Neuroscience and Biobehavioral Reviews 156 (2024) 105480

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

A construct-first approach to consciousness science[☆]

Peter Fazekas[a,b,*], Axel Cleeremans[c], Morten Overgaard[b]

[a] *Aarhus Institute of Advanced Studies, Aarhus University, Høegh-Guldbergs Gade 6B, 8000 Aarhus, Denmark*
[b] *Center of Functionally Integrative Neuroscience, Aarhus University, Universitetsbyen 3, 8000 Aarhus, Denmark*
[c] *Center for Research in Cognition & Neurosciences, Université Libre De Bruxelles, 50 avenue F.D. Roosevelt CP191, 1050 Bruxelles, Belgium*

ARTICLE INFO

Keywords:
Conscious experience
Theories of consciousness
Theoretical constructs
Auxiliary hypotheses

ABSTRACT

We propose a new approach to consciousness science that instead of comparing complex theoretical positions deconstructs existing theories, takes their central assumptions while disregarding their auxiliary hypotheses, and focuses its investigations on the main constructs that these central assumptions rely on (like global workspace, recurrent processing, metarepresentation). Studying how these main constructs are anchored in lower-level constructs characterizing underlying neural processing will not just offer an alternative to theory comparisons but will also take us one step closer to empirical resolutions. Moreover, exploring the compatibility and possible combinations of the lower-level constructs will allow for new theoretical syntheses. This construct-first approach will improve our ability to understand the commitments of existing theories and pave the way for moving beyond them.

1. Theory-based versus construct-first approaches to consciousness research

Everybody who thinks about **consciousness** (see Box 1) has a theory about it, and that has now become a problem for the field as a whole: the discipline struggles with an abundance of alternative theories, all striving to offer a definitive answer about what consciousness is, about what it does, and about its neural basis (Seth and Bayne, 2022). The maturation of the field has resulted neither in the convergence nor in the elimination of theories. Experiments have rarely been designed to test theoretical predictions or to compare competing theories (Seth and Bayne, 2022; Yaron et al., 2022; Melloni et al., 2021, 2023). Instead, empirical findings have typically been interpreted post hoc, from the perspective of a given theoretical framework, leading to claims of confirmation and leaving us with a large number of empirically supported yet incompatible theories that often talk past each other (Yaron et al., 2022). According to an emerging sentiment, the field is now in need of a major change in its approach (Seth and Bayne, 2022; Yaron et al., 2022; Melloni et al., 2021). Two different suggestions have recently been proposed about how to best move the field forward. The first is adversarial collaboration, through which experiments are designed by competing theorists to falsify specific predictions and hence eliminate some theories (Melloni et al., 2021, 2023; Doerig et al., 2021; Cogitate et al., 2023a). The second is to attempt to relate different theories via comparing them on the basis of their explanatory targets (Seth and Bayne, 2022; Northoff and Lamme, 2020).

Our goal here is to offer an alternative strategy. We propose that prior to attempting to derive conflicting testable predictions from the available theories or to analyze what kinds of phenomena these theories have been designed to address, one should be clear about the exact meaning of the **theoretical constructs** (see Box 2) that the main ideas of the very theories in question are couched in. To put it in another way: before turning towards what the different theories say with respect to specific empirical conditions or different target phenomena, we should first try to get to the core of their **central assumptions**.

2. Theories of consciousness and relations in construct space

2.1. Deconstructing theories of consciousness

Understanding theoretical constructs consists in clarifying (1) how they are related to each other and (2) how they could be operationalized and tested empirically. In what follows, we shall argue that both of these aspects of a construct-focused rather than theory-focused approach offers advantages that can advance the field. In this section, our focus will be on the relationship between different constructs and how a construct-

[☆] *Perspective article for Neuroscience & Biobehavioral Reviews*

[*] Correspondence to: Høegh-Guldbergs Gade 6B, 8000 Aarhus C, Denmark.
E-mail address: fazekas.peter@gmail.com (P. Fazekas).

https://doi.org/10.1016/j.neubiorev.2023.105480
Received 7 September 2023; Received in revised form 26 October 2023; Accepted 20 November 2023
Available online 24 November 2023

first approach leads to thinking about the core claims of theories as occupying locations in a **construct space**, the dimensions of which are defined by the constructs in question.

To do so, we first need to go through a 'deconstruction' phase in which our goal will be to clarify how the *central constructs* of different theories of consciousness are related to further theoretical constructs. Plainly put, central constructs are the concepts in which the core claims of theories are couched. These central constructs and their definitions are typically 'higher-level' in the sense that they are more abstract and utilize cognitive or psychological descriptions that are often specific to the given theory (e.g. global workspace). However, as theories of consciousness are increasingly empirically focused, many theorists further explicate their central constructs in terms of lower-level constructs and descriptions utilizing neural terminology that are often deployed as bridges to operationalization and offer a shared vocabulary. To illustrate the deconstruction phase of our construct-first approach, here we offer a brief overview of the clarifications of the central constructs of three prominent and widely discussed theories of consciousness: the global workspace, local recurrence and higher order theories (a fourth prominent account, the **information integration theory** (IIT) is discussed in Box 2; for a more thorough list of available theories, see also Box 2).

2.2. Global workspace theory

According to the main claim of the **global workspace theory**, a piece of information becomes conscious when it gets into the *global workspace (*Baars et al., 2021; Mashour et al., 2020; Baars, 1988, 1998; Dehaene et al., 1998). This idea is often further explained by invoking other terms like *global availability* and *cognitive access*: the information that is in the global workspace is widely available to many local processors (Mashour et al., 2020) including cognitive systems like planning, reasoning and rational control of action (Block, 2005) that can access, monitor and manipulate this information, and this *global broadcasting* promotes conscious phenomenology (Block, 2023). This high-level characterization is anchored at the level of neuroscience by the Global Neural Workspace (GNW) hypothesis in the shape of a widely distributed network of neurons forming interconnected cortical hubs that can either mobilize or suppress local processors and communicate in a reciprocal manner with them via bottom-up feedforward and top-down feedback connections (Dehaene et al., 1998; Deco et al., 2021; Dehaene and Naccache, 2001; Dehaene et al., 2006, 2014, 2017). Due to these *recurrent loops*, the activation of the GNW network—which is a non-linear process often referred to as *ignition (*Dehaene et al., 2003; Dehaene and Changeux, 2011)—"amplifies and sustains neural representations allowing the corresponding information to be globally accessed by local processors" (Mashour et al., 2020, p.776). In this framework, then, contents become part of the global workspace through igniting the GNW network, and global availability depends on the *strong* and *stable signal* that the corresponding neural representation affords.

The constructs the global workspace approach relies on are not independent from each other and are formulated at different levels of abstraction. The higher-level constructs 'global workspace' and 'global availability' are defined at a lower level in terms of the constructs 'ignition', 'recurrent activity', 'signal strength' and 'temporal stability'. *Ignition* promotes the spread of *recurrent activity*, which in turn increases *signal strength* and *temporal stability* rendering the content of these representations *globally available*. From a dynamic systems perspective, ignition can be seen as a shift in the state of the system towards the activation of deeper, longer range feedback loops which then further shift the system towards higher levels of signal strength and temporal stability.

2.3. Local recurrence theory

The **local recurrence theory** (LRT) (Lamme and Roelfsema, 2000; Lamme, 2003, 2006) denies the importance of the global workspace and instead associates the occurrence of conscious experiences with the construct of a *fragile visual short-term memory* (VSTM) store (Sligte et al., 2008; Vandenbroucke et al., 2015; Pinto et al., 2013). Fragile VSTM is activated earlier in time and has a higher capacity than the global

Box 1
Glossary of main terms.

Auxiliary hypotheses: all theories rely on some theoretical constructs to formulate their central assumptions. Specific empirical predictions, however, are typically not derived solely from these constructs and assumptions. The further premises that are used to link the central assumptions of theories to observational data are called auxiliary hypotheses.

Central assumption: the core claim of a theory that defines the theory's target phenomenon in terms of a theoretical construct that the theory relies on. In the case of theories of consciousness, this is the fundamental statement a theory offers to illuminate what consciousness is.

Consciousness: subjective experience. There are major distinctions to be made between state versus creature consciousness (in other terminology: local and global states of consciousness), phenomenal versus access consciousness, and self-awareness and perceptual awareness. Here our focus is on the phenomenology that characterizes the contents of perceptual states, i.e. what it is like to be in those local states.

Construct space: an abstract space the dimensions of which are the theoretical constructs that the different theories of consciousness rely on. Its dimensionality can be reduced by uncovering relationships between the constructs. Theories of consciousness occupy partly overlapping regions within this space. Construct-drivel empirical studies can explore which locations of this space are compatible with the presence of consciousness.

Global workspace theory: claims that consciousness requires a specific kind of cognitive architecture, the global workspace, that distributes information to a broad range of consumer systems. Those pieces of information will become part of the content of consciousness that enter into this global workspace.

Higher-order theories of consciousness: a group of theories claiming that consciousness requires metarepresentation. The content of those first-order representations are conscious that are re-represented by higher-order representations.

Information integration theory: associates consciousness with the information integration capacity of a system. The content of consciousness is determined by the components of that sub-system that has the maximum of integrated information.

Local recurrence theory: claims that local recurrent activity is both necessary and sufficient for the occurrence of consciousness. According to this view, the content of those neural representations that are amplified, stabilized and integrated by local feedback loops are already conscious.

Theoretical constructs: scientific theories are systems of theoretical constructs (specific concepts) and propositions connecting these constructs that together with further assumptions (auxiliary hypotheses) and boundary conditions are able to account for some target phenomena.

workspace, but is fragile in the sense that similar stimuli can overwrite it. According to LRT, information in this fragile VSTM store is already phenomenally conscious. The global workspace, from this perspective, only contributes to making contents available to cognitive processing (Sligte et al., 2008; Vandenbroucke et al., 2015; Lamme, 2010). At a lower level of description, fragile VSTM is accounted for in terms of the construct of *recurrent processing*. Contrary to the GNW hypothesis, LRT claims that *local* recurrent interactions in the form of top-down feedback from higher levels of the perceptual hierarchy and lateral connections within and between processing areas are already sufficient for the occurrence of conscious experiences (Northoff and Lamme, 2020; Lamme, 2018). Crucially, global recurrent loops are not necessary. The immediate bottom-up processing occurring after stimulus onset (the feedforward sweep) already extracts information about the stimulus and can even reach motor regions and control centers, but the corresponding representations are short-lived and their content remains unconscious. With the activation of local horizontal connections and feedback loops, however, these representations get *amplified* and *stabilized*, enabling information exchange between distinct areas that process different properties of the stimulus (Lamme, 2006) and thus support perceptual binding, i.e. the *integration* of distinct stimulus-features into a single unified percept. According to LRT, this kind of perceptual organization is the key feature of conscious experiences, and thus local recurrence is sufficient for the occurrence of phenomenal consciousness (Northoff and Lamme, 2020). To put it differently, the representations produced by the feedforward sweep are too transient and lack integration, hence this initial stage of processing is insufficient for consciousness, whereas the further stability, signal amplification and integration with plans and task-relevant information that ignition provides is unnecessary. The local recurrence theory thus associates consciousness with intermediate levels of temporal stability, signal strength and spread of recurrent processing (Vandenbroucke et al., 2015).

That is, the abstract, higher-level constructs of global workspace and fragile visual short-term memory can be further clarified in terms of the same set of lower-level constructs of temporal stability, signal strength and spread of recurrent processing. Fig. 1/(a-b) illustrates their relation in the construct space defined by these dimensions.

2.4. Higher-order theories

The central construct of **higher-order theories** (HOT) of consciousness (Rosenthal, 2005; Lycan, 1996; Kriegel, 2009; Carruthers, 2000; Lau and Rosenthal, 2011) is *metarepresentation*: the idea that the content of a representation of a stimulus feature (a first-order representation) becomes conscious only if there is another (higher-order) representation present in the brain that indicates the existence of the target first-order representation. There are many varieties of HOT differing in exactly what constructs they rely on to explicate this notion of metarepresentation by a higher-order state (Brown et al., 2019; Gennaro, 2012; Brown, 2015; LeDoux and Brown, 2017). For example, the Self-organizing Metarepresentational Account (SOMA) (Cleeremans et al., 2020) claims that the relevant kind of higher-order representations are *explicit re-descriptions* of the implicit knowledge encoded in how perceptual representations affect other representations deeper in the first-order network that maps perception to action. The function of these representational re-descriptions is to contribute to cognitive control via making this implicit knowledge available as data to further processing. First-order representations that have sufficient ‘quality’, i.e. *strength*, *stability* and *distinctiveness*, activate these higher-order re-descriptions, which in turn render the content of the first-order representations conscious. If the quality of the first-order representations reaches a certain upper limit (as in the case of the automatization of certain skills), then the higher-order redescribing processes disengage and the content of the first-order representations fades out of consciousness (Cleeremans et al., 2020; Cleeremans, 2011).

Note that the lower-level constructs that SOMA relies on to explicate the higher-level construct of metarepresentation partly overlap with the constructs used by GNW and LRT: all three accounts make reference to the strength and stability of the first-order representations when characterizing which content elements can occur in consciousness. SOMA also introduces a new dimension to the overall construct-space: representational re-description, which is implemented as an additional network that operates on the internal representations generated by the first-order network. See Fig. 1/(c-d) for more details, and Box 3 for a comparison of different varieties of HOT from this perspective.

Box 2
Theoretical constructs in the science of consciousness and the information integration theory.

All existing theories of consciousness rely on certain theoretical constructs to characterize those states of affairs that they take to be instrumental in producing subjective experiences. The central constructs of the global workspace, local recurrence and higher-order theories are *global workspace* (Mashour et al., 2020; Baars, 1988; Dehaene and Changeux, 2011), *recurrent processing* (Lamme, 2006; Zeki, 2003) and *metarepresentation* (Rosenthal, 2005; Lau and Rosenthal, 2011; Brown et al., 2019), respectively (see main text). Other central constructs proposed include an *attention schema* (Graziano and Webb, 2015; Graziano, 2020; Graziano et al., 2020; Wilterson et al., 2021; Graziano, 2022), the *attentional amplification of intermediate-level representations* (Prinz, 2007, 2012), *predictive inference of the causes of sensory signals* (Hohwy, 2012, 2013; Clark, 2013, 2015, 2019), *control-oriented predictive regulation of physiological states* (Seth and Tsakiris, 2018; Seth, 2021), *mastery of the laws governing sensorimotor contingencies* (O’Regan and Noë, 2001), *re-entrant interactions among populations of neurons in the thalamocortical system* (Edelman, 1987) and *quantum computations within microtubules inside neurons* (Hameroff and Penrose, 1995, 2014, 2016).

The central construct of the information integration theory (IIT) is a *cause–effect structure that specifies the maximum of integrated information (*Tononi, 2004; Tononi et al., 2016; Albantakis et al., 2022; Oizumi et al., 2014). IIT starts from features of experiences that it identifies as fundamental (experiences exist, are structured, specific, irreducibly unified and definite) and treats them as axioms (self-evident truths, although see (Bayne, 2018)). Then, these axioms are ‘translated’ into postulates about the physical substrate of consciousness (has cause-effect power, which is structured, specific, unitary and definite; for the definition of these constructs see Albantakis et al., 2022; Oizumi et al., 2014). From the correspondence between the phenomenal and causal features, and the precisification of the characteristics of psychical states sharing such causal features IIT concludes that the physical substrate of consciousness has a cause-effect structure that is maximally irreducible, the components of the cause-effect structure correspond one-to-one to the components of experience determining its quality, and the quantity of experience is measured by the maximum of intrinsic, integrated cause-effect power.

Although a full deconstruction (see main text) of IIT is beyond the scope of this paper, it is worth noting the relations between IIT and the neural-level constructs discussed in the main text. Due to the requirement of generating *integrated* information, IIT claims that *recurrence*, in the sense of *feedback connections,* is fundamental for the occurrence of consciousness. However, *stable* and *intense* activity—in fact, *activity* of any kind—is not required, as for IIT it is the potential for interactions among parts of a complex that matters (Oizumi et al., 2014).

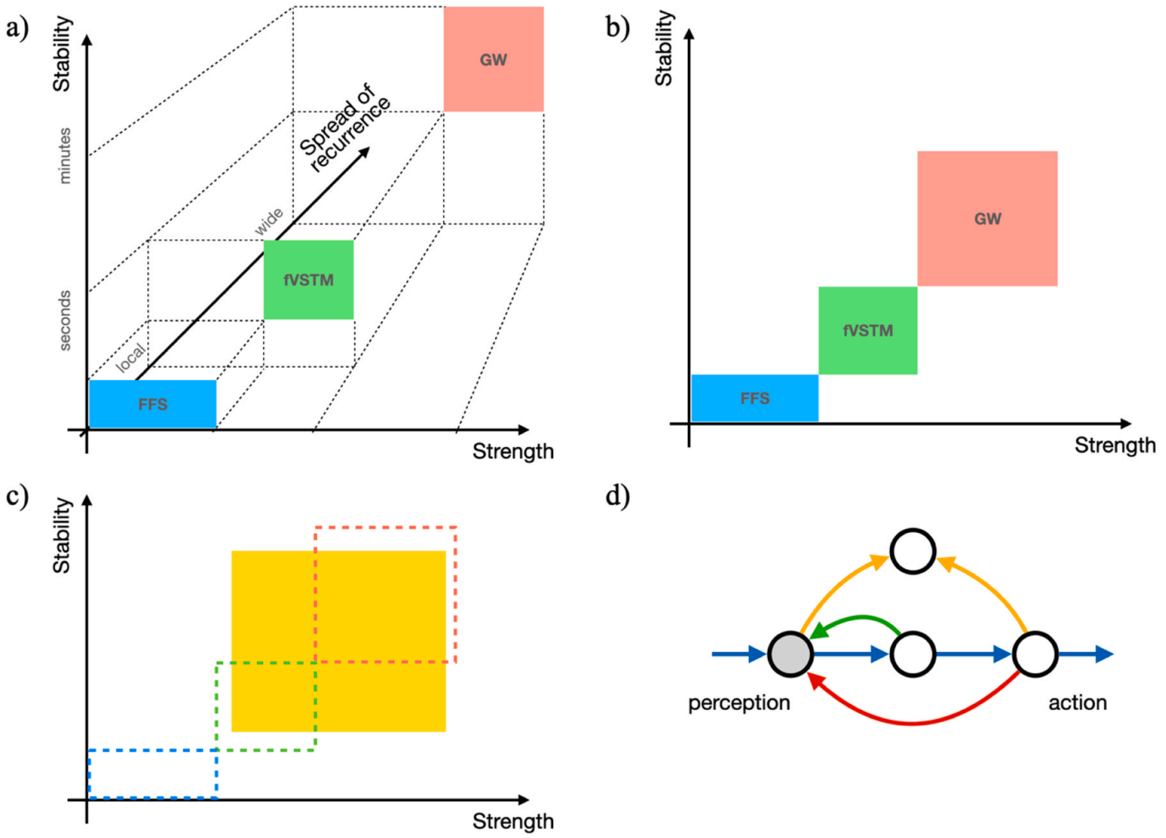

Fig. 1. a-b) The relations of global workspace and local recurrence theories in construct space. a) The 'temporal stability' – 'signal strength' – 'spread of recurrence' subspace of the construct space. GW: global workspace; fVSTM: fragile visual short-term memory; FFS: feedforward sweep. b) Projection onto the 'temporal stability' – 'signal strength' plane. c-d) The SOMA variety of HOT compared to GNW and LRT. c) Same projection as in (b). The yellow area depicts the 'intermediate levels' of signal strength and temporal stability that SOMA associates the activation of higher-order representations with. The relations between these intermediate levels and the levels of strength and stability characteristic of GNW and LRT (red and green outlines, respectively, see (b)) are unclear. d) A schematic illustration of the core claims of GNW, LRT and HOT. Circles: neural representations; blue arrows: bottom-up information processing; red arrow: long-range top-down connection implementing wide-spread recurrence; green arrow: local recurrence; yellow arrows: the re-representation of a first-order representation (grey disk) and its downstream consequences.

2.5. From construct space to new insights

The resulting picture is that of a construct space in which the central assumptions of different theories of consciousness occupy regions that are partly defined by the same lower-level constructs. The construct-first approach recommends that the major high-level constructs that consciousness has been associated with (see Box 2) should be analyzed in a similar fashion to uncover their relations to lower-level constructs, to add new dimensions to the construct space, and to explore via focused empirical investigations their connections to other dimensions and to consciousness itself. Studying which regions of the construct space central assumptions of existing theories of consciousness cluster around and the specificities of their overlaps can lead to novel ideas with regard to what sets of features should theoretical and empirical efforts within consciousness science be re-focused to.

Note that in this framework there are three different levels of descriptions and two separate sets of claims connecting pairs of these levels. The central assumptions of theories of consciousness connect phenomenal descriptions to cognitive descriptions, whereas the further explications anchoring the cognitive constructs to lower-level constructs connect cognitive descriptions to neural descriptions. In the present paper we put the philosophical question regarding the nature of the relationships between these different descriptions (i.e. eliminative, reductive, non-reductive (Fazekas, 2009, 2022) aside. What is important from our point of view, is that the construct space is defined by neural level constructs that are shared among the different theories and thus can offer a common, theory-neutral perspective.

3. The construct-first approach as a research paradigm

3.1. The target phenomena of consciousness science

The study of consciousness is a unique scientific discipline. Its target phenomenon, conscious experience, resists functionalization. This is the source of the 'hard problem of consciousness': unlike in the case of other scientific phenomena, there is a lack of connection between consciousness and functional descriptions (Chalmers, 1995, 1996). Such a functional characterization of a target phenomenon is crucial as standard scientific explanatory practice aims to account for these functional characterizations (by finding a mechanism that fills the functional role). In other words, the functional characterizations in question provide those 'essential properties' of the target phenomenon that scientists then can set out to try to account for.

Although the existence of the hard problem is debated (Dennett, 1991, 2016), and there are claims that it is, after all, possible to associate functions with phenomenal experience (Cleeremans and Tallon-Baudry, 2022), it is safe to say that there is a significant difference in how straightforward it is to connect consciousness versus other scientific target phenomena to functional descriptions. In consciousness science, this is reflected in the fact that there is no consensus regarding what those 'essential properties' of consciousness might be that explanatory attempts should concentrate on. Different theories focus on different

Box 3
Higher-order states, re-representation and consciousness.

Although re-representation is a central construct of higher-order theories of consciousness, different varieties of this approach differ in how they think about the relationship between the content of first- and higher-order states. Some are committed to the explicit re-representation of content, whereas others envisage higher-order states more like monitoring or modelling systems.

According to David Rosenthal's classical version of HOT (Rosenthal, 2005; Lau and Rosenthal, 2011) (see Fig. 2/a), a perceptual representation (P) is conscious to the extent that one is representing oneself (S) as being in that state. The higher-order representation re-represents the first-order content in such a thought-like format. Fig. 3

The perceptual reality monitoring (PRM) account proposes that the higher-order component is implemented by a discriminator system (the PRM) that interprets whether first-order representations are reliable reflections of the external world (see Fig. 2/b). The PRM only indexes the sensory quality space encoded by first-order representations, no re-representation of first-order content occurs (Lau, 2019; Lau et al., 2022). Reality monitoring judgments are driven by the quality of the perceptual representations (P) and by traces of cognitive (C) operations (Fazekas, 2021).

Similarly, the higher-order state space (HOSS) framework (see Fig. 2/c), which combines HOT with the predictive processing approach, proposes that the higher-order state indexes first-order representations with content-invariant 1-dimensional magnitude tags corresponding to the posterior probability of reporting that the content of the representation was present (Fleming, 2020). The state of attention (A) might serve as input (dashed arrow) into resolving ambiguity about the state of awareness (Fleming, 2020).

According to SOMA (Cleeremans et al., 2020; Cleeremans, 2011), higher-order representations are explicit re-descriptions of the knowledge implicit in the cause-effect structure of first-order representations (see Fig. 2/d and main text). Such higher-order representations can occur locally and more globally as well, in fact, SOMA hypothesizes that the global workspace itself might be a result of the interconnected hierarchies of representational re-descriptions.

Finally, although the attention schema theory (Graziano and Webb, 2015; Graziano, 2020; Graziano et al., 2020; Wilterson et al., 2021; Graziano, 2022) is officially not a version of HOT, it relies on a similar higher-order modelling mechanism that re-represents first-order processing features (see Fig. 2/e). In this case, the re-representation is imperfect and connects perceptual representations (P) with the self (S) and with information regarding the allocation of attention (A) to create an attention schema that can be utilized for the control of attention.

aspects of consciousness: GNW on information availability, LRT on perceptual unity, HOT on one's awareness of undergoing a conscious experience, IIT on the irreducibly integrated nature of a conscious experience.

In a milieu like this, where there is no consensus regarding what exactly it is that a theory of consciousness has to account for, the relationships between alternative explanatory attempts are especially unclear. It is not just the central constructs and core claims of the theories that are different, but there is no obvious overlap even between their explanatory targets, i.e. what aspects of consciousness the central constructs are associated with by the core claims.

3.2. *Theory-based experiments and auxiliary hypotheses*

The prospects of comparing theories via their empirical predictions is dubious as well. One reason why the interaction between theoretical frameworks of consciousness and empirical research is prone to criticism (Yaron et al., 2022) is that the central constructs of existing theories of consciousness are conceptually distant from what can actually be empirically tested. The empirical predictions of these theories depend to a great extent on **auxiliary hypotheses** that connect the central assumptions (sometimes also called the 'hard core' (Lakatos, 1970)) of the theories to details regarding implementation and measurement. When a prediction is refuted, it is the conjunction of the central assumptions and the set of associated auxiliary hypotheses that is falsified and needs to be revised. The typical move at this point is to retain the defining core of the theory and to adjust the auxiliary hypotheses creating a new version of the entire theory-complex. That is, auxiliary hypotheses form a 'protective belt' (Lakatos, 1970) around the central assumptions defending them from change. Empirical predictions thus are especially prone to bear almost no consequences whatsoever from the perspective of the possible falsification of the central assumptions themselves (Lakatos, 1970; Quine, 1951; Popper, 1959; Duhem, 1914).

This consideration is true of scientific theories in general. Consciousness science, however, is especially affected, as its target phenomenon lacks straightforward association with functional descriptions, and hence there is no serious 'sanity check' on possible central assumptions. So the fact that classical proposals of empirical tests of particular theories of consciousness (Dehaene and Changeux, 2011; Lau and Rosenthal, 2011), recent debates focusing on the involvement of the prefrontal cortex (Boly et al., 2017; Odegaard et al., 2017; Raccah et al., 2021) and even the current adversarial collaborations that aim to test contradictory predictions of different theories (Melloni et al., 2021, 2023; Cogitate et al., 2023a) all heavily rely on auxiliary hypotheses is worrisome, and casts doubt on the prospects of theory elimination and convergence. In Box 4, we present concrete examples from the recent adversarial collaboration demonstrating how findings incompatible with initial predictions lead to the revision of implementation- and measurement-specific auxiliary hypotheses (Baars et al., 2021; Boly et al., 2017; Odegaard et al., 2017; Raccah et al., 2021; Michel and Morales, 2020) rather than the central assumptions of the theories.

3.3. *Construct-driven experimentation*

The construct-first approach recommends a reorientation of empirical efforts from generating data that feeds into theory-complexes to testing specific construct-driven hypotheses that focus on how the different values of the different dimensions of the construct space are related to the presence or absence of consciousness.

Take the notion of working memory as an example (Baddeley, 2003, 2012; D'Esposito and Postle, 2015; Eriksson et al., 2015; Christophel et al., 2017). References to working memory permeate consciousness science (Vandenbroucke et al., 2015; Baars and Franklin, 2003; Block, 2011; Cohen and Dennett, 2011; Bronfman et al., 2014; Phillips, 2018). It is one of the major constructs that have been used to explicate the meaning of the global workspace (see especially the literature on whether perceptual consciousness overflows cognitive access (Lamme, 2010; Block, 2011; Cohen and Dennett, 2011; Bronfman et al., 2014; Phillips, 2018; Block, 2014). Working memory, however, has multiple functions: it plays a role in information maintenance, monitoring and

Box 4
Auxiliary hypotheses and the adversarial collaboration between GNW and IIT.

To see more specific examples of how the presence of auxiliary hypotheses in contemporary consciousness science calls the prospects of the recent adversarial collaborations into doubt, consider the protocol and results of the Cogitate Consortium that aims to test contrasting predictions of GNW and IIT (Melloni et al., 2023; Cogitate et al., 2023a). The study focuses on five kinds of predictions regarding the location of the NCC, decoding the content of consciousness, the temporal dynamics of NCC, the relations between characteristics of pre-stimulus activity and experience, and functional connectivity.

First, notice that according to the architects of the protocol, "the most viable and testable point of disagreement between the theories" (Melloni et al., 2023, p.4) is the one that concerns to location of the NCC. However, even the architects of the protocol themselves are ready to acknowledge that the core claims of IIT have no relevance in this regard—it is only an "auxiliary prediction of IIT" (Melloni et al., 2023, p.4) that localizes the NCC in the so-called posterior hot zone (Koch et al., 2016a, 2016b).

Equally telling is the reaction of the proponents of IIT to the result that no sustained synchronization between category selective regions and V1/V2 was found, which is "incompatible with IIT's claim that the state of the neural network, including its activity and connectivity, specifies the degree and content of consciousness" (Cogitate et al., 2023a, p.24), i.e. is incompatible with the *core claim* of IIT that the substrate of consciousness is a maximum of integrated information (given that this substrate is supposed to be localized in the posterior hot zone) (Cogitate et al., 2023a, Extended Data Figure 10). Addressing this issue in a supplementary discussion, proponents of IIT consider two possible explanations of this lack of evidence: either that it stems from technical limitations due to a limited electrode coverage, or that such synchrony should be present not in the gamma band where it was originally predicted, but instead in lower frequency ranges (Cogitate et al., 2023b)—both of which defend a central claim of IIT by deducing the failure of a prediction from the failure of measurement-specific auxiliary hypothesis.

Similar observations can be made with regard to the failures of GNW's predictions. For example, one of the predictions that the protocol focused on was that "the content of experience should be present *both* in the prefrontal-parietal network and high-level sensory cortices" (Melloni et al., 2023, p.4, our emphasis). According to the findings, however, whereas category-specific information was found in PFC, no representation of identity or orientation could be detected. Proponents of GNW explain this finding by noting that the prefrontal code is not spatially clustered but is distributed over a large number of intermingled neurons (Cogitate et al., 2023a, p.28; Kapoor et al., 2022; Xie et al., 2022). Another possible explanation—which is still compatible with the core claim of GNW that associates the content of consciousness with the content of the perceptual representations that ignite the prefrontal-parietal network and are amplified and stabilized via long range feedback loops—is that low-level information does not get re-represented in PFC; instead, what PFC represents in this case is only a pointer to the relevant information stored in posterior areas (D'Esposito and Postle, 2015; Christophel et al., 2017; Xu, 2017; Scimeca et al., 2018; Barbey et al., 2013; Mackey et al., 2016; Ivanova et al., 2018). Importantly, both of these explanations deduce the failure of a specific prediction from the failure of implementation-specific auxiliary hypotheses.

manipulation as well (Fazekas and Nemeth, 2018). These functions are underlain by different neural mechanisms, and according to recent empirical findings, not all of these are associated with conscious awareness. For information monitoring and manipulation, the engagement of attention seems to be necessary and the information in question is stored by strong, persistent neural representations (Zylberberg and Strowbridge, 2017; van Vugt et al., 2018; Trübutschek et al., 2019; Sreenivasan et al., 2014). Purely maintaining information in working memory, however, is possible outside the focus of attention using neural representations with weaker persistent activity level (Kamiński and Rutishauser, 2020; Konecky et al., 2017), and even 'activity-silent' storage mechanisms relying on short-term changes of synaptic weights, i.e. without the persistent activation of corresponding neural representations (Trübutschek et al., 2017; Beukers et al., 2021; Wolff et al., 2017; Sprague et al., 2016; Rose et al., 2016; Stokes, 2015). This has led to the revision of the link between working memory and the global workspace: they are different (high-level) dimensions of the construct space with only partial overlap—only those items stored in working memory that are also in the focus of attention and are actively maintained are in the global workspace (Mashour et al., 2020; Trübutschek et al., 2017; Soto and Silvanto, 2014).

Similar construct-driven empirical work is required to explore the relation between further constructs and conscious experiences with as little commitment to existing theories of consciousness as possible. For example, the constructs of intensity and stability of neural representations have general, theory-neutral empirical counterparts at every level of the perceptual hierarchy: the level of neural firing rate and the temporal profile of the firing rate (Fazekas and Overgaard, 2018; Fazekas et al., 2020; Jagadisan and Gandhi, 2022). Furthermore, construct-driven experimentation could be the right tool to clarify the relations between other dimensions of the construct space as well, like for example representational re-description and recurrence in first-order processing (since if a higher-order system influences processing in the first-order system, then certain forms of recurrence are readily implemented) (Cleeremans et al., 2020). Fig. 3 illustrates the difference between the theory-driven and the construct-first approach.

3.4. New theoretical syntheses

The deconstruction phase that we proposed as a way of shifting the focus from theories to the multi-dimensional space of constructs associated with consciousness makes a consecutive *construction* phase possible. Once the relationships between constructs are uncovered and it is clarified how higher-level constructs map onto lower-level constructs, construct-driven empirical work can reveal 'go' and 'no-go' regions within the construct space, i.e. locations in this space defined by values of the different dimensions that correlate with the presence of consciousness and by values that don't. This would be the construct-first analogue of the theory-driven search for the neural correlates of consciousness. Instead of debates—such as for instance the one about whether the seat of consciousness is in the front or the back of the brain (Melloni et al., 2023; Boly et al., 2017; Odegaard et al., 2017; Raccah et al., 2021)—that derive 'top-down', so to speak, from complex theoretical positions, it would be a 'bottom-up' approach that would then allow for a new synthesis. Based on the values of the different dimensions, characteristic of the regions of the construct space that are compatible with the presence of consciousness, so-far unexplored theoretical possibilities could be explored and new theories could be constructed.

4. The construct-first approach and the future of consciousness research

The construct-first approach offers a novel path forward that can

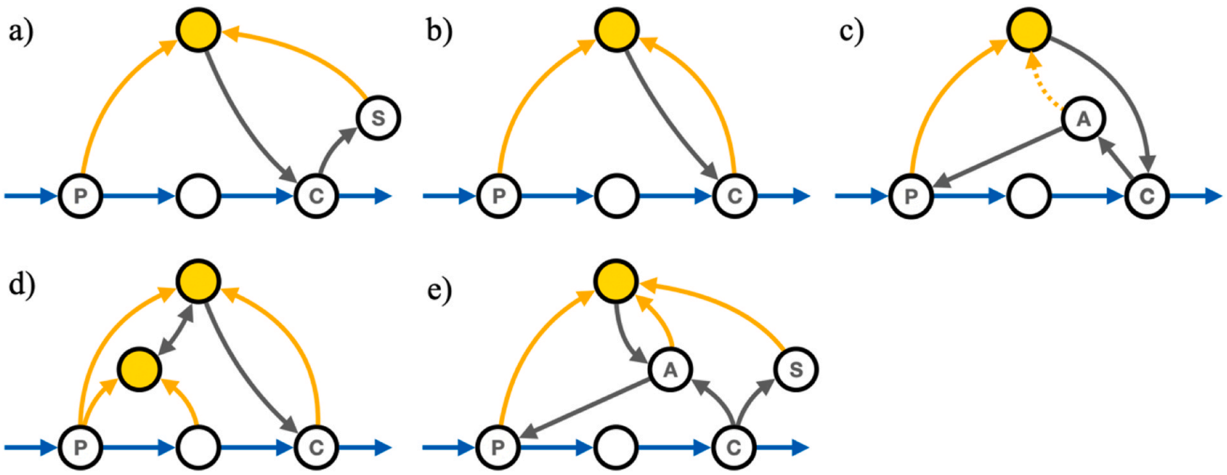

Fig. 2. Varieties of higher-order re-representation. a) Rosenthal's classical version of HOT; b) Perceptual Reality Monitoring Account; c) Higher-order State Space framework; d) Self-organizing Metarepresentational Account; e) Attention Schema Theory. P: perceptual representation; S: self; C: cognition; A: attention.

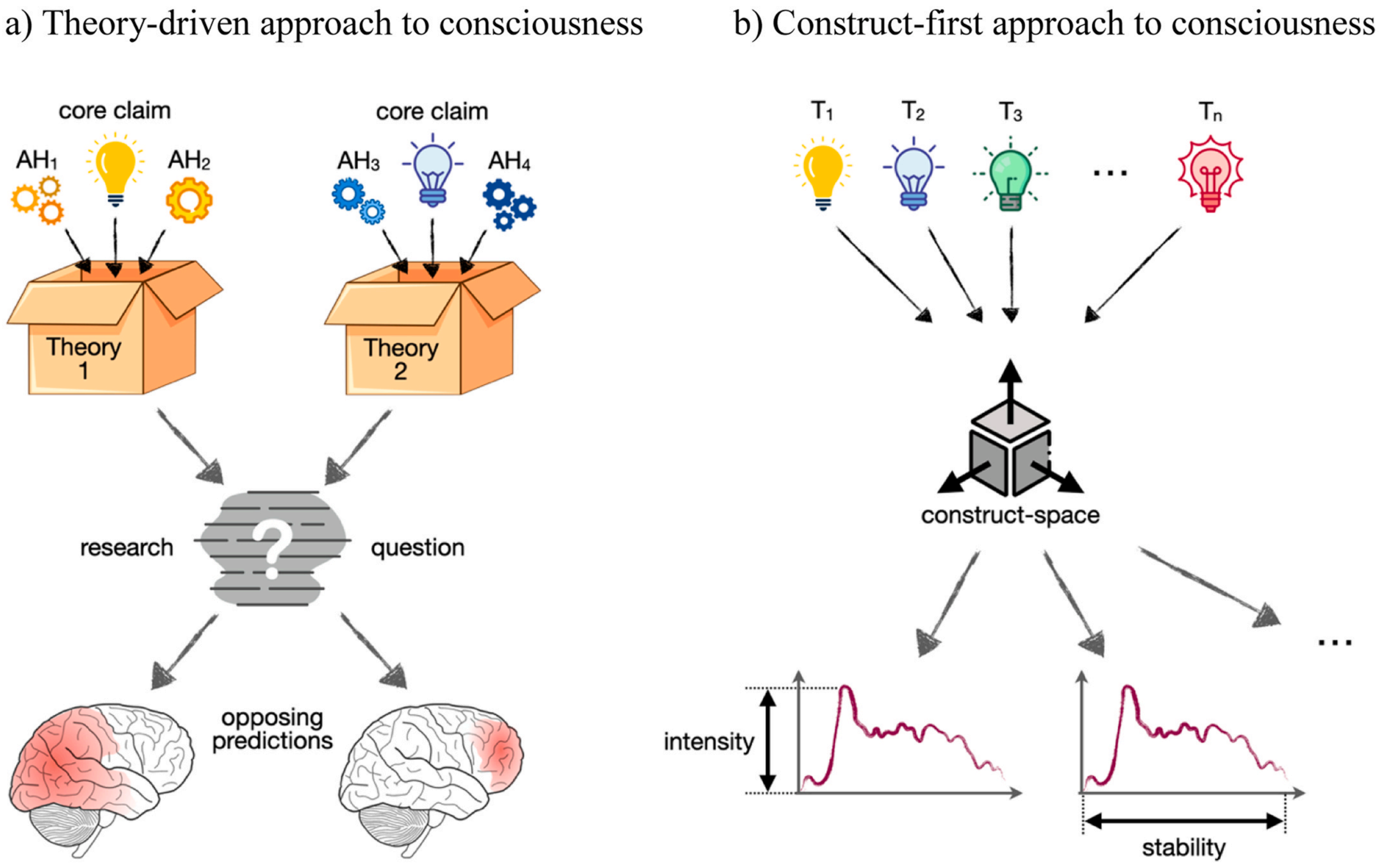

Fig. 3. Theory-driven vs. construct-first approaches to experimentation. a) Theory-driven approaches compare theory complexes (illustrated as boxes), where a central assumption (lightbulb) is supplemented by auxiliary hypotheses (cogwheels, 'AH'), i.e. further commitments and assumptions, such that they jointly become applicable to specific experimental scenarios and can make testable predictions. These auxiliary hypotheses also protect the central assumptions: in case a prediction is not supported by empirical findings, it is typically an auxiliary hypothesis and not the central assumption that gets revised. b) The construct-first approach focuses on the core claims of the theories ('T'), and tries to stripe auxiliary hypotheses away. Via analyzing the main constructs, the individual theories utilize to express their central assumptions, and exploring how they are related to further theoretical constructs that help clarify their meaning, the construct-first approach reveals a partly shared set of key constructs. In the abstract space of these constructs, regions compatible with the presence of consciousness can be uncovered through theory-neutral operationalizations of the constructs and consecutive experimentation. As an example, the purple graphs depict neural activity to illustrate the intensity and the stability of the underlying neural code representing a stimulus feature as the amplitude and the temporal profile (length of maintenance) of the response, respectively (Fazekas et al., 2020; Fazekas, 2023, see also main text). AHi: auxiliary hypothesis #i; Ti: the core claim of Theory i. (This figure has been designed using assets from Freepik.com).

resolve current challenges, mitigate problematic features of the discipline and offer new directions for further research.

4.1. Clarifying the relationships between theories

Different theories of consciousness rely on central constructs and core claims that are formulated using theory-specific vocabularies, and different theories often target different aspects of consciousness. Hence the relationships between alternative explanatory attempts are especially unclear.

The construct-first approach proposes that this challenge can be resolved by analyzing the major high-level constructs that the different theories associate consciousness with in terms of lower-level constructs that form a more common, shared vocabulary. Understanding these lower-level constructs as dimensions of a construct space, such an analysis results in mapping existing theories onto different regions of this space and reveals their relationships along the dimensions.

4.2. Decreasing the reliance on auxiliary premises

The core commitments of extant theories of consciousness are conceptually distant from observational predictions and thus empirical efforts need to rely heavily on auxiliary premises.

The construct-first approach proposes that this problem can be mitigated by refocusing the empirical efforts to testing hypotheses based on lower-level constructs, and studying how the different values of the different dimensions of the construct space defined by these lower-level constructs are related to the presence or absence of consciousness. The lower-level constructs in question can be operationalized with little or no commitment to particular theories of consciousness, and thus construct-driven empirical findings can have more direct relevance to the core claims of different theories and can serve as the basis of exploring uncharted theoretical positions.

4.3. Taking the overlaps seriously

The theory-based approach focuses on the differences between the numerous theories and tries to derive contradicting predictions. A new research direction that the construct-first approach can offer is a reorientation towards the overlaps motivated by the idea that the success of diverse theoretical accounts in finding empirical support despite their surface differences might be due to the fact that there is a common set of phenomena covered by these accounts.

For example, the overlaps in intensity and stability demonstrated above can be interpreted as revealing that the neural processes that major theories of consciousness identify as crucial from the perspective of the occurrence of conscious experiences are, in fact, either *preconditions* or *consequences* of the formation of strong and stable neural representations. In an environment that provides rapidly changing input signals, only recurrent systems (that can decouple from input conditions) can achieve stability, hence local recurrence is a precondition. The integration of strong and stable representations into a global workspace is a consequence of their strength and stability—and also what enables their longer-term stability; therefore, it is either a precondition or a consequence, depending on what levels of strength and stability will turn out to be required for the occurrence of conscious experiences. Similarly, the re-description of these representations by monitoring networks is a consequence triggered by the strength and stability of the representations in question. Finally, although a system, in principle, integrates information even if its parts are not active (see Box 2), maximal information integration might be interpreted as an emergent property of a network that can amplify and stabilize perceptual signals in the most effective way.

From here, one possibility for how the field could move forward is to consider that the strength and stability of these signals might be more important for the occurrence of consciousness than the processes that current theories of consciousness focus on (see (Zeki, 2007; O'Brien and Opie, 1999) for similar claims resulting from different motivations).

5. The possibility of multiple non-overlapping construct-space-correlates of consciousness

Another option is that the empirical study of how the different values of the different dimensions of the construct space are related to consciousness will reveal that there are multiple non-overlapping regions in the construct space that all correlate with the presence of consciousness.

This could point towards the possibility that consciousness can be genuinely multiply realizable at the level of the known dimensions of the construct space, i.e. that different mechanisms can equally give rise to conscious experiences. The result would be a pluralistic view with regard to the processes underlying consciousness (see (He, 2023) for a recent exposition of such a view).

That is, the construct-first approach is a general research program that although recommends a fundamental theoretical and empirical reorientation, can nevertheless respect both the assumption of existing theories of consciousness that a unifying account might be possible, and the spirit of pluralistic views according to which different neural processing architectures might be responsible for different types of conscious experiences.

6. Concluding remarks

Confirmation biases permeate consciousness science at many different levels (Yaron et al., 2022). Our goal in this opinion piece was to argue for a reorientation of theoretical and empirical efforts from a theory-driven approach to a construct-first approach. The construct-first approach invites a 'back to basics' attitude that focuses on dissecting existing theories to analyze their central assumptions and the theoretical constructs they utilize. Using these constructs as dimensions of a construct space offers a new tool that can be a true alternative of existing approaches that focus on theory comparison from the perspectives of empirical predictions (Melloni et al., 2021, 2023; Doerig et al., 2021; Cogitate et al., 2023a), explanatory targets (Seth and Bayne, 2022; Northoff and Lamme, 2020) or specific cases and conditions like infant consciousness (Bayne et al., 2023) or mental disorder (Stefanelli, 2023). The construct-first approach neither assumes nor is motivated to establish the primacy of any one existing theoretical framework over the others. On the contrary, the construct space it proposes offers am impartial viewpoint for exploring the relationships between the central assumptions of the theories. Moreover, formulating empirical hypotheses that focus on the individual constructs and favoring theory-neutral methodologies offer a way to reduce biases. The resulting framework is bottom-up, exploratory, allows for new theoretical syntheses, offers motivation for new research directions, and holds the promise of a data-driven turn within consciousness science.

Funding statement

Peter Fazekas was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement (No. 754513), and The Aarhus University Research Foundation. Axel Cleeremans is a research director with the F.R.S.-FNRS (Belgium) and a senior fellow of the Canadian Institute for Advanced Research. This work was partially supported by ERC AdG Grant #101055060 "EXPERIENCE" to Axel Cleeremans.

Declaration of Competing Interest

The authors declare no conflicts of interest.

Data availability

No data was used for the research described in the article.

References

Albantakis, L., et al., 2022. Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *Arxiv* 1–53.

Baars, B.J. (1988) *A Cognit*ive Theory of Consciousness Cambridge University Press.

Baars, B.J., 1998. Metaphors of consciousness and attention in the brain. *Trends Neurosci.* 23, 58–62.

Baars, B.J., et al., 2021. Global Workspace Theory (GWT) and Prefrontal Cortex: Recent Developments. Front. Psychol. 12 https://doi.org/10.3389/fpsyg.2021.749868.

Baars, B.J., Franklin, S., 2003. How conscious experience and working memory interact. Trends Cogn. Sci. 7, 166–172. https://doi.org/10.1016/S1364-6613(03)00056-1.

Baddeley, A., 2003. Working memory: looking back and looking forward. Nat. Rev. Neurosci. 4, 829. https://doi.org/10.1038/nrn1201.

Baddeley, A., 2012. Working Memory: Theories, Models, and Controversies. Annu. Rev. Psychol. 63, 1–29. https://doi.org/10.1146/annurev-psych-120710-100422.

Barbey, A.K., et al., 2013. Dorsolateral prefrontal contributions to human working memory. Cortex 49, 1195–1205.

Bayne, T., 2018. On the axiomatic foundations of the integrated information theory of consciousness. niy007 Neurosci. Conscious. 2018. https://doi.org/10.1093/nc/niy007.

Bayne, T., et al., 2023. Consciousness in the cradle: on the emergence of infant experience. Trends Cogn. Sci. https://doi.org/10.1016/j.tics.2023.08.018.

Beukers, A.O., et al., 2021. Is Activity Silent Working Memory Simply Episodic Memory? Trends Cogn. Sci. 25, 284–293. https://doi.org/10.1016/j.tics.2021.01.003.

Block, N., 2005. Two neural correlates of consciousness. Trends Cogn. Sci. 9, 46–52. https://doi.org/10.1016/j.tics.2004.12.006.

Block, N., 2011. Perceptual consciousness overflows cognitive access. Trends Cogn. Sci. 15, 567–575. https://doi.org/10.1016/j.tics.2011.11.001.

Block, N., 2014. Rich conscious perception outside focal attention. Trends Cogn. Sci. 18, 445–447.

Block, N., 2023. The Border Between Seeing and Thinking. Oxford University Press,.

Boly, M., et al., 2017. Are the Neural Correlates of Consciousness in the Front or in the Back of the Cerebral Cortex? Clinical and Neuroimaging Evidence. J. Neurosci. 37, 9603–9613.

Bronfman, Z.Z., et al., 2014. We See More Than We Can Report: "Cost Free" Color Phenomenality Outside Focal Attention. Psychol. Sci. 25, 1394–1403. https://doi.org/10.1177/0956797614532656.

Brown, R., 2015. The HOROR theory of phenomenal consciousness. Philos. Stud. 172, 1783–1794.

Brown, R., et al., 2019. Understanding the higher-order approach to consciousness. Trends Cogn. Sci. 23, 754–768. https://doi.org/10.1016/j.tics.2019.06.009.

Carruthers, P., 2000. Phenomenal Consciousness: A Naturalistic Theory. Cambridge University Press,.

Chalmers, D., 1995. Facing up to the problem of Consciousness. J. Conscious. Stud. 2, 200–219.

Chalmers, D. (1996) The conscious mind: in search of a fundamental theory (Philosophy of mind series, Oxford University Press.

Christophel, T.B., et al., 2017. The Distributed Nature of Working Memory. Trends Cogn. Sci. 21, 111–124.

Clark, A., 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci. 36, 181–204. https://doi.org/10.1017/S0140525X12000477.

Clark, A., 2015. Surfing Uncertainty: Prediction, Action, and the Embodied Mind. Oxford University Press,.

Clark, A., 2019. Consciousness as Generative Entanglement. J. Philos. 116, 645–662.

Cleeremans, A., 2011. The radical plasticity thesis: how the brain learns to be conscious. Front. Psychol. 2, 1–12. https://doi.org/10.3389/fpsyg.2011.00086.

Cleeremans, A., et al., 2020. Learning to Be Conscious. Trends Cogn. Sci. 24, 112–123. https://doi.org/10.1016/j.tics.2019.11.011.

Cleeremans, A., Tallon-Baudry, C., 2022. Consciousness matters: phenomenal experience has functional value. Neurosci. Conscious. 2022, niac007 https://doi.org/10.1093/nc/niac007.

Cogitate, C., et al., 2023a. An adversarial collaboration to critically evaluate theories of consciousness. *2023.2006.2023* bioRxiv, 546249. https://doi.org/10.1101/2023.06.23.546249.

Cogitate, C., et al., 2023b. Supplementary information to "An adversarial collaboration to critically evaluate theories of consciousness", 2023.2006.2023 bioRxiv, 546249. https://doi.org/10.1101/2023.06.23.546249.

Cohen, M., Dennett, D., 2011. Consciousness cannot be separated from function. Trends Cogn. Sci. 15, 358–364. https://doi.org/10.1016/j.tics.2011.06.008.

Deco, G., et al., 2021. Revisiting the global workspace orchestrating the hierarchical organization of the human brain. Nat. Hum. Behav. 5, 497–511. https://doi.org/10.1038/s41562-020-01003-6.

Dehaene, S., et al., 1998. A neuronal model of a global workspace in effortful cognitive tasks. Proc. Natl. Acad. Sci. 95, 14529.

Dehaene, S., et al., 2003. A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proc. Natl. Acad. Sci. USA 100, 8520–8525. https://doi.org/10.1073/pnas.1332574100.

Dehaene, S., et al., 2006. Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends Cogn. Sci. 10, 204–211. https://doi.org/10.1016/j.tics.2006.03.007.

Dehaene, S., et al., 2014. Toward a computational theory of conscious processing. Curr. Opin. Neurobiol. 25, 76–84. https://doi.org/10.1016/j.conb.2013.12.005.

Dehaene, S., et al., 2017. What is consciousness, and could machines have it? Science 358, 486.

Dehaene, S., Changeux, J., 2011. Experimental and Theoretical Approaches to Conscious Processing. Neuron 70, 200–227. https://doi.org/10.1016/j.neuron.2011.03.018.

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

Dennett, D. (1991) Consciousness explained 1st edn), Little, Brown and Co.

Dennett, D., 2016. Illusionism as the Obvious Default Theory of Consciousness. J. Conscious. Stud. 23, 65–72.

D'Esposito, M., Postle, B., 2015. The Cognitive Neuroscience of Working Memory. Annu. Rev. Psychol. 66, 115–142.

Doerig, A., et al., 2021. Hard criteria for empirical theories of consciousness. Cogn. Neurosci. 12, 41–62. https://doi.org/10.1080/17588928.2020.1772214.

Duhem, P. (1914/1954) The Aim and Structure of Physical Theory Princeton University Press.

Edelman, G.M. (1987) *Neural Darwinism: The Theory of Neuronal Group Selection* Basic Books.

Eriksson, J., et al., 2015. Neurocognitive architecture of working memory. Neuron 88, 33–46. https://doi.org/10.1016/j.neuron.2015.09.020.

Fazekas, P., 2009. Reconsidering the Role of Bridge Laws In Inter-Theoretical Reductions. Erkenntnis 71, 303–322. https://doi.org/10.1007/s10670-009-9181-y.

Fazekas, P., et al., 2020. Perceptual representations and the vividness of stimulus-triggered and stimulus-independent experiences. Perspect. Psychol. Sci. 15, 1200–1213.

Fazekas, P., 2021. Hallucinations as intensified forms of mind-wandering. Philos. Trans. R. Soc. B: Biol. Sci. 376, 20190700.

Fazekas, P., 2022. Flat mechanisms: a reductionist approach to levels in mechanistic explanations. Philos. Stud. https://doi.org/10.1007/s11098-021-01764-4.

Fazekas, P., 2023. Vividness and content. *Mind Lang.* Online first, 02 March 2023. https://doi.org/10.1111/mila.12455.

Fazekas, P., Nemeth, G., 2018. Dream experiences and the neural correlates of perceptual consciousness and cognitive access. Philos. Trans. R. Soc. B: Biol. Sci. 373, 20170356.

Fazekas, P., Overgaard, M., 2018. A Multi-Factor Account of Degrees of Awareness. Cogn. Sci. 42, 1833–1859. https://doi.org/10.1111/cogs.12478.

Fleming, S.M., 2020. Awareness as inference in a higher-order state space. Neurosci. Conscious. 2020, niaa011 https://doi.org/10.1093/nc/niaa011.

Gennaro, R.J., 2012. *Consciousness Paradox: Consciousness, Concepts, and Higher-Order Thoughts* MIT Press. The,.

Graziano, M.S.A., 2020. Consciousness and the attention schema: Why it has to be right. Cogn. Neuropsychol. 37, 224–233. https://doi.org/10.1080/02643294.2020.1761782.

Graziano, M.S.A., et al., 2020. Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. Cogn. Neuropsychol. 37, 155–172. https://doi.org/10.1080/02643294.2019.1670630.

Graziano, M.S.A., 2022. A conceptual framework for consciousness. Proc. Natl. Acad. Sci. 119, e2116933119 https://doi.org/10.1073/pnas.2116933119.

Graziano, M.S.A., Webb, T.W., 2015. The attention schema theory: a mechanistic account of subjective awareness. *Front. Psychol.* 6. https://doi.org/10.3389/fpsyg.2015.00500.

Hameroff, S. and Penrose, R. (1995) Orchestrated reduction of quantum coherence in brain microtubules: a model for consciousness? In Scales in Conscious Experience, Is the brain too important to be left to specialists to study? (King, J. and Pribram, K., eds), pp. 243–274, Lawrence Erlbaum.

Hameroff, S., Penrose, R., 2014. Consciousness in the universe: A review of the 'Orch OR' theory. Phys. Life Rev. 11, 39–78. https://doi.org/10.1016/j.plrev.2013.08.002.

Hameroff, S.R. and Penrose, R. (2016) Consciousness in the universe: An updated review of the "ORCH OR" theory. In Biophysics of Consciousness, pp. 517–599, WORLD SCIENTIFIC.

He, B.J., 2023. Towards a pluralistic neurobiological understanding of consciousness. Trends Cogn. Sci. 27, 420–432. https://doi.org/10.1016/j.tics.2023.02.001.

Hohwy, J., 2012. Attention and conscious perception in the hypothesis testing brain. Front. Psychol. 3, 1–14. https://doi.org/10.3389/fpsyg.2012.00096/abstract.

Hohwy, J., 2013. The Predictive Mind. Oxford University Press,.

Ivanova, M.V., et al., 2018. Neural mechanisms of two different verbal working memory tasks: A VLSM study. Neuropsychologia. https://doi.org/10.1016/j.neuropsychologia.2018.03.003.

Jagadisan, U.K. and Gandhi, N.J.. (2022) Population temporal structure supplements the rate code during sensorimotor transformations. Current Biology 32, 1010–1025. e1019. 10.1016/j.cub.2022.01.015.

Kamiński, J., Rutishauser, U., 2020. Between persistently active and activity-silent frameworks: novel vistas on the cellular basis of working memory. Ann. N. Y. Acad. Sci. 1464, 64–75. https://doi.org/10.1111/nyas.14213.

Kapoor, V., et al., 2022. Decoding internally generated transitions of conscious contents in the prefrontal cortex without subjective reports. Nat. Commun. 13, 1535 https://doi.org/10.1038/s41467-022-28897-2.

Koch, C. *et al.* (2016a) Neural correlates of consciousness: progress and problems. Nature Reviews Neuroscience 17, 307–321. 10.1038/nrn.2016.22https://www.nature.com/articles/nrn.2016.22#supplementary-information.

Koch, C. *et al.* (2016b) Posterior and anterior cortex — where is the difference that makes the difference? Nature Reviews Neuroscience 17, 666–666. 10.1038/nrn.2016.105.

Konecky, R.O., et al., 2017. Monkey prefrontal neurons during Sternberg task performance: full contents of working memory or most recent item? J. Neurophysiol. 117, 2269–2281. https://doi.org/10.1152/jn.00541.2016.

Kriegel, U., 2009. Subjective Consciousness: A Self-Representational Theory. Oxford University Press UK,.

Lakatos, I., 1970. Falsification and the Methodology of Scientific Research Programmes. In: Musgrave, A., Lakatos, I. (Eds.), *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965*. Cambridge University Press, pp. 91–196.

Lamme, V., 2003. Why visual attention and awareness are different. Trends Cogn. Sci. 7, 12–18.

Lamme, V.A.F., 2006. Towards a true neural stance on consciousness. Trends Cogn. Sci. 10, 494–501. https://doi.org/10.1016/j.tics.2006.09.001.

Lamme, V.A.F., 2010. How neuroscience will change our view on consciousness. Cogn. Neurosci. 1, 204–220. https://doi.org/10.1080/17588921003731586.

Lamme, V.A.F., 2018. Challenges for theories of consciousness: Seeing or knowing, the missing ingredient, and how to deal with panpsychism. Philos. Trans. R. Soc. B: Biol. Sci. 373, 20170344.

Lamme, V.A.F., Roelfsema, P.R., 2000. The distinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci. 23, 571–579. https://doi.org/10.1016/S0166-2236(00)01657-X.

Lau, H., 2019. Consciousness, Metacognition, & Perceptual Reality Monitoring. PsyArXiv. https://doi.org/10.31234/osf.io/ckbyf.

Lau, H., et al., 2022. The mnemonic basis of subjective experience. Nat. Rev. Psychol. 1, 479–488. https://doi.org/10.1038/s44159-022-00068-6.

Lau, H., Rosenthal, D., 2011. Empirical support for higher-order theories of conscious awareness. Trends Cogn. Sci. 15, 365–373. https://doi.org/10.1016/j.tics.2011.05.009.

LeDoux, J.E., Brown, R., 2017. A higher-order theory of emotional consciousness. Proc. Natl. Acad. Sci. 114, E2016–E2025. https://doi.org/10.1073/pnas.1619316114.

Lycan, W.G. (1996) Consciousness and Experience MIT Press.

Mackey, W.E., et al., 2016. Human Dorsolateral Prefrontal Cortex Is Not Necessary for Spatial Working Memory. J. Neurosci. 36, 2847–2856.

Mashour, G.A., et al., 2020. Conscious Processing and the Global Neuronal Workspace Hypothesis. Neuron 105, 776–798. https://doi.org/10.1016/j.neuron.2020.01.026.

Melloni, L., et al., 2021. Making the hard problem of consciousness easier. Science 372, 911–912. https://doi.org/10.1126/science.abj3259.

Melloni, L., et al., 2023. An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory. PLOS ONE 18, e0268577. https://doi.org/10.1371/journal.pone.0268577.

Michel, M., Morales, J., 2020. Minority reports: Consciousness and the prefrontal cortex. Mind Lang. 35, 493–513. https://doi.org/10.1111/mila.12264.

Northoff, G., Lamme, V., 2020. Neural signs and mechanisms of consciousness: Is there a potential convergence of theories of consciousness in sight? Neurosci. Biobehav. Rev. 118, 568–587. https://doi.org/10.1016/j.neubiorev.2020.07.019.

O'Brien, G., Opie, J., 1999. A connectionist theory of phenomenal experience. Behav. Brain Sci. 22, 127–148. https://doi.org/10.1017/S0140525X9900179X.

Odegaard, B., et al., 2017. Should a Few Null Findings Falsify Prefrontal Theories of Conscious Perception? J. Neurosci. 37, 9593–9602.

Oizumi, M., et al., 2014. From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. PLoS Comput. Biol. 10, e1003588 https://doi.org/10.1371/journal.pcbi.1003588.

O'Regan, K., Noë, A., 2001. A sensorimotor account of vision and visual consciousness. Behav. Brain Sci. 24, 883–917.

Phillips, I., 2018. The Methodological Puzzle of Phenomenal Consciousness. Philos. Trans. R. Soc. B: Biol. Sci. 373, 20170347.

Pinto, Y., et al., 2013. Fragile visual short-term memory is an object-based and location-specific store. Psychon. Bull. Rev. 20, 732–739. https://doi.org/10.3758/s13423-013-0393-4.

Popper, K., 1959. The logic of scientific discovery. Basic Books.

Prinz, J., 2007. The Intermediate Level Theory of Consciousness. In: Velmans, M., Schneider, S. (Eds.), *The Blackwell Companion to Consciousness*. Blackwell, pp. 247–260.

Prinz, J., 2012. The Conscious Brain: How Attention Engenders Experience. Oxford University Press,.

Quine, W.V., 1951. Main Trends in Recent Philosophy: Two Dogmas of Empiricism. Philos. Rev. 60, 20–43. https://doi.org/10.2307/2181906.

Raccah, O., et al., 2021. Does the Prefrontal Cortex Play an Essential Role in Consciousness? Insights from Intracranial Electrical Stimulation of the Human Brain. J. Neurosci. 41, 2076–2087. https://doi.org/10.1523/jneurosci.1141-20.2020.

Rose, N.S., et al., 2016. Reactivation of latent working memories with transcranial magnetic stimulation. Science 354, 1136–1139. https://doi.org/10.1126/science.aah7011.

Rosenthal, D.M. (2005) Consciousness and Mind Oxford University Press UK.

Scimeca, J.M., et al., 2018. Reaffirming the Sensory Recruitment Account of Working Memory. Trends Cogn. Sci. 22, 190–192.

Seth, A.K. (2021) *Being You: A New Science of Consciousness* Faber & Faber.

Seth, A.K., Bayne, T., 2022. Theories of consciousness. Nat. Rev. Neurosci. 23, 439–452. https://doi.org/10.1038/s41583-022-00587-4.

Seth, A.K., Tsakiris, M., 2018. Being a Beast Machine: The Somatic Basis of Selfhood. Trends Cogn. Sci. 22, 969–981. https://doi.org/10.1016/j.tics.2018.08.008.

Sligte, I.G., et al., 2008. Are there multiple visual short-term memory stores? PLOS ONE 3, e1699. https://doi.org/10.1371/journal.pone.0001699.

Soto, D., Silvanto, J., 2014. Reappraising the relationship between working memory and conscious awareness. Trends Cogn. Sci. 18, 520–525. https://doi.org/10.1016/j.tics.2014.06.005.

Sprague, T.C., et al., 2016. Restoring Latent Visual Working Memory Representations in Human Cortex. Neuron 91, 694–707. https://doi.org/10.1016/j.neuron.2016.07.006.

Sreenivasan, K.K., et al., 2014. Revisiting the role of persistent neural activity during working memory. Trends Cogn. Sci. 18, 82–89.

Stefanelli, R., 2023. Theories of consciousness and psychiatric disorders – A comparative analysis. Neurosci. Biobehav. Rev. 152, 105204 https://doi.org/10.1016/j.neubiorev.2023.105204.

Stokes, M.G., 2015. Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn. Sci. 19, 394–405. https://doi.org/10.1016/j.tics.2015.05.004.

Tononi, G., 2004. An information integration theory of consciousness. BMC Neurosci. 5 (1), 22. https://doi.org/10.1186/1471-2202-5-42.

Tononi, G., et al., 2016. Integrated information theory: from consciousness to its physical substrate. Nat. Rev. Neurosci. 17, 450–461. https://doi.org/10.1038/nrn.2016.44.

Trübutschek, D., et al., 2017. A theory of working memory without consciousness or sustained activity. eLife 6, e23871. https://doi.org/10.7554/eLife.23871.

Trübutschek, D., et al., 2019. Probing the limits of activity-silent non-conscious working memory. Proc. Natl. Acad. Sci. 116, 14358–14367. https://doi.org/10.1073/pnas.1820730116.

Vandenbroucke, A.R.E., et al., 2015. Neural Correlates of Visual Short-term Memory Dissociate between Fragile and Working Memory Representations. J. Cogn. Neurosci. 27, 2477–2490. https://doi.org/10.1162/jocn_a_00870.

van Vugt, B., et al., 2018. The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. Science 360, 537–542. https://doi.org/10.1126/science.aar7186.

Wilterson, A.I., et al., 2021. Attention, awareness, and the right temporoparietal junction. Proc. Natl. Acad. Sci. 118, e2026099118 https://doi.org/10.1073/pnas.2026099118.

Wolff, M.J., et al., 2017. Dynamic hidden states underlying working-memory-guided behavior. Nat. Neurosci. 20, 864–871. https://doi.org/10.1038/nn.4546.

Xie, Y., et al., 2022. Geometry of sequence working memory in macaque prefrontal cortex. Science 375, 632–639. https://doi.org/10.1126/science.abm0204.

Xu, Y., 2017. Reevaluating the Sensory Account of Visual Working Memory Storage. Trends Cogn. Sci. 21, 794–815. https://doi.org/10.1016/j.tics.2017.06.013.

Yaron, I., et al., 2022. The ConTraSt database for analysing and comparing empirical studies of consciousness theories. Nat. Hum. Behav. 6, 593–604. https://doi.org/10.1038/s41562-021-01284-5.

Zeki, S., 2003. The disunity of consciousness. Trends Cogn. Sci. 7, 214–218. https://doi.org/10.1016/S1364-6613(03)00081-0.

Zeki, S., 2007. A theory of micro-consciousness. *Black Companion Conscious.* 580–588.

Zylberberg, J., Strowbridge, B.W., 2017. Mechanisms of persistent activity in cortical circuits: possible neural substrates for working memory. Annu. Rev. Neurosci. 40, 603–627. https://doi.org/10.1146/annurev-neuro-070815-014006.