empirical studies of consciousness theories. *bioRxiv* 2021. https://www.biorxiv.org/content/10.1101/2021.06.10.447863v1.

Yoshimi J. Mathematizing phenomenology. *Phenomenol Cogn Sci* 2007;**6**:271–91.