Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

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

# Unpacking the complexities of consciousness: Theories and reflections

Liad Mudrik [a,b,c,*], Melanie Boly [d], Stanislas Dehaene [c,e,f], Stephen M. Fleming [c,g,h,i], Victor Lamme [j], Anil Seth [c,k], Lucia Melloni [c,l]

[a] School of Psychological Sciences, Tel Aviv University, Israel
[b] Sagol School of Neuroscience, Tel Aviv University, Israel
[c] Program on Brain, Mind, and Consciousness, Canadian Institute for Advanced Research, Toronto, Canada
[d] University of Wisconsin-Madison, Madison, WI, USA
[e] Institut National de la Santé et de la Recherche Médicale (INSERM), Gif-sur-Yvette, France
[f] Collège de France, Paris, France
[g] Department of Experimental Psychology, University College London, England, United Kingdom
[h] Functional Imaging Laboratory, University College London, London, England, United Kingdom
[i] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, England, United Kingdom
[j] Amsterdam Brain and Cognition (ABC), Dept of Psychology, University of Amsterdam, Amsterdam, the Netherlands
[k] Sussex Centre for Consciousness Science, Department of Informatics, University of Sussex, Brighton, United Kingdom
[l] Max Planck Institute for Empirical Aesthetics, Frankfurt am Main Germany

ARTICLE INFO

*Keywords:*
Theories of consciousness
Global neuronal workspace Theory
Integrated information theory recurrent processing theory
Higher order thought theories
Predictive processing theory

ABSTRACT

As the field of consciousness science matures, the research agenda has expanded from an initial focus on the neural correlates of consciousness, to developing and testing theories of consciousness. Several theories have been put forward, each aiming to elucidate the relationship between consciousness and brain function. However, there is an ongoing, intense debate regarding whether these theories examine the same phenomenon. And, despite ongoing research efforts, it seems like the field has so far failed to converge around any single theory, and instead exhibits significant polarization. To advance this discussion, proponents of five prominent theories of consciousness—Global Neuronal Workspace Theory (GNWT), Higher-Order Theories (HOT), Integrated Information Theory (IIT), Recurrent Processing Theory (RPT), and Predictive Processing (PP)—engaged in a public debate in 2022, as part of the annual meeting of the Association for the Scientific Study of Consciousness (ASSC). They were invited to clarify the explananda of their theories, articulate the core mechanisms underpinning the corresponding explanations, and outline their foundational premises. This was followed by an open discussion that delved into the testability of these theories, potential evidence that could refute them, and areas of consensus and disagreement. Most importantly, the debate demonstrated that at this stage, there is more controversy than agreement between the theories, pertaining to the most basic questions of what consciousness is, how to identify conscious states, and what is required from any theory of consciousness. Addressing these core questions is crucial for advancing the field towards a deeper understanding and comparison of competing theories.

## 1. Introduction

In recent years, as the field of consciousness studies matures, the research program is gradually transitioning from its initial focus on the search for the neural correlates of consciousness (Crick and Koch, 2003), and the accumulation of empirical observations, to developing theories of consciousness (ToCs). A shared research program of these ToCs is to provide an explicit account of the physical and psychological mechanisms of conscious experience (e.g., Doerig et al., 2020; Kuhn, 2024; Sattin et al., 2021; Seth and Bayne, 2022; Storm et al., 2024; Yaron et al., 2022). That is, these theories all strive to explain the mechanisms that underlie conscious, phenomenal experiences: many of them accordingly target what differentiates between processes that are accompanied by a conscious experience and those that are not, and some also ask what differentiates between two conscious experiences, though this question has been by and large less thoroughly dealt with thus far

* Corresponding author at: School of Psychological Sciences, Tel Aviv University, Israel.
*E-mail address:* mudrikli@tauex.tau.ac.il (L. Mudrik).

https://doi.org/10.1016/j.neubiorev.2025.106053
Received 29 October 2024; Received in revised form 31 January 2025; Accepted 5 February 2025

(Seth and Bayne, 2022). Yet despite substantial empirical and theoretical progress, there is still no clear path towards a unified account (Lepauvre and Melloni, 2021; Seth, 2018; Seth and Bayne, 2022).

This failure to establish an agreed-upon account might stem from at least two reasons. First, the explanandum of the theories (i.e., the phenomena they aim at explaining) might differ (Sattin et al., 2021; Signorelli et al., 2021). If so, seeking a unified theory is likely to fail, and a more promising strategy would be to first delineate the different explananda and establish the boundary conditions for these theories, as a step towards developing a more integrative account (Evers et al., 2024). Second, the methodologies and measures used in the empirical work around the theories has been shown to differ, potentially leading to a confirmatory bias in the literature. Indeed, a recent study showed that support for ToCs can be predicted from the methodological choices of the experiments, irrespective of the obtained result (Yaron et al., 2022; see also Promet and Bachkann, 2022).

In response to this fragmented landscape, the Association for the Scientific Study of Consciousness (ASSC) in 2022 hosted the "Great Consciousness Debate," featuring proponents of four theories: Stanislas Dehaene presented the Global Neuronal Workspace Theory (GNWT; Dehaene and Naccache, 2001; Mashour et al., 2020), Melanie Boly presented the Integrated Information Theory (IIT; Albantakis et al., 2023; Tononi, 2008), Stephen Fleming presented Higher Order theories (HOTs; Brown et al., 2019; Lau and Rosenthal, 2011), and Victor Lamme presented the Recurrent Processing Theory (RPT; Lamme, 2006; Lamme and Roelfsema, 2000). Anil Seth had planned to present the Predictive Processing Theory (PPT; Hohwy and Seth, 2020; Seth, 2021), but was unable to attend. Thus, his contribution to this paper was provided in retrospect. Importantly, as this paper is based on the 2022 debate, it only features the five theories invited to participate in that debate. These theories only subtend some of the theoretical landscape in the field, which includes other theories (for reviews, see again Doerig et al., 2020; Kuhn, 2024; Sattin et al., 2021; Seth and Bayne, 2022; Storm et al., 2024). For example, all the theories discussed here are primarily corticocentric, as they rely mainly on cortical mechanisms (and some also on underlying thalamocortical loops ), unlike other theories that focus on subcortical substrates (e.g., Merker, 2007; Ward, 2011; see also Solms, 2019). Accordingly, we do not consider them to fully represent the breadth of theories and views, and we welcome similar future debates amongst other prominent theories.

The debate aimed at updating the academic community on each theory, identifying commonalities that might pave the way towards a unified theory, and highlighting conceptual differences to guide empirical testing. Each proponent was asked to define their theory's explananda, its core mechanisms and foundational premises, as well as to suggest results which, if established, could potentially alter their stance on their theory.

In this paper, we first provide summaries of each proponent's presentation. These summaries are largely based on the talks given during the debate, with few modifications introduced during the writing process by each of the proponents to improve readability. Table 1 further provides a concise summary prepared by the proponents, outlining for each theory its core focus, metaphysical commitments, and explanations of non-conscious processes and key neural mechanisms. It also highlights assumptions regarding the interplay between consciousness, attention, cognition, and evolutionary factors, as well as potential challenges for each theory. We then follow with a section written primarily by the mediators of the debate, highlighting what they considered key insights and challenges drawn from the open discussion between the proponents. There, we primarily focus on the ongoing effort to clarify concepts, detect points of disagreement as well as find common ground among the leading theories of consciousness (for other attempts to do so, see Chis-Ciure et al., 2024; Doerig et al., 2020; Signorelli et al., 2021; Storm et al., 2024).

## 2. Presentation of the debated theories

### 2.1. Global neuronal workspace / Stanislas Dehaene

The Global Neuronal Workspace theory (GNWT), which was proposed more than 25 years ago (Dehaene et al., 1998), holds that what we call consciousness is a computational property, characteristic of a certain type of information processing (Dehaene and Naccache, 2001). Tying the theory with current developments in Artificial Intelligence (AI) and Machine Learning (ML), the theory holds that in principle, the architecture underlying conscious processing could be implemented in machines. However, consciousness is a polysemic term. To clarify the situation, together with Sid Kouider and Hakwan Lau, we distinguished three different types of computations, with the latter two representing two different dimensions of consciousness (Dehaene et al., 2017): C0, which is characteristic of many existing algorithms, refers to *nonconscious computations* (i.e., information processing operations that can be carried out without consciousness). C1 represents the level of *conscious access* where a workspace globally broadcasts and amplifies a specific piece of information, akin to a routing system. C2 refers to the ability of a system to represent itself. A system with both self-representation and conscious access to it will have knowledge about itself: it knows what it knows, and it knows what it doesn't know.

What is the scope of the GNW theory? The theory aims to account for two different things: One is *conscious processing* (i.e., which processes require or do not require consciousness), and the other is *conscious phenomenology* (i.e., what a participant experiences at a given moment). In the last 25 years, our understanding of conscious vs. nonconscious processing has expanded substantially. Many experiments have revealed that nonconscious processing is extensive, so that much can be done without awareness (Kouider and Dehaene, 2007; Mudrik and Deouell, 2022; Weiskrantz, 1997). Yet, we have also learned that something special happens when a piece of information becomes conscious. According to GNWT, this 'something special' is a non-linear ignition, leading to a global availability of information. There is a system of neurons that is able to select a piece of information, amplify it and broadcast it across the brain, and across modules. This broadcasting allows the sharing of information across otherwise nonconscious processing pathways, which enables the flexible recombination and routing of information, including the ability to send it to circuits for verbal report. Thus, reportability is just one marker for information being available in the workspace. Global availability is the key difference between conscious and unconscious processes.

According to GNWT, even detailed visual content, say a fine visual Vernier pattern, if it is within the focus of attention, should be accessed by GNW neurons and broadcasted to the rest of the brain. This is what allows us to report even on very fine perceptual details, if and only if we access them consciously.

Another important clarification should be made here; the workspace neurons are distributed. GNW is not a localist theory of consciousness. A common error is to equate it with claims about prefrontal areas. Much to the contrary, the GNW is postulated to rely on a robust, redundant, highly distributed system of neurons with long-distance axons, which incorporates both parietal and prefrontal cortices. During ignition by an external sensory stimulus, a subset of those neurons become activated (a subspace of the high-dimensional set of all possible GNW states), thus encoding the details of the conscious contents, while the other neurons, reflecting what the stimulus is not, are inhibited. In this manner, GNW areas amplify, hold online, and propagate the information originating for processors in sensory areas, and these regions in turn receive top-down signals from fronto-parietal neurons. Therefore, a key claim is that any conscious contents should be represented by a distributed assembly of neurons, present in both prefrontal cortex and these related areas. This prediction is well-supported by anatomical findings of long-distance connections interconnecting those regions and forming a central cortical core also supported by thalamo-cortical and basal ganglia

**Table 1**
Key commitments of each theory outlining their core focus, metaphysical commitments, and explanations of non-conscious processes and key neural mechanisms. It also highlights the role of attention, cognition, and evolutionary factors within each theory, as well as potential challenges. This comparison offers a structured insight into the strengths and weaknesses of different theoretical approaches to consciousness.

| Theory | Explanandum | What needs to be explained | Metaphysical commitment | Non-conscious information & associated mechanism | Key neural feature | Key computational feature | Evolutionary commitments | Role of attention/ cognition | Potential challenges |
|---|---|---|---|---|---|---|---|---|---|
| GNWT | Conscious processing (i.e., which processes require or do not require consciousness), and conscious phenomenology (i.e., what a participant experiences at a given moment) | Any reportable perceptual or non-perceptual experience (e.g., experience of knowing that one just made an error). Any delay or change in this experience (e.g., attentional blink or psychological refractory period). Any contrast between reportable versus non-reportable aspects of perceptual and cognitive processing | Functionalism | Dehaene and Naccache (2001) and Dehaene et al., (2006)(Box 2) introduced a taxonomy of non-conscious information in the brain, including (1) information not represented in cortical patterns (e.g. dormant synaptic connectivity); (2) information not represented by an explicit pattern of firing; (3) information encoded in regions disconnected from GNW areas; (4) activation evoked by weak sensory stimuli; (5) activation evoked by strong stimuli that failed to mobilize top-down attention and remain associated with feedforward processing or confined to local interactions within sensory regions, without triggering ignition | Global ignition defined as non-linear activation of neurons in Prefrontal-Parietal and Cingulate cortex, and broadcasting of those signals back to other distant areas | Flexible sharing of information across otherwise modular brain systems | The basic mechanisms of conscious access (non-linear ignition and global broadcasting) are the same in human and non-human species (e.g., crows); but some conscious contents (of a symbolic or compositional nature) may be uniquely human | Attention is necessary for consciousness, and phenomenology is sparse | The theory would be challenged if experiments could show that participants experience two incompatible conscious contents at the same moment |
| RPT | Conscious phenomenology (i.e., what a participant experiences at a given moment) | Sensation (vision in particular), perceptual organization and unification or integration, perceptual inference, perceptual richness, cognitive impenetrability | Functionalism (with structural implementation: the to-be-explained functions happen to be mediated by recurrent interactions in most brains) | Feedforward processing sweep, extracting low (shape, color, motion etc.) and high-level (faces, objects, scene gist) features, and activating sensorimotor reflexes up to | The distinction between Fast Feedforward Sweep (FFS) and Recurrent Processing (RP), the first unconscious, the latter yielding interactions between visual areas and hence conscious sensation. These interactions | Perceptual organization and integration, and inference by means of interactions between units representing individual features. Possibly perceptual learning | Basic mechanisms of conscious phenomenology could be shared with other animals (up to insects), but content depends on sensory and neural architecture | Sufficiency of recurrent processing for conscious phenomenology, independent of attention, access, cognition or response (first-order representation is sufficient) | Departure from introspection, cognition and report as the 'gold standard' evidence for having conscious experience |

(*continued on next page*)

**Table 1** (*continued*)

| Theory | Explanandum | What needs to be explained | Metaphysical commitment | Non-conscious information & associated mechanism | Key neural feature | Key computational feature | Evolutionary commitments | Role of attention/ cognition | Potential challenges |
|---|---|---|---|---|---|---|---|---|---|
| | | | | cognitive control functions | are for example expressed in feedback mediated modulations of early visual cortex (contextual modulation), effects of Transcranial Magnetic Stimulation (TMS) on early visual cortex, effects of masking, anesthesia, etc. showing a selective disruption of RP but not FFS. Possibly selective N-methyl-D-aspartate (NMDA) receptor activation by RP | | | | |
| HOTs | Conscious phenomenology (i.e., presence vs. absence of phenomenal experience) | Dissociations between phenomenology and behavioral performance; relationships between phenomenology and perceptual metacognition; inner awareness of mental function; reality monitoring; subjective inflation | Functionalism | Any first-order representation without an accompanying higher-order representation | Neural substrates of subpersonal (implicit) metacognition within parietal and prefrontal cortex; for joint determination variants of HOT, joint contributions of first-order (e.g., visual) and higher-order areas to consciousness | HOT variants differ according to their computational features. The Higher-Order State Space / Perceptual Reality Monitoring variants propose "lean" higher-order states, pointing to the precision or reliability of first-order representations, whereas the Higher-Order Representation Of a Representation variant posits "full-content" higher-order representations | HOT variants are agnostic regarding the status of consciousness in animals; lean HOTs are compatible with basic mechanisms of phenomenology being shared with all animals capable of perceptual metacognition / reality monitoring, with content dependent on sensory architecture | Requires cognitive machinery for higher-order monitoring or re-representation; may be lean / subpersonal / automatic | Translating psychological / philosophical distinctions between higher-order and first-order representations into neural / computational terms |
| IIT | What determines whether consciousness is present vs. absent; and what determines why specific experiences feel the way they do | First, what determines whether consciousness is present vs. absent? With respect to anatomy, why do certain part of the corticothalamic system seem to contribute directly to consciousness, while many other parts of the brain, such as the cerebellum and certain | The starting point of IIT is the existence of experience (0th axiom). This truth is not the result of an inference. It is immediate and irrefutable. The five axioms of IIT identify the essential properties of phenomenal existence—those that | The contents of experience are fully specified by the shape of the cause-effect structure specified by the substrate of consciousness (a main complex with its particular spatiotemporal grain) in a particular state. Every other | Dense, divergent-convergent hierarchical 3D lattice of specialized units in a state of causal 'readiness,' regardless of activity | None. An experience is a cause-effect structure, not a computation, a function, or a process (Findlay et al., 2024; Tononi et al., in press; Zaeemzadeh and Tononi, 2024) | IIT provides a principled explanation as to why consciousness and adaptation to a complex environment may evolve in parallel, though they do not have to ( Albantakis et al., 2014; Mayner et al., 2024) | Attention/cognition are not necessary for consciousness. In many instances cognition can be performed unconsciously by slave systems as a consequence of consciousness | Evidence in contrast with key tenets of the theory about the presence and contents of consciousness. For example, that the substrate of consciousness does not have the anatomical / physiological properties of a |

(*continued on next page*)

**Table 1** (*continued*)

| Theory | Explanandum | What needs to be explained | Metaphysical commitment | Non-conscious information & associated mechanism | Key neural feature | Key computational feature | Evolutionary commitments | Role of attention/ cognition | Potential challenges |
|---|---|---|---|---|---|---|---|---|---|
| | | cortical areas do not? With respect to physiology, why is it that consciousness vanishes during deep sleep even though neurons continue to be active, and during generalized tonic-clonic seizures, even though neurons fire maximally and in a highly synchronous manner? Second, what determines why specific experiences feel the way they do? Why do visual space and body space feel extended, why does time feel flowing, and why do objects, colors, sounds or touch, feel the way they do? | are true of every conceivable experience (every experience is intrinsic, specific, unitary, definite and structured). These essential phenomenal properties are then formulated in physical terms—in terms of a substrate we can observe and manipulate. For other assumptions, see Albantakis et al., (2023), and IIT WIKI (www.iit.wiki) | part of the brain (excluded from the main complex) is excluded and does not directly contribute to the contents of the experience specified by the system | | | | | maximum of cause-effect power, that it does not have a definite border and grain, that inactive neurons within the maximum of cause-effect power do not contribute to experience, that changes in connectivity with no or minimal changes in activity should modify experience, that spatial experiences are not supported by lattice-like substrates or temporal flow by directed grids, and so on (Tononi et al., in press) |
| PP | Conscious phenomenology (i.e., presence and absence of consciousness and the specific phenomenal character of a given experience) | Phenomenological character of subjective experiences of world and self, their function, and their presence vs. absence. More specifically, how expectations shape or give rise to conscious contents, the functional role of top-down connectivity, what are the sufficient conditions for conscious perception in terms of predictive inference | Materialism (subtypes may have other commitments or suggest more specific metaphysical commitments, such as biological naturalism, computational functionalism, or non-computational functionalism) | Open question; potentially predictive inferences without an active component, or without a top-down generative component | Neural activity corresponding to predictive inferences about the causes of sensory signals; potentially the active resolution of sensory prediction error through overt or covert action | Hierarchical minimization of precision weighted prediction error; more generally, active minimization of variational or expected free energy. Some versions of PP may not be computational per se, but may instead be dynamical theories (expressed in terms of differential equations etc.) | Varied; basic mechanisms shared with other animals; these mechanisms may have evolutionary roots in allostatic regulation of physiological essential variables | Attention is interpreted in PP as precision-weighting of prediction error, shaping conscious contents; cognition corresponds to higher-level predictive inferences. Under most versions of PP there is a smooth continuum from perception to cognition | (i) evidence that changes in conscious content are not accompanied by changes in Bayesian (approximate) posterior beliefs; (ii) lack of evidence for (or evidence against) PP (or active inference) as a core brain process, and / or that it is empirically poorly related to conscious contents. See Whyte et al. (2024) for more on what would refute a minimal active inference theory of consciousness |

loops (Markov et al., 2013). What should be common to all conscious-processing states is a processing style, or a type of neural trajectory within those areas, with non-linear ignition leading to metastable activity in these high-level areas lasting for about 100–200ms (King et al., 2014; Schurger et al., 2015).

GNW theory makes precise empirical predictions (Dehaene and Changeux, 1997; Dehaene et al., 2006; Dehaene and Naccache, 2001), which are also supported by simplified simulations of thalamo-cortical activity (Dehaene and Changeux, 2005; Dehaene et al., 1998; Dehaene et al., 2003; Klatzmann et al., 2023), many of which have already been borne out by the data.

First, the theory predicts that nonconscious processing can run very deep, along specialized and automatized processing chains, which may be subcortical as well as cortical (Fig. 1). According to the theory, knowledge encoded within the nervous system can remain nonconscious for a variety of reasons, and a taxonomy of nonconscious states has been proposed (Dehaene et al., 2006). The theory predicts that information necessarily remains nonconscious and cannot be consciously accessed (1) if it is merely encoded in latent form, by matrices of synaptic weights; (2) if it is not explicitly condensed in the firing of small specialized groups of neurons; (3) if those neurons are functionally disconnected from GNW neurons (e.g., those located in the brainstem); (4) if processing is confined to a brief, unstable traveling wave of firing, insufficient strong to cause ignition; or (5) if processing occurs while top-down attention is focused on another stimulus or task set. The theory stresses that information becomes conscious only when its mental representation receives top-down amplification and gets expanded into a global parietofrontal metastable state.

Second, there should be a sudden global and non-linear ignition, similar to a first-order phase transition, particularly but not exclusively in prefrontal cortex, whenever a novel content gains access to consciousness. This ignition should only be seen when the information is new and leads to a refresh of information within this global workspace.

Third, a crucial point for the theory is that such conscious ignition can be temporally decoupled from the external world: the onset and duration of conscious experience is not determined by the external stimulus, but by the availability of the GNW. Thus, on the one hand, we can be presented with an extremely brief stimulus, and still consciously experience it for an extended duration, if it gets amplified and stabilized by top-down signals – once the relevant neural assembly is ignited within the global workspace, the information can remain stable over time, maintaining it "in mind" for as long as needed. On the other hand, even a powerful external stimulus can remain nonconscious if the GNW is distracted elsewhere, thus leading either to complete invisibility (inattentional blindness; Mack and Rock, 1998) or to a transient delay in conscious perception (psychological refractory period; Marti et al., 2012). In a related phenomenon, "retro-cueing", a weak and otherwise invisible stimulus can become conscious, hundreds of milliseconds after its onset, if it is consciously cued (Sergent et al., 2013).

Fourth, given the architecture of the global workspace, we predict that one should be able to decode *any* conscious contents from the workspace neurons, based on the pattern of ignited GNW neurons, which involves a subset of active neurons while the rest are inhibited. In a nutshell, GNW activity holds a neural code for anything that we are conscious of. Notably, GNWT is not merely a theory of conscious perception, but a theory of consciousness in the broader sense. Whether you are conscious of seeing a face, of making an error (Charles et al., 2013) or of being sad, that information must be in the global workspace. Though our field has made substantial progress by focusing on bottom-up visual perception and visual illusions, many other paradigms are available to study other non-perceptual forms of conscious processing (e.g., automatized versus effortful cognitive tasks).

A fifth prediction of the theory refers to the notion of a central bottleneck. The GNW imposes a cognitive bottleneck because it serves as a central core which is shared by all conscious processing tasks, and which can only process once such task at a time. While the workspace is occupied by a specific content, other contents are blocked from entering the workspace, and hence, from accessing consciousness. Those contents may stay at the pre-conscious level, temporarily held in various peripheral buffers, but they are not yet conscious. This idea is supported by several experiments on dual tasking, attentional blink, inattentional blindness, and the psychological refractory period (Marti et al., 2010; Marti et al., 2012; Sergent et al., 2005; Sigman and Dehaene, 2008), where a piece of information has to wait a certain amount of time before

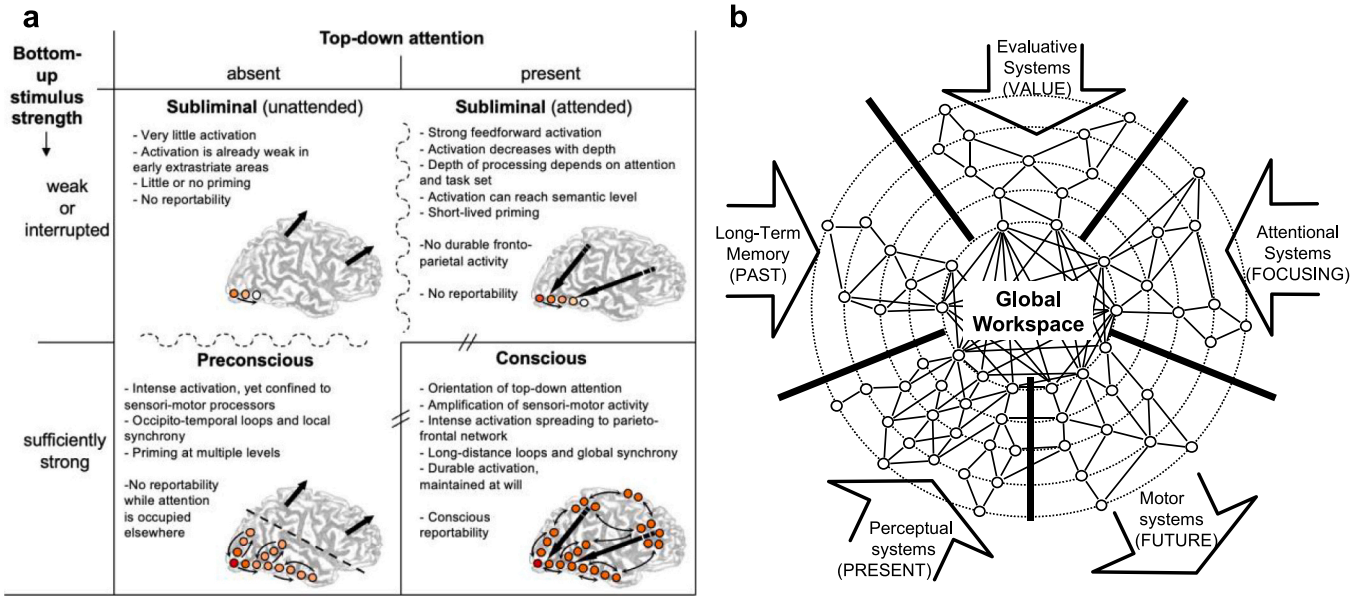



**Fig. 1.** (a) This classic figure adapted from Dehaene et al. (2006) shows the joint need for bottom-up stimulus strength (vertical axis) and top-down amplification (horizontal axis) in creating the conditions needed for access of a visual stimulus to the global neural workspace. While manu chains of processing can occur non-consciously, the theory stipulates that a distributed system of workspace neurons with long-distance axons must be activated for a sufficient duration (minimun 100-200 ms) for conscious appraisal to occur. As illustrated in the bottom right cell, what we experience as consciousness is the availability of information. (b) No longer confined to a narrow specialized circuit, information which is consciously accessed becomes broadcasted globally and flexibly to many other brain processors, thus allowing the information to be named, memorized, and acted upon in a deliberate, intentional manner (adapted from Dehaene et al., 1998).

becoming conscious. Notably however, the fact that we are always conscious of a single mental object does not imply that this object cannot comprise multiple features. Rather, the theory is compatible with a factorized neuronal representation in which distinct axes of neural space encode different dimensions of knowledge. For instance, in a single moment of conscious access, it is possible to become aware of a person speaking, including multimodal integration of their face, voice, and speech content.

Finally, a sometimes-overlooked property of the workspace model is spontaneous ignition. Conscious ignition does not require external stimulation, but can be also triggered from inside, for instance when accessing a memory or when performing a series of mental calculations. Indeed, simulations show that, even in the absence of external stimuli, the global workspace is constantly traversed by a series of ever-changing global patterns (spontaneous ignitions; Dehaene and Changeux, 2005). This mimics the fact that one can close the eyes and rest while still having a rich conscious flow of information (cf. William James' stream of consciousness; James, 1892). In fact, in patients with disorders of consciousness, the dynamics of spontaneous brain activity can be one of the clearest signatures that the patient is still having a conscious experience. We have made quite some progress in identifying such signatures and using them in the clinic or during anesthetic sedation, as means to detect consciousness (Barttfeld et al., 2015; Demertzi et al., 2019).

These predictions of GNWT have been supported by experimental observations. For example, Pieter Roelfsema and his team (Van Vugt et al., 2018) simultaneously recorded the activity of V1, V4 and prefrontal neurons in macaque monkeys presented with liminal stimuli – a dot, such that sometimes monkeys saw the dot and sometimes they did not. Areas in the visual cortex were strongly activated both when the monkeys reported seeing the dot and also when they reported not seeing it. This suggests that these areas cannot be the basis of conscious perception as they do not discriminate between conscious and unconscious content. Prefrontal cortex neurons, on the other hand, showed the expected non-linear ignition whenever the monkeys reported seeing the dot. This was true whether it was presented or not (i.e., false alarm). This lends initial evidence for the theory; yet, to provide much stronger support we should go beyond that and show that *any* conscious content has a neural assembly in prefrontal cortex that responds non-linearly to it when it is consciously perceived. This is one of our current research directions. Marie Bellet and other collaborators at Neurospin, including Fanis Panagiotaropoulos, Timo van Kerkoerle, Joaquin and Marie Bellet, and Marion Gay recorded from prefrontal neurons as monkeys were presented with a visual version of the local global test (Bekinschtein et al., 2009; Bellet et al., 2024). Here, we habituated monkeys to a particular type of sequence, such as three identical pictures that are followed by a different one (AAAB). Then, sometimes we presented rare sequences in which all images are identical (AAAA; so they violate the AAAB pattern). Contrary to some popular beliefs, we found that detailed visual information can be decoded from prefrontal cortex (see also Bellet et al., 2022; Panagiotaropoulos et al., 2012). Moreover, we found that whatever the monkey consciously knows, whether abstract or concrete, is encoded by a dimension of neuronal firing in prefrontal cortex. In this experiment, from Utah array recordings in prefrontal cortex (PFC), we could decode which image had been presented (identity), and also what its ordinal position was (number), whether it was different from the previous ones (local effect), which abstract sequence pattern it was included in (global pattern: AAAA or AAAB), and whether this pattern was occasionally violated (global effect). Thus, not only sensory contents, but also abstract knowledge was encoded in PFC firing. In other words, this experiment provides evidence for a key prediction of GNW theory, namely that whatever we consciously know is being encoded by the firing of a distributed population of neurons in prefrontal cortex. Interestingly, these features are encoded by directions in vector space that are almost entirely orthogonal to each other (see Tian et al., 2024; Xie et al., 2022 for relevant work suggesting that there are subspaces in prefrontal cortex for conscious contents). Future work should further test GNWT's prediction that the information encoded by PFC neuronal populations should be just as detailed as the phenomenal content of consciousness – if, for instance, the objective perceptual image is subjectively distorted by a visual illusion, then GNWT predicts that the subjective, rather than the objective information, should be decodable from PFC.

As a final point, I would like to focus on consciousness in humans and other animals. GNW theory proposes that the basic *mechanisms* of conscious access (non-linear ignition and global broadcasting) are the same in human and non-human species (for evidence of a threshold for ignition in crows, see Nieder et al., 2020). The basic operations of conscious access are similar: both have a global workspace and show very similar signatures of consciousness, including spontaneous activity, the presence of visual illusions, central collision etc. However, such parallels in the *mechanisms* of conscious access do not imply that the *contents* of consciousness are identical. There is growing evidence that monkeys may not be able to represent nested compositional structures, unlike humans. In the past five years, I have reoriented some of my research to ask what differentiates monkeys from humans. My hypothesis is that only humans have access to a language of thought: a compositional, syntactic ability which is manifested not only in natural language but also in mathematics or music (Dehaene et al., 2022). While both human and non-human primates share some neural mechanisms for representing sequences, the level of nested symbolic structures may be unique to humans (Dehaene et al., 2015). This would allow us to entertain more complex, symbolic, recursive conscious contents, including the representation of our thoughts and those of others (theory of mind).

### 2.2. *Recurrent processing theory / Victor Lamme*

Perhaps the best way to present recurrent processing theory (RPT) is to define it in contrast to other theories and explain where it surpasses these. As a starting point, I would like to note that RPT simply began as a set of empirical observations.

The first empirical observation pertains to the pattern of information processing when a visual stimulus is presented, which involves four stages of processing (Lamme, 2010; see also Fig. 2 below): until about 100 ms post stimulus presentation, feedforward feature extraction of different attributes takes place, like motion, orientation, color and so on (Lamme and Roelfsema, 2000). Notably, at ~200 ms, this feedforward sweep may then proceed up to frontal motor areas, where actions are being prepared, allowing deeper feedforward processing (Lamme, 2018). Third, around the same time, neurons in visual cortex start to engage in local recurrent interactions: higher level neurons that have been activated by feedforward processing send signals back to lower levels, and such repeated back-and-forth interactions enable low and high-level feature extraction to interact. Neurophysiologically, this is accompanied by phenomena like 'contextual modulation', where lower-level neurons modulate their activity depending on the global perceptual context of the features they encode. In this way, recurrent interactions enable functions like perceptual grouping, binding and perceptual organization (Lamme, 2020). Finally, we observe what within GNWT is called 'global ignition', where the whole brain gets engaged into widespread interactions, enabling working memory, recognition and so on (Dehaene et al., 2006).

The second empirical observation was that the feedforward sweep, no matter how deep it penetrates, is unconscious (Lamme, 2018): It is well-known that one can record both low- and high-level feature responses from anesthetized animals. Studies using manipulations rendering stimuli invisible (like masking) show that feedforward processing may unconsciously activate high-level visual areas (like the Fusiform Face Area (FFA); Fahrenfort et al., 2012), or even frontal regions, thereby evoking unconsciously triggered inhibitory control (van Gaal et al., 2008).

The third empirical observation, on the other hand, was that having

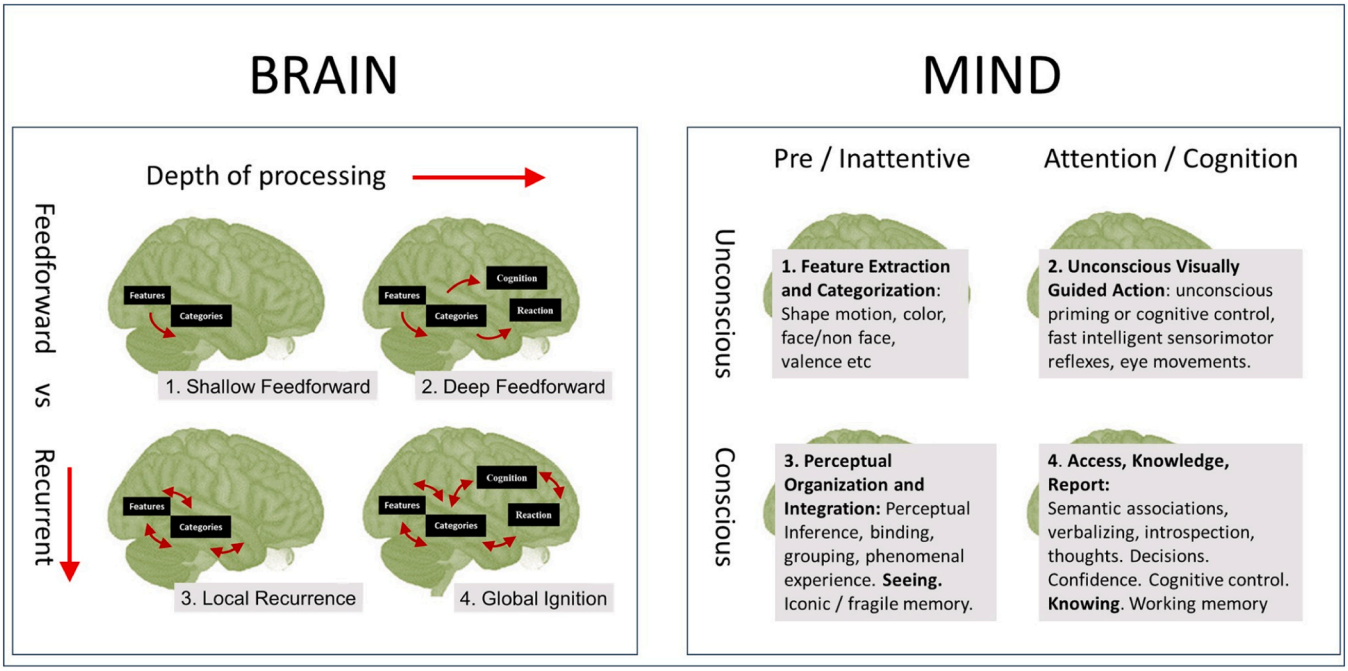



**Fig. 2.** Recurrent Processing Theory (RPT) identifies four stages of processing on orthogonal axes. The one axis pertains to the depth of cortical processing, or the extent information travels and is processed along the cortical hierarchy of visual areas and towards frontal regions involved in cognition, report and action. This depth is determined by attention; only (bottom-up or top-down) attended stimuli reach the whole depth. The second axis pertains to whether processing is strictly feedforward or involves recurrent interactions (mediated by horizontal connections within areas or feedback from higher to lower areas). The right panel lists the mental / functional equivalents of these four stages of cortical processing. Note how recurrent processing yields two varieties of conscious experience, the one summarized as 'seeing' (Phenomenal -conscious), the other as 'knowing' (Access-conscious), depending on the extent / depth of recurrent interactions. Also note how consciousness is orthogonal to and independent of attention or cognition, so that in RPT, consciousness gets a proper and fully independent ontology.

conscious sensations invariably goes along with observing recurrent processing: it is present in awake animals, but not anesthetized ones (Lamme et al., 1998), it marks the transition between visible versus invisible stimuli in masking paradigms (Fahrenfort et al., 2007; Lamme et al., 2001), as well as the difference between spontaneously seen or not seen stimuli (Super et al., 2001).

The fourth empirical observation was that as soon as recurrent interactions within visual cortex grow more widespread and include frontal regions, the transition towards attending the stimulus begins, storing it in working memory, or being able to report it (Scholte et al., 2006). One can say: "yeah, I saw that stimulus".

These observations – basically finding that recurrent interactions go along with conscious states and events, while feedforward processing does not - were the empirical foundations for RPT. The main theoretical point that RPT tries to make is very simple: the local recurrence might already be sufficient for us to have a conscious visual percept, to have a phenomenal experience (Lamme, 2018). So that implies that consciousness arises at stage 3 (Fig. 2), rather than at stage 4, where for example GNWT would position it. RPT states that global recurrence might actually only be needed for cognitive access: to be able to report of that conscious experience. In a similar way, there is no need for any higher-order re-representation of the content, as Higher-Order theories hold: first-order representation is sufficient for consciousness to take place (Fig. 2).

This view is very much in line with the empirical agenda of finding the 'true NCC', instead of its consequences, which include reports (Aru et al., 2012). Recently, there have been quite a few experiments suggesting that much of the observed frontal involvement, including the P3b component, are more tightly related to the report and task performance rather than to the phenomenal experience itself (Tsuchiya et al., 2015; Pitts et al., 2014). Notably, already in the early experiments we did with monkeys, I noticed that recurrent interactions in visual cortex, while related to stimulus saliency and visibility and conscious experience, can easily be dissociated from the report the monkey is making (usually via eye movements) by using different decision criteria (Super et al., 2001). In fact, this exact finding inspired me to posit that recurrent activation in itself is sufficient for having the phenomenal experience, which is different from reporting having that experience. It would imply that local recurrence is independent of attention, access and report (Lamme, 2020), and so may be a better candidate for being a 'true NCC' (Aru et al., 2012).

Yet one could present counter-arguments, mainly because people have different criteria for defining consciousness. Accordingly, GNWT proponents might say that what I refer to as phenomenal experience is in fact, pre-conscious because it precedes access, and because there is no cognition involved. HOT supporters might claim that these states are unconscious, because there is no higher-order thought that points at them. These deep disagreements about the definitions themselves might be one of the reasons for the current stalemate between theories, with 'back of the head' theories – RPT and IIT – claiming that visual activation itself is sufficient for consciousness, and the 'frontal theories' – GNWT and HOT (although there's more to global workspace than just frontal activations) – stating that frontal activations are necessary for consciousness to take place. In a way, this entire debate becomes almost a matter of taste, hinging on which criterion you use to define consciousness.

So, how can we find a way out of here? I believe a promising attitude is to focus on what needs to be explained about conscious vision. First, a very obvious and prominent hallmark of conscious visual perception, and all sorts of perception, is that experience is unified. Although there are about 30–40 different visual areas in the brain, each processing a different type of visual information, we have one visual percept: we don't see its aspects (e.g., shapes, colors or objects) in isolation, but as an integrated whole. Second, conscious vision is a type of inference. We usually go beyond the physical information of the stimulus itself to what we interpret based on that information. We go from luminance to gray

shade, from wavelength to color, from features to inference. Third, consciousness is generally cognitively impenetrable. Even if we gain cognitive information about a specific illusion, we cannot but see the illusory percept. Fourth, consciousness involves integration. When we see a face, there are neural mechanisms (e.g., face neurons, or the FFA) that detect the difference between faces and other stimuli, seeing a face is way more than that: we integrate this 'faceness property' with the shape of the face, its color, its structure and all sorts of other properties it has. Thus, seeing involves a large degree of integration. To me, this is a key feature that we must explain about consciousness. In a way, this is the explanandum.

This point is nicely illustrated by a study conducted by Johannes Fahrenfort (Fahrenfort et al., 2012), showing that the FFA is activated by invisible faces, and that only when the FFA interacts with lower visual areas, you get a conscious percept of that face. I claim that local recurrence itself is fully capable of explaining this unity of consciousness, or this integration, because such local recurrence mediates those functions: perceptual inference, incremental Gestalt grouping, long-range and complex organizations, figure-ground organization etc. (Lamme, 2014). I really don't see how any frontal or higher order theories have any explanatory power towards explaining this aspect of consciousness - the aspect of unity, integration and binding. The same goes for cognitive impenetrability: I argue that frontal theories would predict that it should be entirely possible for knowledge to affect perception (e.g., to overcome perceptual illusions). If all types of processing get integrated into one coherent whole by this global ignition, why wouldn't you be able to also integrate your thoughts into the perception? I believe that frontal theories would also predict dependence of perceptual functions on attention, but that is not the case: there is ample evidence that all these above-mentioned functions happen whether you attend to a stimulus or not. Attention may modulate or strengthen these functions, but they do not critically depend on it (Lamme, 2020). This has been demonstrated by yet another nice experiment of Johannes Fahrenfort (Fahrenfort et al., 2017), where the effects of attentional blink (AB) and masking on processing Kanizsa figures were compared. He decoded the contrast (high vs. low) of Kanizsa inducers from visual cortex activity regardless of AB or masking, showing contrast is a feature typically processed by feedforward processing (Fahrenfort et al., 2017). Conversely, the 'integration contrast' (i.e., whether the Kanizsa configuration yielded the typical illusory percept) was disrupted by masking (known to interfere with visibility and recurrent interactions in visual cortex), yet not by the attentional blink.

Another feature of visual perception is that it is phenomenally very rich. This poses a problem for frontal theories, which are so tightly linked to attention, having strong capacity limits. Thus, these theories should in fact predict that visual perception is fairly limited and sparse, having an attentional bottleneck capacity. Now, one can claim, for example, that change blindness experiments show exactly that: if you show an array of objects twice, and in the second presentation you change one of the items, it is indeed very difficult to notice the difference between the two. Yet we have shown in many experiments that if you cue these arrays between the change, participants clearly recognize the item that was replaced (Landman et al., 2003).

This suggests that their representation of the first image is in fact very detailed, precise and rich (Sligte et al., 2010). These iconic and fragile memory experiments have clearly shown that in change blindness, the first scene is overwritten by the second scene as soon as it is presented. Accordingly, change blindness is a failure of memory, rather than of perception (Lamme, 2010).

As a bonus argument for RPT, I will take a metaphysical standpoint: I think that RPT is ideally suited to give consciousness an independent ontology. The four stages of processing described above can be put in a $2 \times 2$ scheme where attention or cognition can be seen as one axis, basically depending on how deep information gets processed in the brain, whether frontal regions are involved etc. The other axis differentiates between unconscious and conscious processing, which is an entirely different and orthogonal axis, and in the brain corresponds to the difference between feedforward and recurrent processing. This scheme readily explains how shallow versus deep feedforward processing makes it possible to have unconscious priming depend on attention (Naccache et al., 2002), or why strictly feedforward frontal activation may yield unconscious cognitive control (van Gaal and Lamme, 2012). Similarly, it explains the difference between phenomenal and access consciousness (Block, 1995), indexing 'shallow' (visual) versus 'deep' (global ignition) recurrent processing. Most importantly, it shows how in RPT consciousness is thought to be independent and orthogonal to other cognitive functions such as attention or cognition, functions that are often conflated with consciousness in frontal theories like GNWT or HOT. One could say that RPT is a theory that takes consciousness seriously, and gives it its own and independent ontological status.

### 2.3. *Higher Order theories / Stephen Fleming*

Higher Order theories (HOTs) are already well-established in philosophy (Carruthers, 2001), but are somewhat more of a new kid on the block when it comes to neuroscience. As previous work shows (Yaron et al., 2022), there is less empirical evidence for or against these theories compared to the others. But I am still excited by HOTs, because I think they offer us a way of charting an empirically productive path between global and local theories. I will try to demonstrate that using the guiding questions of this debate.

First, what is the explanandum? We are trying to explain the presence versus absence of phenomenal experience: for example, if our visual systems are processing information about a sunset over London, we are usually aware of, and able to communicate, our experience of the sunset to others, while remaining unaware of other perceptual inputs that are nevertheless being processed, such as the feeling of clothes on our skin, changes in our posture and so on. HOTs seek to explain the difference between these two types of information processing.

Second, what do HOTs claim? In a nutshell, HOTs say that a perceptual representation of some content X – for instance, a red apple – is not sufficient for a conscious experience of X to arise. These kinds of representations are known as "first-order" as they are directed at the world. Having a first-order representation of the object is often crucial for guiding behavior – for instance, allowing us to pick up and eat the apple. However, according to HOTs, such first-order representations can also occur nonconsciously and are insufficient for phenomenally conscious experiences. Instead, HOTs hold that the phenomenal consciousness of X depends on an organism being in some way aware of being in state X. This, in turn, entails that the first-order state is in some way monitored or meta-represented by a higher order representation (Brown et al., 2019; Lau and Rosenthal, 2011; Fig. 3). This higher-order representation can take many forms, as described below. HOTs account for unconscious influences on behavior by positing that first-order states can still drive task performance unconsciously. In particular, those first-order states might be quite widely broadcast and facilitate unconscious control – which is where HOTs start to diverge from GNW.

Importantly, there are multiple versions of HOT, just like there are multiple versions of first-order theories. We are actively engaged in an adversarial collaboration to test distinct empirical predictions made by the different HOTs, led by myself and Axel Cleeremans and supported by the Templeton World Charity Foundation. This project is seeking to test two key axes of disagreement between different HOTs (Fig. 3b). One key difference is whether a higher-order theory allows the first-order states to be doing any of the work involved in creating conscious experience. On the one hand, proponents of "sparse" views suggest that consciousness arises when first-order and higher-order states work together (also referred to as joint determination; Fleming, 2020; Lau, 2019). According to these views, the content of experience is carried by the first-order states, with the higher-order states having a more subtle (but nevertheless critical) role in monitoring the precision, intensity or reliability

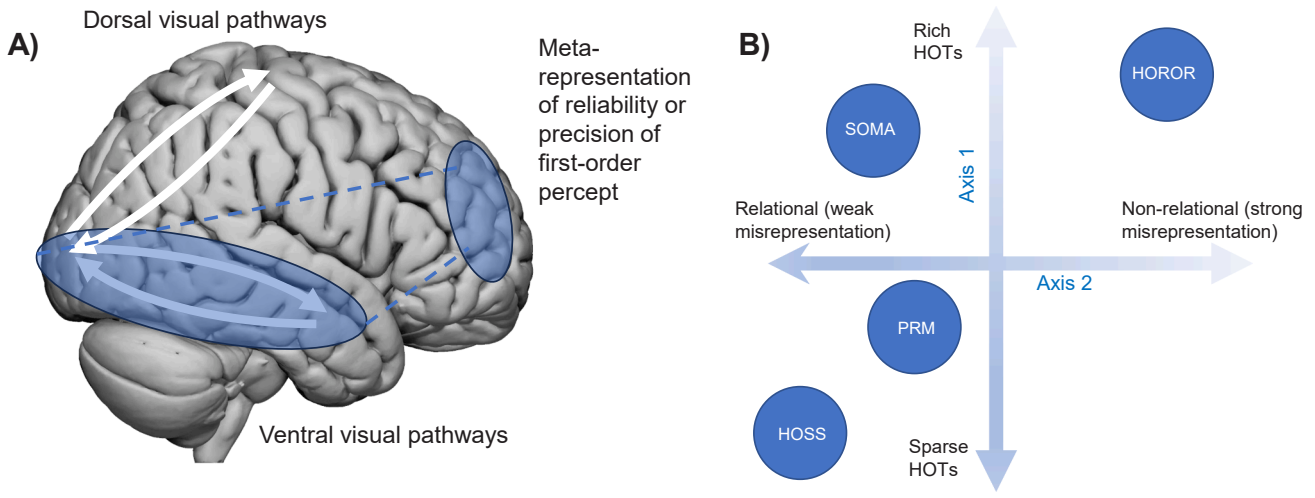



**Fig. 3.** **a.** Illustration of hypothetical neural substrates supporting lean higher-order accounts such as HOSS. Prefrontal areas support implicit meta-representations of the reliability, or confidence in, first-order percepts built along ventral visual pathways. The meta-representation is hypothesized to track the reliability of an entire first-order representation, which itself may be supported by hierarchical processing along ventral visual pathways. Under HO theories, global broadcast and action guidance may still proceed unconsciously, for instance via the dorsal visual pathways. **b.** Two axes of disagreement between higher-order theory variants: HOSS, Higher-Order State Space account; PRM, Perceptual Reality Monitoring account; SOMA, Self-Organizing Metarepresentational Account; HOROR, Higher-Order Representation of a Representation account. A first axis indicates to what degree meta-representations are rich or sparse. Sparse HO theories propose that phenomenal experience is jointly determined by first-order (FO) and HO states, with HO states playing a lean role in tracking the precision, intensity or reliability of FO states. Rich HO theories propose that phenomenal experience is fully determined by higher-order states, such that higher-order representations are just as rich and detailed as perceptual experience itself. A second axis of disagreement is on whether HO representations can "misrepresent" their FO targets, and in what way. The position of theory variants along these axes is often correlated, with the rich non-relational theories permitting stronger misrepresentation.

of the first-order state. These types of higher-order representation are sometimes referred to as "pointers" or "indexes". At the other end of the spectrum, we have "rich" HOTs that predict that the phenomenal experience is fully determined by the higher-order states, which are held to be as rich as perception itself (Brown, 2015; Rosenthal, 2005). Irrespective of the question of perception being sparse or rich, these variants hold that the higher-order states should fully determine the sparse or rich content of conscious perception. Another axis of disagreement concerns whether different variants of HOT allow for the possibility of misrepresentation of first-order states – cases in which the higher-order and first-order representations disagree. Under relational views (which are also often rich; Fig. 3b) what matters for consciousness is the content of the higher-order, not first-order states – and so consciousness should change in tandem with the former and not the latter. There are also intermediate views, in which the first-order states contribute to conscious experience, but the higher-order states are more than mere pointers or indexes, and also furnish independent content (Cleeremans et al., 2020; LeDoux, 2020a; Van Gulick, 2004).

HOTs also hold that it is particularly important to control experimentally for performance when adjudicating between different theories of consciousness. Plants, cameras and thermometers all have some sensitivity (or, in signal detection terminology, some level of d') to aspects of the environment, yet we typically do not think they are conscious of what they are sensitive to. Similarly, in the human brain, there may be mechanisms that create a form of behavioral sensitivity to the stimulus, without any accompanying conscious experience. This creates an empirical problem in human experiments, because if first-order representations can both drive performance and contribute to consciousness, then if we alter consciousness by also changing performance or signal strength, we will not know whether to ascribe the effects we find to performance or to consciousness. This is a pervasive problem for the field, but it is particularly problematic for testing predictions of HOTs, because if our experiments do not control for performance, we might unfairly stack the deck in favor of first-order theories by misattributing a change in consciousness to the accompanying change in first-order states that may in fact be (unconsciously) driving performance (see discussion in Lau, 2022).

HOTs are also often linked to metacognition – the ability to reflect on and monitor other aspects of cognition and perception (Fleming, 2024). Metacognition is often investigated by asking people to judge confidence (or error) in aspects of behavioral and cognitive performance – which in perceptual tasks is referred to as "perceptual confidence". Because in studies of metacognition, people are usually explicitly asked to reflect on and judge their performance, these experiments are tapping into "explicit" forms of metacognition. This can often be a source of misunderstanding. Higher-order theorists do not typically hold that explicit metacognition is the same thing as phenomenal consciousness. Instead, some variants of HOT claim that there are theoretical reasons to think that the computations that are important for instantiating higher-order states, such as monitoring the precision of first-order representations, share mechanisms with perceptual confidence (Fleming, 2020; Lau, 2019). This is because both may depend on a subpersonal implicit form of metacognition – aspects of metacognition which are not necessarily available for subjective report.

Empirically, studying perceptual metacognition is useful for two reasons: first, it allows us to study the neural basis of perceptual confidence, and second, it allows us to tackle the performance issue, by using statistics such as metacognitive efficiency which isolate metacognitive capacities controlling for first-order task performance (Maniscalco and Lau, 2012). By analyzing metacognitive efficiency in experiments probing conscious experience, we can identify situations where performance (d') is matched, but there is an isolated change in confidence. In the past few years, a growing body of work has implicated prefrontal cortex, and particularly anterior prefrontal sub-regions, in perceptual metacognition. In our lab, we find early implicit signatures of perceptual confidence tracked in the activity of the medial prefrontal cortex, particularly the perigenual anterior cingulate cortex, followed by a later involvement of the lateral frontopolar cortex (Bang et al., 2020; Bang and Fleming, 2018). This division of labor between medial and lateral subregions of prefrontal cortex in perceptual metacognition has also been shown by Marios Philiastides' group using EEG-informed fMRI (Gherman and Philiastides, 2018), and in studies of the role of intra-prefrontal connectivity in supporting metacognitive efficiency (De Martino et al., 2013). One interpretation of these data is that the

coupling supports a more explicit form of metacognition, so that confidence estimates become available for communication and control (Fleming, 2024). Other studies have found that when activity within subregions of the prefrontal cortex is manipulated using Transcranial Magnetic Stimulation (Shekhar and Rahnev, 2018) or decoded neurofeedback (Cortese et al., 2020) it is possible to modulate perceptual confidence without changing performance. These data support the notion that prefrontal and parietal regions are important for perceptual metacognition.

A key implication of these findings is that there is a psychologically and neurally meaningful distinction between implicit and explicit metacognition: perceptual confidence can be formed automatically, independent of the ability to communicate metacognitive estimates to others (Fleming, 2024; Shea et al., 2014). Studying the role of implicit confidence estimation in perceptual experience will be helpful for adjudicating between the different variants of HOT. But even though HOTs traditionally emphasize prefrontal cortex as supporting higher-order states, there is increasing recognition that the critical distinction here is computational rather than anatomical. As Joseph LeDoux has pointed out, mapping a philosophical division between first-order and higher-order states onto the functional anatomy of the human brain is necessarily going to be much more complicated than two interacting neural populations (LeDoux, 2020b). Thus, the 'front versus back of the brain debate' (e.g., Boly et al., 2017; Odegaard et al., 2017) should be viewed as more of a starting point than endpoint for adjudicating between the different theories.

In that vein, I want to briefly highlight a research program that we have been pursuing, developing a minimal computational implementation of a sparse HOT within a predictive coding framework. We call this the higher-order state space (HOSS) model (Fleming, 2020). We start with a first-order generative model that performs inference on perceptual inputs, such as whether a visual stimulus is an apple or an orange. This process of building first-order representations of the world may proceed unconsciously, as in Helmholtz's famous phrase of "unconscious inference" (Gregory, 1970; Helmholtz, 1867/1962) – we perceive the product of perceptual inference, rather than the machinery of inference itself. Within HOSS, additional higher-order layers monitor the precision of these first-order representations to allow the system to track whether the first-order representations reflect external reality, or just noise (similar architectures underpin the perceptual reality monitoring variant of HOT; Lau, 2019). The operation of these higher-order representations underpins conscious experience by endowing the agent with beliefs about the assertoric force and reality of its perceptual representations – beliefs which have what philosophers refer to as "assertoric force" (Lau, 2022).

These different types of computation make broad qualitative predictions about the nature of the neural representations at these different levels: higher-order states should be lower dimensional and symmetrically encode the high vs. low precision of the first-order states, whereas first-order states should be higher dimensional and encode first-order content (note that a first-order state may be supported by an anatomical network that spans sensory and association cortices, rather than a single node in early sensory areas). We have also been exploring how classical neural signatures associated with conscious awareness such as global ignition can be reinterpreted within this model architecture: specifically, we see that ignition-like signatures emerge as a natural consequence of an asymmetry in the first-order state prediction errors. When the representations within a first-order generative model are precise (on trials associated with conscious awareness of stimulus contents), strong prediction errors from the stimulus propagate through the system (Whyte et al., 2022). Thus, predictive processing versions of HOT may be able to recapitulate the neural signatures previously highlighted as supporting GNWT. However, HOSS interprets these ignition-like signatures differently: rather than reflecting global broadcast, under HOSS, conscious trials are associated with greater (average) prediction error. This is due to an asymmetry in the magnitude of prediction errors associated with "aware" and "unaware" trials in an experiment. This interplay between predictions and prediction errors is consistent with the broader class of predictive processing (PP) models discussed earlier, but now focused on explaining differences that ensue between conscious and unconscious perception.

In work together with Nadine Dijkstra, Peter Kok and colleagues at UCL we have started to test some of these predictions of the model using neuroimaging. We trained participants to generate predictions not only about the content of the stimulus (first-order) but also whether or not they will be aware of that content (higher-order). For instance, on some trials, you might be cued to expect that you will not see anything (an expectation of stimulus absence), but that if you did see something, the stimulus would be a face (as opposed to a house). By confirming and violating these content and awareness expectations with face and house stimuli embedded in noise, simulations show that we can nicely orthogonalize the prediction errors at the two levels (first-order and higher-order) of our computational model. We find that behaviorally, these two types of prediction error both affect response times, slowing down identification of the stimulus (Dijkstra et al., 2024), and modulate subjective reports and confidence in stimulus identity (Haarsma et al., 2024). We then combined a no-report version of this experimental design with fMRI: participants were simply presented with the cues and stimuli without requiring overt responses. In the brain, we find that while first-order state prediction errors are tracked in patterns of activity in ventral visual cortical areas, higher-order state prediction errors are tracked by activity patterns in medial PFC (Dijkstra et al., 2023). Taken together, these findings support the dissociation in prediction errors posited by the HOSS model.

Finally, I would like to end by highlighting common ground between some of the theories under consideration, which have more in common than is typically assumed: HOTs and GNWT (see section by Dehaene above) both claim that consciousness depends on interactions between first-order states and higher-order computational processes, and under the predictive coding implementations of HOT that we have been pursuing, ignition and broadcast signatures reflect asymmetries in global prediction error as a function of precision. HOTs and RPT (see section by Lamme above) agree on some things too: recurrent message passing may be needed for higher-order and first-order states to interact, so this approach is especially compatible with the joint determination variants of HOT. There is obvious common ground with PP theories (see section by Seth below): as we have seen, it is possible to implement versions of HOT within generative model architectures. Under PP there seems to be no single criterion by which conscious and unconscious contents are differentiated; HOSS, however, provides a clear proposal (see above). Lastly, with respect to IIT (see section by Boly below): in a recent review (Lau et al., 2022) we suggested that the organization and structure of the first-order states themselves – the sensory code, the fact that it's smooth and sparse – may enable higher-order representations to support the relational comparisons that might underlie the ability to communicate what it is like to have an experience: the capacity to implicitly represent that scarlet is more similar to crimson than it is to blue, for instance. To me, this is what a functional explanation of the "what it is likeness" of conscious experience may look like, and it has some commonalities with the idea that the organization and cause-effect structure of the first-order states themselves are important for phenomenology, as proposed by IIT (although under HOT, this organization alone would not be sufficient for conscious experience). Thus, different elements within HOTs share common ground with other theories of consciousness.

### 2.4. *Integrated Information Theory / Melanie Boly*

A theory of consciousness must answer two key questions. First, what determines whether consciousness is present versus absent; and second, what determines why specific experiences feel the way they do. I will start with the first question. With respect to anatomy, why do certain parts of the corticothalamic system seem to contribute directly to

consciousness, while many other parts of the brain, such as the cerebellum and certain cortical areas, do not (Tononi et al., 2016)? With respect to physiology, why is it that consciousness vanishes during deep sleep even though neurons continue to be active, and during generalized tonic-clonic seizures, even though neurons fire maximally and in a highly synchronous manner? (Juan et al., 2023).

The second question a theory of consciousness should answer is what determines why specific experiences feel the way they do. This includes explaining why visual space or body space feel extended, why time feels flowing, and why objects, colors, sounds or touch, feel the way they do.

IIT does not ask how the physical world 'gives rise' to experience—it does not try to 'squeeze' consciousness out of the gray matter of the brain. Instead, IIT starts from consciousness itself—from phenomenology—which is the starting point for everything, including science. IIT's approach is to characterize the essential properties of consciousness—those that are immediately and irrefutably true of every conceivable experience—and formulate a principled, coherent account of those properties based on the causal powers of a physical substrate (Fig. 4; see also Albantakis et al., 2023; Tononi et al., 2016).

What do we mean by physical? IIT defines the physical in purely operational terms, as cause-effect power—the ability of a substrate's units to take and make a difference. In line with the traditional scientific method, it employs systematic manipulations and observations of a physical substrate to obtain a Transition Probability Matrix (TPM) that summarizes the substrate's powers in terms of causes and effects: if we do X, what is the probability that we reliably observe Y (Albantakis et al., 2023).

IIT identifies five properties that hold true for every conceivable experience (Albantakis et al., 2023). These are intrinsicality (experience exists for itself), information (it is specific), integration (it is irreducible), exclusion (it is definite), and composition (it is structured). The last property (composition) refers to the fact that every experience is composed of phenomenal distinctions – such as faces, the left and right corner of space and so on – and phenomenal relations that bind them together in various ways – for example, the face is in the left corner.

This phenomenal structure composed of distinctions and relations now needs to be accounted for in physical terms. IIT does so by developing a mathematical framework that allows us to 'unfold' in full the cause-effect power of a substrate, leading to a cause-effect structure (also called Φ structure) composed of causal distinctions (cause-effects) and relations (overlaps among causes and/or effects). According to IIT, this is the physical structure that can account for all the properties of consciousness, and for the quality of a specific experience, here and now.

Within this mathematical framework, which we have refined over the years, we can aim to achieve an explanatory identity such that all of the properties of experience can be expressed explicitly in terms of a specific Φ-structure expressing a substrate's cause-effect power (Albantakis et al., 2023). In this way, we can generate predictions about which physical substrates can support consciousness (and which cannot), as well as about the quality of experience (the composition of the Φ-structure) and its quantity (measured by Φ). Importantly, given a substrate in its current state, the associated Φ-structure should account for the properties of specific experiences with no additional ingredients.

What does the theory explain empirically? With respect to anatomy, IIT explains why certain substrates, and not others, can account for the essential properties of consciousness. For example, most of posterior-central cerebral cortex is organized roughly like a hierarchy of 2D grids of neurons (i.e., pyramid of grids; Maruoka et al., 2017). According to IIT, this kind of substrate is very well suited for supporting rich cause-effect structures of high Φ. In contrast, the modular architecture of the cerebellum (D'Angelo and Casali, 2013) necessarily fragments into many small substrates, each supporting disjoint cause-effect structures of minimal Φ. With respect to physiology, IIT can explain why the same anatomical substrate – the cerebral cortex – can support large Φ-structures during wakefulness, but disintegrates into many small structures during dreamless non-REM sleep, when changes in neuromodulation lead to the breakdown of cortical effective connectivity. Indeed, after direct electrical stimulation of the cortex, intracranial recordings show complex, recurrent interactions during wakefulness, which are interrupted by the stereotypical occurrence of an off-state during deep non-REM sleep (Pigorini et al., 2015).

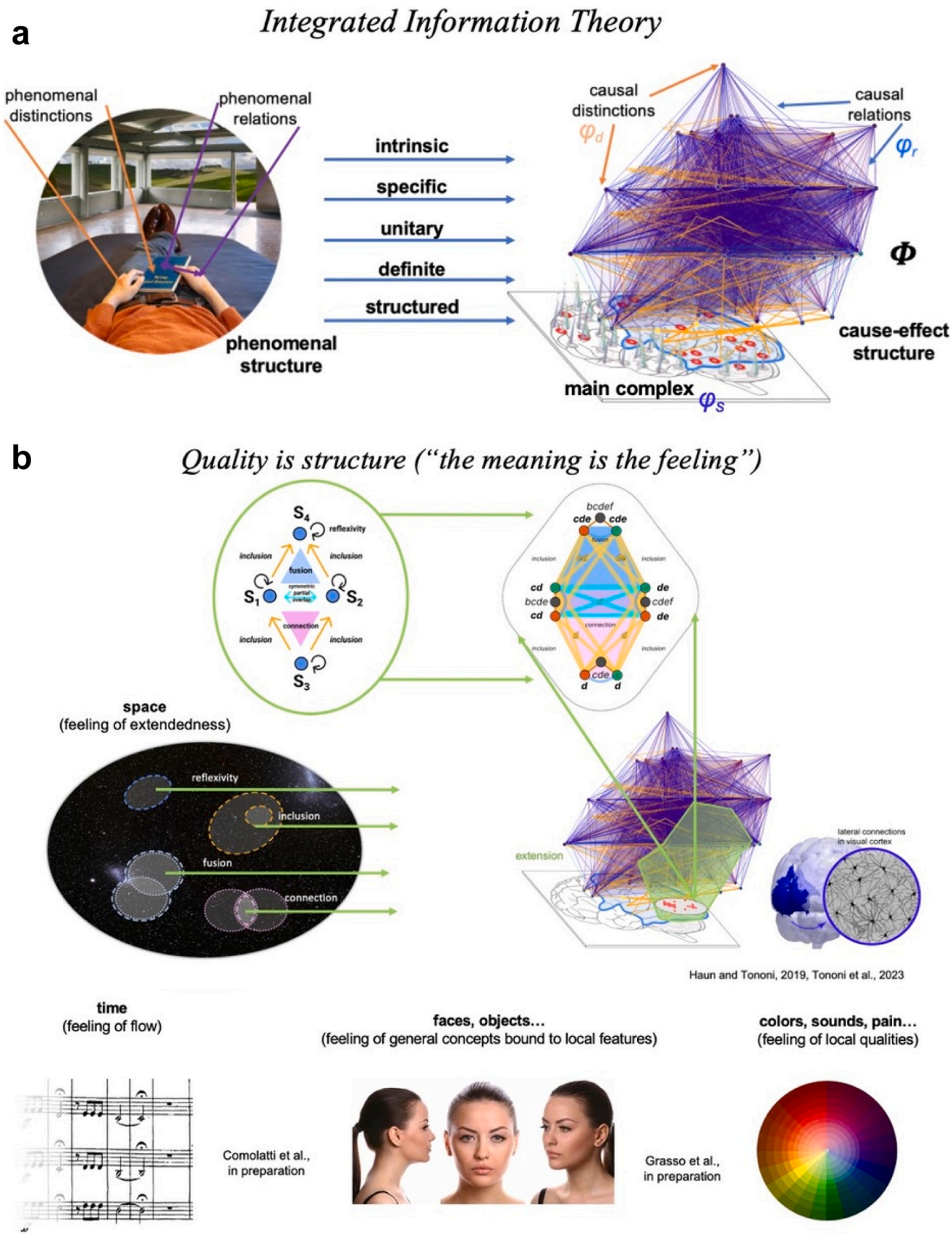
IIT's principles have also inspired a Transcranial Magnetic Stimulation - Electroencephalography (TMS-EEG) method for assessing the presence or absence of consciousness through a Perturbational Complexity Index (PCI): in wakefulness, the EEG response to TMS is complex (high PCI), revealing the induction of differentiated activity patterns, while in deep non-REM sleep the response becomes local, short, and stereotyped (Massimini et al., 2005; Tononi et al., in press). Notably, PCI is high in states where subjects are conscious but unresponsive, including REM sleep or ketamine anesthesia (Casarotto et al., 2016).

Going back to the question of what determines why specific experiences feel the way they do, we started approaching this question with spatial extendedness, because the experience of space is both pervasive and partially penetrable, in the sense that we can use introspection—especially spatial attention—to dissect its phenomenal structure, as opposed to color or pain (Haun and Tononi, 2019). We then showed that the kind of Φ-structures specified by 2D grids can account for how space feels. Briefly, the fundamental property of spatial experience is extendedness: the visual field, for example, is composed of phenomenal distinctions ('spots') that overlap according to a distinctive set of phenomenal relations (reflexivity, inclusion, connection, and fusion). It turns out that the Φ-structures specified by 2D grids, such as those found in much of posterior-central cortex (Wang et al., 2015), are composed of causal distinctions and relations of exactly the same kind. Given that experience is pervasively spatial (in both the visual and somatosensory domains), it is no surprise that both in humans and in non-human animals, large parts of the posterior-central cortex indeed constitute of pyramids of grid structures. Ongoing work (including an adversarial collaboration, https://osf.io/4rn85) is testing some of the predictions that follow from IIT's account of spatial experience and its neural substrate. In the meantime, we are actively pursuing a research program that aims at accounting for the feeling of time flowing (Comolatti et al., 2024), and of objects binding general concepts with particular features (for more details, see IIT WIKI).

While IIT focuses on the contents of an experience triggered by a stimulus or occurring during a dream, it is obviously meant to deal with both how we perceive the external world and how we learn about it (for an in-depth discussion of how IIT formalism can reconceptualize perception as the triggering of intrinsic meanings/feelings by stimuli, see Mayner et al., 2024). Through the notion of matching, IIT can also explain the extent to which intrinsic feelings/meanings capture regularities due to causal processes in the environment, which are internalized in the connectivity of the brain. Finally, there are some simple demonstrations using animats (small simulated organisms evolving in a simple environment) showing that learning about a complex environment is typically associated with an increase in integrated information (Albantakis et al., 2014), thus providing an indication as to why selective pressure might favor the evolution of greater and greater consciousness. Accordingly, IIT reconceptualizes perception as interpretation, rather than as information processing or representation.

### 2.5. *Predictive processing / Anil Seth*

Predictive processing (PP) theories occupy a unique region in the space of theories of consciousness. This is mainly because they are, in general, not (yet) 'theories of consciousness' in the sense of proposing necessary and sufficient conditions for conscious experience (or specific conscious contents) to occur. Instead, they are more general theories of brain and mind that can be applied to account for various properties of consciousness – such as the specific phenomenological properties associated with different kinds of conscious experience. As Howhy & Seth




(caption on next page)

**Fig. 4. a.** The starting point of IIT are the essential properties of consciousness – those that are true of every experience. In short, every experience is intrinsic, specific, unitary, definite, and structured by phenomenal distinctions bound by relations (left side). IIT then formulates these properties in physical terms, understood operationally as cause-effect power, or the ability to take and make a difference. Accordingly, the substrate of consciousness, or main complex, must have cause-effect power that is intrinsic, specific, irreducible, definite, and structured (right side). The cause-effect structure fully unfolds the complex's cause-effect by considering causal distinctions, i.e., the cause-effect of each of its subsets, and causal relations, i.e., the overlaps among cause-effects. The content of experience should be accounted for in full by the complex's cause-effect structure, with no additional ingredients. **b.** Accordingly, for IIT all quality is structure. Moreover, the meaning of contents of experience is fully intrinsic, being specified by the cause-effect structure ("the meaning is the feeling"). IIT's research program aims to account for the quality of different kinds of experiences. As a first example, an analysis of spatial experiences indicates that space feels "extended" because phenomenal distinctions (called "spots") are bound by relations that capture reflexivity, inclusion, connection, and fusion. An analysis of cause-effect structures unfolded from substrates connected in a grid-like manner, as is the case in much of posterior cortex, can account for phenomenal extendedness in physical terms. Ongoing work aims at accounting for why time feels flowing, why faces and objects feel like general concepts bound to local features, and why colors and sounds feel the way they do. IIT's explanatory identity claims that a complex's cause-effect structure should account for all properties of an experience, essential and accidental, with no additional ingredients. Note that the explanatory identity is between the quality (structure) of an experience and the cause-effect structure (composed of causal distinctions and relations) specified by the complex in its current state; not with the complex's connectivity or activity patterns. The cause-effect structure of the complex is obtained by unfolding the cause-effect power of all its subsets in their current state, following IIT's composition postulate.

(2020) put it, PP is not a theory *of* consciousness, it is – or least started out as – a theory *for* consciousness science.

According to PP, the brain is continually minimizing sensory "prediction error" signals, either by updating its predictions about the causes of sensory signals or by performing actions to bring about predicted or desired sensory inputs (the latter process being called "active inference"; Clark, 2013; Friston et al., 2010). This ongoing process of prediction error minimization provides a mechanism by which the view of perception as a process of Bayesian inference (Helmholtz, 1867/1962) can be implemented. In this view, the aim of perception is to infer the most likely causes of sensory signals (the Bayesian posterior, or 'best guess'), given some 'prior' belief or expectation about these causes, and the new information provided by the sensory data (the Bayesian likelihood). Importantly, this process can also provide a means of predictive regulation of physiological variables, in which interoceptive perceptual priors can serve as set-points or target ranges for homeostasis and allostasis (Barrett and Simmons, 2015; Seth, 2015).

In its most ambitious and all-encompassing version, the "free energy principle" (FEP), the mechanism of prediction error minimization arises out of fundamental constraints regarding control and regulation. These constraints apply to all physical systems that maintain their organization over time in the face of external perturbations, with living systems being a particularly expressive example (Friston et al., 2010). In this view, prediction error emerges as a proxy for sensory entropy, which organisms are mandated to minimize, through active inference, in order to remain in the statistically expected states that are compatible with their survival. While the FEP, being a principle, is not itself testable, 'process theories' that fall within the FEP – such as PP – are testable.

Generally, in PP, predictions are proposed to flow in a top-down (or 'inside-out') manner, with prediction error signals flowing in a bottom-up (or 'outside-in') direction (Fig. 5). This challenges the usual ascription of the labels 'feedback' and 'feedforward' to top-down and bottom-up connections respectively, since 'feedback' usually connotes transmission of an error signal, whereas in PP these signals are typically associated with bottom-up or 'feedforward' connectivity.

This functional architecture for PP is not fixed. Tscshantz et al. (2023) recently described a 'hybrid predictive coding' architecture in which predictions and prediction errors flow in both directions, but at different time scales, implementing a flexible balance between learned (amortized) mappings from sensory data to Bayesian posteriors and standard (iterative) inference. Under the FEP, recent extensions have illustrated how minimization of 'expected free energy' (i.e., approximate future prediction error under some action policy) can optimally balance a trade-off between exploratory (epistemic) and goal-directed actions (Friston et al., 2015; Tschantz et al., 2020). I mention these extensions to underline the richness of the computational resource provided by the still-evolving PP (and, more broadly, active inference) framework. This resource is further enriched by the hierarchical nature of PP, and – in particular – by the key role of *precision-weighting*, in which sensory prediction error signals with high (estimated) precision have

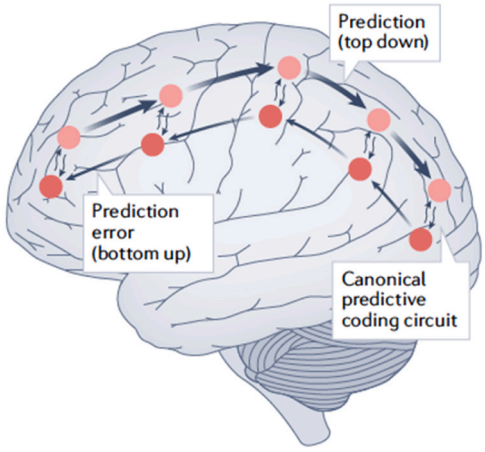



**Fig. 5.** Predictive processing (and related theories like active inference; Friston et al., 2010) are not *prima facie* theories of consciousness (Hohwy and Seth, 2020). They are more general theories of perception, cognition, and action that can be used to derive predictions relating to aspects of consciousness, perhaps leading to some minimal theory of consciousness *per se* (Whyte et al., 2024). The core claim of predictive processing is that conscious mental states are associated with top-down signaling (thick arrows) that, for predictive processing, convey predictions about causes of sensory signals (thin arrows signify bottom-up prediction errors), so that continuous minimization of prediction errors implements an approximation to Bayesian inference. Conscious contents are specified — in most predictive processing theories — by the content of top-down predictions. Adapted from Seth and Bayne (2022).

greater influence on updating predictions; a mechanism that has been argued to map onto the role of attention in perception (Feldman and Friston, 2010).

The broad strategy here is to leverage the rich resources of PP to develop systematic mappings between properties of conscious perception and properties of the underlying neural architectures – where by 'systematic' I mean being systematically guided by theoretical considerations of some sort (Hohwy and Seth, 2020). In virtue of this systematicity, such mappings should also have explanatory and predictive power that can be experimentally tested. The larger ambition is that, through iterating this process, and by examining shared features across diverse individual phenomena, a core set of theoretical commitments will emerge. This set will constitute a PP theory of consciousness *as such*, rather than merely being a theory 'for' consciousness science (Whyte et al., 2024).

Space limitations preclude a detailed review of all the ways this iterative refinement of PP is underway (see Hohwy and Seth, 2020; Seth and Bayne, 2022; Whyte et al., 2022). I will highlight just a few, to illustrate both breadth and depth. An emerging theme in the more recent models is their emphasis on *action* as crucial in shaping conscious

contents and transitions between contents, where these actions can be either overt (i.e., expressed via the body) or covert (i.e., mental actions, such as paying attention).

Perhaps the most well-developed PP studies of conscious perception focus on visual experience. An early model tackled binocular rivalry by proposing the existence of two competing perceptual hypotheses, one of which 'wins', leading to perceptual dominance (Hohwy et al., 2008). Sensory signals from the alternative hypothesis accumulate as prediction error, which eventually leads to a perceptual transition, at which the source of unexplained prediction error switches and the cycle repeats. Parr et al. (2019) extended this model, using expected free energy, to propose that the accumulation of uncertainty makes the suppressed sensory signals progressively more epistemically attractive, until attention (considered as a covert action) shifts to these signals, which makes the corresponding perceptual belief sufficiently precise that it dominates perception. Notably, this model accounts for experimentally observed features of binocular rivalry which are hard to account for in alternative 'passive' models: these include the slowing of rivalry in the absence of attention, the modulation of dominance durations by reward, and the regularities relating stimulus features such as luminance contrast with dominance duration (Whyte et al., 2022).

Versions of this model have also been adapted to account for Troxler fading, where peripheral visual content fades from experience (Parr et al., 2019), and to bistable perception of Necker cube stimuli (Novicky et al., 2024), both again emphasizing the idea of policy selection to minimize expected free energy, with perceptual switches being primarily driven by epistemic actions.

The PP approach also has the potential to integrate experimental findings which may challenge some existing theories. In one example using computational modeling, the researchers developed a PP model of conscious access in which the gating of working memory is treated – similarly to attention – as a covert cognitive action (Whyte et al., 2022). Applying this model to a simulated visual masking task, they showed how late P3b-like event-related potentials (ERPs) and increased PFC activity could be induced by the working memory demands of report generation. But when the model is modified to simulate a no-report condition, these (simulated) late ERPs and PFC activities diminish or go away entirely - an empirical observation that has often been taken to challenge theories which emphasize anterior processing (e.g., GNWT and HOT). However, the model shows that even without reporting demands, simulated PFC activity still reaches the threshold for reportability on some trials – maintaining the link between PFC activity and conscious access that is central to theories like GWT and HOT, and speaking to empirical finding from animal studies in which conscious contents can be decoded from PFC in no-report conditions (Kapoor et al., 2022).

The core idea implicit in these examples – and perhaps the kernel of a PP theory of consciousness *as such* – is that perceptual content is given by the brain's 'best guess' of the causes of its sensorium – this 'best guess' being the approximately optimal Bayesian posterior. The experiential character of this content is specified by the nature of the perceptual predictions in play. On this proto-theory, any change in conscious content must result from a change in the inferred state of the body, brain, and/or world (Hohwy and Seth, 2020; Seth, 2021; Whyte et al., 2022). (The reverse may not be true: changes in inferred state need not always lead to changes in conscious content.)

One appealing aspect of PP is its potential to go beyond highly constrained psychophysical environments to shed light on broader aspects of perceptual phenomenology that may be harder to account for by other theories. A good example is provided by perceptual (e.g., visual) hallucinations. The idea that hallucinations may result from overly strong perceptual priors dates back at least to Fletcher and Frith (2009) and was given computational precision by Suzuki et al. (2017) who used a neural network modeling approach to simulated the altered phenomenology that results from an imbalance between top-down predictions and bottom-up prediction errors. More recently, Suzuki et al., (2023) extended this approach to model the specific phenomenological characteristics distinguishing different kinds of visual hallucination – those arising from neurological disorders, visual loss, and psychedelics. By paying attention to more detailed aspects of phenomenology – beyond the simple detection and discrimination judgements widespread in most experimental settings – the PP approach can fulfill its promise to build explanatory bridges between (neural) mechanism and conscious phenomenology, with important clinical implications including but not limited to hallucinations.

Moving on, PP models have been extensively applied to aspects of conscious *self* – perhaps in contrast to other theories of consciousness (see Metzinger, 2004 for an early and influential attempt). The same strategy applies: the relevant conscious content is associated with an inferred state of the brain, body, or world – now with an emphasis on the body. An early example – albeit one that remains difficult to experimentally test – is the proposal that emotional experience corresponds to inference about the physiological condition of the body; a process of 'interoceptive inference' (Barrett and Simmons, 2015; Seth, 2013). In this view, the broad phenomenology of emotional experience – being dominated by valence – can be linked to a role for interoceptive inference in allostasis. This in turn connects closely to the FEP, according to which the entire edifice of perceptual inference can be derived from a fundamental imperative to continue existing – to stay alive. As I put it, "we perceive the world around us, and ourselves within it, with, through and because of our living bodies" (Seth, 2021). This perspective raises the intriguing but as yet untestable notion that consciousness may be a property of only (but not necessarily all) living systems – a form of biological naturalism (Searle, 2017; Seth, 2021; Seth, 2024b).

Grand claims aside, some progress has been made in computational models of specific aspects of self-related conscious experience. These range from early work by Stephan et al. (2016) linking disorders of allostasis to fatigue and depression to recent models of meta-awareness central to the flow of experience in contemplative states (Sandved-Smith et al., 2021). Rigorous testing of these and other models remains a challenge. As does extending PP to account for changes in conscious level, as in sleep and anesthesia. Here, it would be natural to appeal to the integrity of core mechanisms of prediction error minimization (Boly et al., 2011).

Whether PP succeeds as a theory of (or for) consciousness will depend both on evidence that prediction error minimization is indeed a core brain operation, and on its ability to draw explanatory and predictive links between elements of PP and properties of experience. While substantial evidence links top-down signaling to conscious perception (Hardstone et al., 2021), evidence for explicit sensory prediction error signals remains mixed (Solomon et al., 2021). And, while abundant evidence shows that participant expectations can shape conscious perception (de Lange et al., 2018) much remains to be done to causally connect the computational entities and dynamical processes of PP with specific forms of consciousness.

## 3. Where do we go from here: some insights about ToCs from the open discussion / Liad Mudrik & Lucia Melloni

Following the four talks (as mentioned, PP was unfortunately not presented at the meeting), an open discussion ensued. Below, we offer a selective overview of the key points of agreement and disagreements, open questions, and challenges to the theories that emerged from the debate. Overall, the discussion revealed some of the controversies and disagreements between the theories, while yielding few points of consensus.

### 3.1. What are we trying to explain?

One surprising point of agreement among the theories, considering previous discussions in the literature (Evers et al., 2024; Northoff and Lamme, 2020; Seth and Bayne, 2022), was their shared objective of

explaining the phenomenology of the experience. That is, all theory proponents defined the explanandum of their theory as including both what distinguishes the presence vs. absence of an experience, and the phenomenal character of those experiences – why a certain experience feels the way it does, and how experiences differ. In that respect, an important clarification in the discussion was that both for GNWTand HOT, reports are not the explanandum itself, but rather a medium to capture the explanandum, which is phenomenology. Dehaene also urged the community to broaden its typically narrow focus on bottom-up visual perception and visual illusions, to other types of conscious experiences (e.g., feelings of knowing, awareness of making an error, etc.). It became apparent, however, that not all theories, in their present form, have provided explanations, or theory-based empirical investigations, of the phenomenal character of an experience. This holds true for HOT (though for a recent attempt, see Lau et al., 2022; Fleming and Shea, 2024), PP, but also to some extent to RPT and GNWT, where most empirical work has focused on the difference between consciously and unconsciously processed information.

The theories diverge however with respect to what that phenomenology is (e.g., detailed and rich but fleeting and non-captured by report, vs. sparse, low dimensional and fully captured by report) and what counts as phenomenological data (e.g., if an observation goes against folk-psychology, does it still count as data?).

Similarly, the theories strive to explain both *state consciousness* – what makes a person, or a system, conscious of its environment and its 'self' within that environment, as opposed to not being conscious – and *content consciousness* – what makes a person (or a system) conscious of a specific content at a specific time. This commitment to explaining the multiple aspects of consciousness is important, yet again, it does not fully align with how theories of consciousness have been studied thus far. As previously shown (Yaron et al., 2022), empirical support for the theories is highly unbalanced with respect to studies focusing on exploring state vs. content consciousness. Based on the updated ConTraSt database, which at present includes 503 experiments, but is mute with respect to PP due to lack of records (https://contrastdb.tau.ac.il), RPT has been almost exclusively studied with respect to content consciousness (97 %), and a similar bias also exists to some degree for GNWT 72 % of the experiments and HOT (71%), while IIT has been mostly studied with respect to state consciousness (81 % of experiments). Thus, for the theories to live up to their proclaimed explanandum, more diverse empirical research work is needed, focusing on the understudied domains for each theory. Similarly, for some theories, more theoretical work is needed, as acknowledged for both PP and HOTs above (notably, some claim that under specific frameworks of consciousness, the same explanation can account for both state and content consciousness; Aru et al., 2019; Bachmann and Hudetz, 2014. Yet these claims still require further empirical investigation).

### 3.2. What is consciousness?

Some of the differences between the theories appear to boil down to their different definitions of consciousness, which sometimes depart from common intuitions – or folk-psychological beliefs. For example, RPT explains away cases of inattentional and change blindness by claiming that participants actually did consciously see the unattended stimuli. That is, participants arguably had the conscious experience, although they don't know about it or don't remember it and so are unable to report it.

As Lamme himself acknowledged in the discussion, persuading the community that this counterintuitive interpretation is correct is a major challenge for the theory. Doing so depends on non-empirical arguments: if one finds the explanatory power of RPT high, they should also accept some of its non-intuitive claims, including this one. To him, for the field to stop going in circles, theories and their associated findings, particularly in neuroscience, must be able to trump some of our first-person perspective, folk psychological intuitions (Lamme, 2010), akin to the way researchers accepted non-intuitive claims in other fields, like quantum physics, when compelled to do so by empirical findings.

Yet this was far from being a consensual claim: Fleming argued that while quantum physics might reveal a counterintuitive truth about reality, a counterintuitive claim in consciousness science could take us away from the very same phenomenon we are trying to explain, which in turn might not even count as an explanation of the phenomenon. Thus, an open question is to what extent should theories of consciousness comply with folk-psychological concepts of consciousness. More specifically, the possibility of participants consciously experiencing unattended stimuli despite reporting otherwise was contested by Dehaene and Fleming, in line with their theories, while Boly argued that one might be conscious of some low-level features of the unattended stimulus, without necessarily experiencing an associated category.

Another example of the lack of agreement between the different theoretical stands relates to what happens when we are in a 'flow' state (Csikszentmihalyi et al., 2005) – for instance, reading 'Crime and Punishment', being immersed in challenging math exercises or even driving. What do we actually experience in these cases? And what differential predictions do the theories make about them (e.g., should there be a higher-order state for the experience of the environment while driving)? In the debate, we again failed to reach an agreement around these points, and it was suggested that there is no empirical way to solve some of these disagreements (e.g., about inaccessible experiences; Block, 2011).

Importantly, this discussion showcased that theories do not even agree on which states are conscious and which ones are not: for the very same experimental manipulation – inattentional blindness – RPT considers the information fully consciously perceived, IIT considers that some low-level features may be perceived while the category is not, while GNWT and HOT consider the information fully unconsciously processed (due to the lack of attention or higher-order representation, respectively). The divide went even further, as theorists were not able to agree on the criterion for detecting a conscious state: while for GNWT, some sort of reportability is required, for RPT the mere presence of perceptual organization suffices, and for PP the content needs to be encompassed in an inference about the state of the brain, body, or world. And, critically, we were unable to agree on a justification for either of these measures.

As Lamme put it during the debate, the current state of affairs seems more akin to differences in taste and preferences, where the very same data is held by one theory to reflect phenomenal experience, while another theory considers it to reflect unconscious processing. The lack of agreement between theories might render some of the claims almost tautological, such that they are always true, yet only within the framework of the theory. It is evident that this stalemate should be resolved for the field to move forward and to converge on a satisfactory explanation of consciousness. Yet no clear proposals were offered for how such resolution could be obtained.

On the one hand, this lack of agreement might be detrimental to the field: If the theories cannot even agree on the definition of what counts as a conscious event, it seems almost impossible to directly compare and test them. For that, the theories should be at least somewhat commensurable, both at the logical and at the empirical level (Evers et al., 2024). On the other hand, this might actually be a way to arbitrate between the theories: if one could present a conceptual, or an empirical, argument showing that inattentionally blind stimuli are either consciously or unconsciously processed, or that perceptual organization can (or cannot) take place unconsciously, this could serve as an argument against one or more theories.

More broadly, as a field, it might be worthwhile to explore consensus-establishing methodologies (e.g., Delphi studies; Barrett and Heale, 2020) aimed at agreeing on the definitive properties of consciousness and the associated data that could demonstrate their existence. Another constructive strategy might be to press theories of consciousness to provide clearer and more objective standards for what

will count as evidence for or against their preferred definition. An alternative possibility is to accept that our understanding of the concept of consciousness is still within the prescientific stage. Accordingly, consciousness should be treated for now as a multidimensional entity, with different theories explaining different aspects of it. Then, more experimental and theoretical efforts will hopefully yield a more precise conceptualization. Under this approach, it would be advisable to systematically explore the parameter space of possible conscious experiences, and accept a more pluralistic view of this phenomenon (He, 2023). Such efforts might lead to the refinement and improvement of existing theories (Lakatos, 1978), or to a new account that would explain where existing ones are wrong. In parallel, bottom-up studies, collecting more evidence regarding the neural correlates of consciousness in a theory-neutral manner, could also be constructive in the attempt to establish a better, potentially novel, explanation for consciousness.

### 3.3. Are ToCs even theories?

According to Dehaene, the entire debate would not have taken place had ToCs been real theories, in the full-fledged sense of the term, as none of the current theories has achieved a level of theoretical development enabling precise formulations, leading to clear and explicit testable predictions that fully explain all aspects of consciousness. Dehaene suggested that, at present, perhaps a more modest term such as "framework" or "hypothesis" should be used – also for GNWT. According to him, the proliferation of theories stems from the partiality of each of them. What is lacking, he argued, is a precise formalization of the theories, leading to explicit simulations (as attempted for GNWT) that would allow us to test the theories in a much more rigorous manner. Consciousness research should accordingly move from descriptions to mechanisms, taking a much more mechanistic view. This could take place in two possible axes: at the anatomical level, we should strive for a more fine-grained description of the brain (e.g., at the synaptic or dendritic level; for a recent example, see Aru et al., 2020; Phillips et al., 2024). At the functional level, for theories that assume some form of computationalism, we should provide more detailed and accurate descriptions of the types of computations that are related to consciousness, specifying the relevant grain of the functional description, so that if they are implemented, consciousness should arise (e.g., Dehaene et al., 2017). If computationalism is not assumed, then theories should also be explicit about the non-computational functions associated with consciousness (Godfrey-Smith, 2016; Piccinini, 2020; Seth, 2024a).

Fleming considered the proliferation of theories and models a positive aspect of the field, rather than a disadvantage – according to him, it is a testament to the development of these accounts and their translatability into testable models, much like in other fields (e.g., working memory). Yet he too acknowledged the limited nature of our theories; for HOTs, for example, the focus has mostly been so far on explaining presences vs. absence of stimuli, rather than accounting for the specific phenomenal aspects of experience (the same argument can be made towards GNWT and most theories of consciousness; Seth and Bayne, 2022). More recently, HOTs have been trying to develop that aspect of the theory as well, introducing the notion of a quality space that aims to model the relational properties of conscious experience (Lau et al., 2022; Rosenthal, 2010). Seth noted, retrospectively, that the (current) status of PP as a theory for consciousness science rather than a full-fledged theory of consciousness is an advantage here, mitigating against the temptation to overclaim for a theory while offering a continuity with an understanding of how brains and minds work in general.

More generally, this discussion highlighted the lack of critical criteria for *what a theory must explain, to be counted as a theory of consciousness*; what does it have to provide to be regarded as a theory, as opposed to a description or a hypothesis (Doerig et al., 2020; Kuhn, 2024; Schurger and Graziano, 2022; Seth and Bayne, 2022)? In the same vein, it remains unclear what its explanandum/explanada should be. That is, should a theory of consciousness be required to explain *all aspects of consciousness* (to name a few: states of consciousness, contents of consciousness, the maintenance of a percept over time, the phenomenality of consciousness, its functions and its relations with other mental states)? Here, an interesting conflict arose between Lamme and Dehaene. The former claimed that GNWT or HOT do not explain perceptual organization, and hence cannot explain unity and integration, which he considers a key property of conscious experience. Conversely, Dehaene insisted, following Baars (1997), that this should not be an essential part of a theory of consciousness, which instead should only explain *conscious* perceptual organization (that is, it should not explain all aspects of the brain and its functions, but only the specific mechanisms that single out conscious vs. unconscious processing). Similarly, one can ask if a theory of consciousness must explain *all types* of conscious experiences, or whether it can be confined to one type, or one modality. And, perhaps most importantly, what level of prediction must it be able to generate, to be counted as a meaningful theory?

This question has been especially directed at IIT, in view of the claim that some of its aspects may be untestable (Barrett and Mediano, 2019; Doerig et al., 2019; IIT Concerned et al., 2025). For example, it has been claimed that Phi (Φ) is not computable for large systems like the brain. To that, Boly replied in saying that for smaller systems, Φ is computable exactly (Albantakis et al., 2023), and for larger systems one can resort to approximations (she referred to a work in progress using fMRI focused at the voxel level - see Tononi et al., in press, though see Mediano et al., 2022, for a criticial discussion of approximations vs. proxies of IIT). She also referred to the PCI method as crudely capturing a proxy of Φ (though we note here that the results of PCI experiments are also compatible with other theories of consciousness, like GNWT; but see discussion in Tononi et al., in press). Finally, Boly argued that IIT makes many other testable predictions, about the factors leading to a loss of consciousness (Tononi et al., in press), or the substrate of specific kinds of experiences (such as space; https://osf.io/4rn85, and time; Comolatti et al., 2024), including counterintuitive predictions about the neural substrate of states of pure presence (Boly et al., 2024), and so on.

Seth later added that PP again offers an interesting contrast and alternative, providing a route towards a full theory of consciousness, rather than – as is presently the case – a set of resources for explaining properties of consciousness in terms of neural mechanisms and dynamics. One possible scenario for PP theories of consciousness is that the so-called hard problem of consciousness will not be directly solved (e.g., by a consensus agreement that process X generates or is identical to consciousness) but dissolved, as explanatory and predictive bridges between mechanism and phenomenology are built, tested, and refined (Searle, 2007; Seth, 2021).

### 3.4. What can refute the theories?

Perhaps the hardest question raised in this debate was "what would make you change your mind". Dehaene identified a central assumption of GNWT, according to which there should be only one central state of consciousness at a given moment. With this assumption, the theory tries to overcome a challenge facing local theories of consciousness – the problem of integration. Dehaene contrasted that with RPT, which focuses on the visual system but does not account for what happens within the other sensory cortices, or for non-sensory conscious contents (e.g. the sudden consciousness of having made an error). Arguably, RPT allows each sensory cortex to have its own feedback loops, potentially creating parallel conscious experiences for each modality. For GNWT, conscious experience must be integrated and central, as opposed to parallel. If one attends to audition, one cannot perceive a distinct competing visual stimulus at the same time, unless those two sensory inputs are integrated into a single unified percept (e.g. the McGurk illusion; McGurk and MacDonald, 1976). This is a key prediction of the theory, and if it is found to be wrong, the theory will be substantially challenged. According to Dehaene, this is a feasible test: experiments might show that GNWT underestimates the amount of parallel

processing in the conscious brain, demonstrating that such processing can give rise to a conscious experience, as RPT claims. Or they could show that there is more than one central state or one central sharing system in the brain. All these outcomes would be highly informative for GNWT. We note that in this context, studies on agenesis of the corpus callosum (Paul et al., 2007) could be illuminating. Unlike the more classical studies on split brain patients (Gazzaniga, 1967), here the dissociation between the hemispheres is present from birth, providing a remarkable opportunity to investigate the possibility of multiple, concurrent, conscious experiences (using methods that do not rely on verbal report, given hemispheric disconnection). However, it still remains to be seen to what extent this pathological situation sheds light on the unity of consciousness in the neurologically intact brain.

Fleming maintained that we should not expect to find one critical piece of evidence that would refute HOT (or any other theory), and instead we should appeal to a plurality of measures and tests. According to him, there are many possible outcomes that would make him change his mind. For example, if it were possible to somehow inactivate or remove the neural substrate of the relevant higher-order representations, and participants would still claim awareness of the content that the higher-order state is pointing to, that would require a 'reformulation' of HOT, as he put it. Of course, this is currently not a technically feasible experiment, but according to Fleming, this might be achievable in the (near?) future. However, the real challenge with this proposal is that the theory currently does not specify an area that could be inactivated, as Fleming explained; since the higher-order state is a network property, it is less likely to be subserved by a specific region in PFC. This casts doubts on the feasibility of the proposed experiment, beyond the technical challenges. However, Fleming did mention that there might be a confidence 'code' that is expressed in localized activity patterns (e.g., Cortese et al., 2016; Masset et al., 2020) which one could then discover and potentially knock out. To the extent that subpersonal metacognition supports the relevant higher-order representations (e.g., Lau, 2019), this could provide a strong test of (a variant of) HOT, but so far this is still more of a theoretical conjecture than an empirical reality.

Boly specified several types of evidence that would challenge IIT: first, if one computed an approximation of Φ (based on the IIT 4.0 framework) and showed it is higher during generalized-clonic seizures, when brain activity increases but you lose consciousness, that would be a serious challenge for IIT. Similarly, if the approximations of Φ were found to be higher at the grain of molecules rather than neurons or microcolumns, or at the grain of microsecond or tens of seconds rather than the temporal grain of experience, that would seriously challenge the theory (Tononi et al., in press). The theory could also be indirectly challenged by results obtained through cruder proxies of integrated information, such as the perturbational complexity of EEG responses to TMS (PCI). PCI is meant to capture three of the five postulates of IIT: intrinsic existence (causation within the corticothalamic system), information (as estimated by differentiation), and integration (owing to the deterministic spread of perturbations within the corticothalamic system; Tononi et al., in press). Yet, as we further claim below, the PCI measure does not capture the full set of axioms of IIT and it is also compatible with other theories. Thus, in isolation it cannot be taken as either supporting or fully challenging IIT (but see again the discussion in Tononi et al., in press). Another line of investigation that could potentially refute IIT, according to Boly, focuses on the quality of consciousness: if the feeling of extendedness that characterizes spatial experiences is not subserved by grid-like networks in the cerebral cortex, that would be problematic for IIT (https://osf.io/4rn85).

Lamme reminded the discussants that he already changed his mind, as his original claim was that feedback must revert all the way back to primary visual cortex for consciousness to occur, something that he no longer holds as a necessary condition for consciousness. But the critical evidence that would make the theory false is twofold. First, finding recurrent interactions while the person is not conscious of the information (after excluding possible processes interfering with access or report to the conscious sensation, such as lack of attention, working memory etc.). To some extent, this is an almost trivial challenge, as it can be safely assumed that even in deeply unconscious states, some recurrent interactions between some neurons will remain. And also in subsystems that are generally considered not taking part in conscious experience, such as the spinal cord, recurrent interactions between neurons are present. Therefore, obviously, RPT will need some adjustments to better specify what the precise nature and extent of recurrent interactions have to be for conscious experience to arise. This has been referred to as the 'missing ingredient problem' in earlier publications (e.g., Lamme, 2018, where potential solutions have also been proposed). Second would be the finding of conscious experiences without recurrent interactions. If purely feedforward processes can support conscious experience, that would be a big challenge to the theory. This could be perhaps done using drugs or optogenetics, trying to selectively block feedback connections (Kirchberger et al., 2021). A similar challenge would arise if the signals that are typically considered markers of recurrent processing (such as delayed 'contextual modulation' signals, neuronal interaction measurements, etc.) are in fact due to feedforward processing. For example, delayed modulation of responses could be due to slow ascending subcortical arousal systems. Any mechanism where parallel streams operate at different speeds (such as magno- and parvocellular LGN inputs, for example) could mimic effects of feedback. It is therefore important to notice that the primary empirical observations on which RPT is based used contextual modulation signals for which a feedback origin has been confirmed using lesion experiments in monkeys (Lamme et al., 1998; Super and Lamme, 2007).

Finally, Seth argues that the utility of PP as a theory for, or eventually of, consciousness will turn on whether the core process of prediction error minimization can be experimentally verified or refuted. This is not trivial to do: the very flexibility of PP that endows it with the resources to address diverse aspects of conscious phenomenology also means that it is hard to identify any single (or set of) experimental tests that would be up to the job. For example, one might think that evidence against PP would be provided by experiments showing that conscious contents are constituted or deeply shaped by bottom-up signals, rather than by top-down signals. But the recent 'hybrid' extension of PP undermines this simple hypothesis, licensing instead a different set of predictions linking different types of phenomenology (e.g. focal vs gist) to different types (and directions) of predictive signaling (Tscshantz et al., 2023). Having said this, it should eventually be experimentally tractable to determine whether prediction error minimization is a core neural process. But this by itself is not enough. PP could also fail as a relevant theory if its resources – even if based on a sound empirical footing – do not, in fact, provide a flow of explanatory and predictive links between mechanism and phenomenology. In the terminology of Imre Lakatos (1978), PP will need to justify itself as *progressive* rather than *degenerate*, as a theory for, or of, consciousness.

Altogether, the proponents of these theories have offered potential experiments that could be carried out to refute their theories, given appropriate technical conditions. However, leaving technical feasibility aside for a moment, the major question is how these experiments are connected to the central ideas of the theories. The fact that the theories are still being developed and modified, and with the brain being the dumbfoundingly complex system that it is, theory proponents enjoy a high degree of freedom when interpreting data and making theoretical claims. For all theories, the bridging principles between the theoretical/computational claims (i.e., higher order thoughts, global broadcasting, integrated information, recurrent processing and predictive processing) and their biological realization are not fully specified. Thus, facing new data, the theories could be modified to accommodate these new findings, even when these do not confirm their predictions.

For GNWT, the functional neuroanatomy and physiology of the global neuronal workspace are still not fully clear and need to be further specified. How global and distributed should brain activity be to be considered as forming a conscious cell assembly within the global

workspace? How many neurons should be activated and deactivated, and across how many areas? Which layers and cell types should be privileged? Since multiple superimposed active cell assemblies can co-exist in orthogonal prefrontal subspaces (Xie et al., 2022), what determines which one is conscious? To what extent should thalamic activity also be present? To what degree should these distant cortical and subcortical sites be synchronized, within which band (most likely beta and theta), and over which time period, in order to count as a conscious representation? Beyond existing findings on long-distance beta synchrony and causality (Gaillard et al., 2009), global information sharing (King et al., 2013) and vector stability (Schurger et al., 2015), this work on the formation of a conscious assembly should be brought down to the level of parallel single-cell measures, using both recording and causal stimulation methods.

For RPT, a similar ambiguity surrounds the scope of the required recurrency: how many neurons and areas should participate in a feedback loop for it to be conscious? Is V1-V2 feedback enough, or should the feedback encompass the whole visual system up to the IT cortex? If so, how is the unity of consciousness achieved, considering that feedback will occur earlier and at different temporal intervals between closely connected areas and across the hierarchy? And how can integration be obtained between different contents, and between different modalities?

For HOT, it is still unclear how to map a theoretical distinction between first-order and higher-order representations onto anatomical predictions. Specifically, the theory does not provide explicit arguments for which properties of the brain should be regarded as the neural basis of consciousness. Thus, when facing a null result, it will always be possible to claim that some other brain area, or another type of processing, might still "host" the higher-order representation, yet it failed to be detected given the specific experimental conditions. More generally speaking, as explained by Fleming, the neuroscientific implementation of the theory is still being developed, so these questions might be answered by future formulations.

For IIT, the key challenge is that exact measurements of Φ cannot be obtained for large, multi-level systems. So far, tests of integrated information have relied on indirect measures, such as the PCI, which lack specificity because they are also compatible with other theories (for replies, see Tononi et al., in press). Approximations of Φ (based on IIT 4.0) that could be applied at the level of large-scale neuronal networks are being developed, though they have not been presented yet. Thus, currently, experimental attempts to test the theory (e.g., Cogitate Consortium et al., 2023; see also the INTERPID preregisteration: https://osf.io/4rn85) have accordingly focused on neural implementations of the theory, rather than on directly probing Φ. Other challenges are more conceptual: some criticized the axiomatic approach (Bayne, 2018), other claimed that the theory has consequences that they deemed improbable (e.g., that large grids could be highly conscious; Aaronson, 2014; see also the IIT discussion about grids above), to which IIT proponents replied (Tononi, 2014).

Finally, for PP, since the theory is far from being developed specifically for the purpose of explaining consciousness, it has yet to generate specific and clear predictions that can be directly tested. More generally speaking, if almost any circuit motif is compatible with PP and PE minimization, it is unclear what current finding can refute the theory.

This theoretical flexibility leaves too much room for the theories to adjust in light of counterevidence. To make progress, we accordingly hold that the theories should provide much more detailed and explicit explanations of their core principles and how they translate into the more auxiliary predictions (i.e., how central they are to the theory; Chis-Ciure et al., 2024). Only with such specifications will arbitrating between theories be truly meaningful, pushing the theories towards substantial revisions – and even refutations – and not only minor ones. This process, which will surely take time, would allow the field to move from "surrogates for theories" (Gigerenzer, 1998) to fully-fledged theories that are well-specified and accurately defined.

### 3.5. Consciousness: matter/life or function?

A key challenge for the field in future years would be to face the question of what phenomenology actually is and which systems, biological or artificial, might have it. For computational functionalist theories (e.g., GNWT and variants of HOT), instantiating the right kind of computation(s) would be enough to instantiate consciousness (Dehaene et al., 2017; see also Butlin et al., 2023; Seth, 2024b). Dehaene offered an illuminating example, whereby under GNWT, a system such as a phone endowed with information sharing capabilities would instantiate consciousness. The defining factor, according to GNWT, is whether the apps in the phone are independent of each other (as they are right now) or have the ability to flexibly exchange information across apps in order to fulfill a certain goal (as a global workspace would permit). Quoting Dennett, Dehaene argued "*consciousness is a functional property that has all sorts of gradations and all sorts of adaptations*". Yet, other theories (e.g., IIT) strongly oppose such computational/functionalist claims and instead offer structural explanations, according to which what matters for structure is how a system is built and not the functions that it performs, regardless of how sophisticated or smart those functions are (Findlay et al., 2024; Tononi and Raison, 2024, in press.). Specifically, IIT claims that computers that may replicate our behaviors or cognitive functions will not replicate our experiences (Findlay et al., 2024). Thus, according to IIT, a computer vision system in a self-driving car could act as if it did "see" functionally—recognize scenes and objects and move around the world much like we would—yet would not see anything phenomenally.

Notably, computational functionalism (Butlin et al., 2023) is a stronger assumption than functionalism in general, such that non-computational forms of functionalism are possible (Piccinini, 2004). Some flavors of predictive processing, such as advocated by Seth, indeed challenge the assumption of computational functionalism and suggest that consciousness may depend on the material properties of living systems – a form of 'biological naturalism' (Searle, 2017; Seth, 2021; Seth, 2024b). According to some versions of biological naturalism, consciousness is a high-level emergent process that evolved in the complex, hierarchically organized living systems (Feinberg, 2024). How to test these highly divergent stands was left open and unanswered. When tackling this issue, a key question is what function – if any – does consciousness serve. Taking a broader, evolutionary, perspective in the years to come might help illuminate the question of the adaptative function of consciousness and its dependence or independence from biological systems (Cleeremans and Tallon-Baudry, 2022; Feinberg and Mallatt, 2016; Cabral-Calderin et al., 2025).

### 3.6. So, what are we left with?

A clear conclusion arising from the discussion is that there is currently significantly more disagreement than agreement between the theories. And the disagreement pertains to the most basic questions that define the scientific investigation to begin with: what consciousness is, which states are conscious and which are not, what is required from a theory of consciousness, and what exactly should be explained. As the theories themselves either do not provide full answers to these questions, or provide conflicting answers, it seems like it is up to the scientific community to establish a consensus on these basic questions, which might be a necessary condition for making progress in addressing them and in comparing and arbitrating among theories. Such consensus can be achieved in one of two ways; inter-theoretical dialogue between theory proponents, or – conversely – an initiative that stems from theory-neutral experts, who will lay out the grounds for future investigations and for a more informed evaluation of the theories. Once we agree on what we are trying to explain, and what is considered as an acceptable explanation, we can start weighing in on which explanations are better than others.

To do so, it would be useful to widen the scope of our investigation;

thus far, experimental paradigms used to investigate consciousness have only probed some types of conscious experiences, with an over-representation of studies in the visual system (Yaron et al., 2022), mostly targeting perceptual consciousness using paradigms that are quite far from everyday conscious experiences (Mudrik et al., 2024). Thus, while much progress has been made in the field, we are just at the beginning. Much more empirical work is needed both to characterize the minimal properties of consciousness and to understand how those properties evolve and what function they serve. Alongside theoretical development and refinement, it would be beneficial to invest empirical efforts in accruing more data, exploring the parameter space of the phenomena, and defining the boundary conditions of the observations.

Here, more than ever, patience might pay off. Developing coherent and systematic theories beyond frameworks and hypotheses takes time. Premature refutation might stifle theory development and progress. Like the taste of a good, mature wine, theories might also offer more satisfactory explanations if they are allowed time to mature. During the maturation process, these theories could make their assumptions, bridging principles across levels, and derivations clear, while embracing intellectual humility, given the high number of unknowns

## Acknowledgments

LMu is supported by the European Research Council (ERC; grant number: 101077144). MB is supported by Templeton World Charity Foundation TWCF0646. MB also wishes to thank Dr. Matteo Grasso and Mr. Jeremiah Hendren for their assistance with preparing the IIT Figure. AS is supported by ERC Advanced Investigator Grant (101019254). SF is supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (selected as ERC Consolidator, grant number 101043666). LMe was supported by Templeton World Charity Foundation TWCF0389, and the Max Planck Society. We thank the ASSC and the local organizers of the Amsterdam meeting in 2022, Johannes Fahrenfort, Simon van Gaal and Steven Scholte. We would also like to thank the five reviewers of this manuscript for their highly thoughtful and valuable feedback that allowed us to substantially improve the manuscript.

## Data availability

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

## References

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A.M., Marshall, W., Mayner, W. G.P., Zaeemzadeh, A., Boly, M., Juel, B.E., 2023. Integrated information theory (IIT) 4.0: formulating the properties of phenomenal existence in physical terms. PLOS Comput. Biol. 19 (10), e1011465.

Albantakis, L., Hintze, A., Koch, C., Adami, C., Tononi, G., 2014. Evolution of integrated causal structures in animats exposed to environments of increasing complexity. PLOS Comput. Biol. 10 (12), e1003966.

Aru, J., Bachmann, T., Singer, W., Melloni, L., 2012. Distilling the neural correlates of consciousness. Neurosci. Biobehav. Rev. 36 (2), 737–746.

Aru, J., Suzuki, M., Larkum, M.E., 2020. Cellular mechanisms of conscious processing. Trends Cogn. Sci. 24 (10), 814–825.

Aru, J., Suzuki, M., Rutiku, R., Larkum, M.E., Bachmann, T., 2019. Coupling the state and contents of consciousness. Front. Syst. Neurosci. 13, 43.

Baars, B.J., 1997. In the theatre of consciousness. J. Conscious. Stud. 4, 292–309.

Bachmann, T., Hudetz, A.G., 2014. It is time to combine the two main traditions in the research on the neural correlates of consciousness: C= L× D. Front. Psychol. 5, 940.

Bang, D., Ershadmanesh, S., Nili, H., Fleming, S.M., 2020. Private–public mappings in human prefrontal cortex. eLife 9, e56477.

Bang, D., Fleming, S.M., 2018. Distinct encoding of decision confidence in human medial prefrontal cortex. Proc. Natl. Acad. Sci. USA 115 (23), 6082–6087.

Barrett, D., Heale, R., 2020. What are Delphi studies? Evid. -Based Nurs. 23 (3), 68–69.

Barrett, A.B., Mediano, P.A.M., 2019. The Phi measure of integrated information is not well-defined for general physical systems. J. Conscious. Stud. 26 (1-2), 11–20.

Barrett, L.F., Simmons, W.K., 2015. Interoceptive predictions in the brain. Nat. Rev. Neurosci. 16, 419–429.

Barttfeld, P., Uhrig, L., Sitt, J.D., Sigman, M., Jarraya, B., Dehaene, S., 2015. Signature of consciousness in the dynamics of resting-state brain activity. Proc. Natl. Acad. Sci. USA 3, 887–892.

Bayne, T., 2018. On the axiomatic foundations of the integrated information theory of consciousness. Neurosci. Conscious. 2018 (1), niy007.

Bekinschtein, T.A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., Naccache, L., 2009. Neural signature of the conscious processing of auditory regularities. Proc. Natl. Acad. Sci. USA 106 (5), 1672–1677.

Bellet, M.E., Gay, M., Bellet, J., Jarraya, B., Dehaene, S., van Kerkoerle, T., Panagiotaropoulos, T.I., 2024. Spontaneously emerging internal models of visual sequences combine abstract and event-specific information in the prefrontal cortex. Cell Rep. 43 (3).

Bellet, J., Gay, M., Dwarakanath, A., Jarraya, B., van Kerkoerle, T., Dehaene, S., Panagiotaropoulos, T.I., 2022. Decoding rapidly presented visual stimuli from prefrontal ensembles without report nor post-perceptual processing. Neurosci. Conscious. 2022 (1), niac005.

Block, N., 1995. On a confusion about a function of consciousness. Behav. Brain Sci. 18 (2), 227–287.

Block, N., 2011. Perceptual consciousness overflows cognitive access. Trends Cogn. Sci. 15 (12), 567–575.

Boly, M., Garrido, M.I., Gosseries, O., Bruno, M.A., Boveroux, P., Schnakers, C., Massimini, M., Litvak, V., Laureys, S., Friston, K., 2011. Preserved feedforward but impaired top-down processes in the vegetative state. Science 332 (6031), 858–862.

Boly, M., Massimini, M., Tsuchiya, N., Postle, B.R., Koch, C., Tononi, G., 2017. Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and neuroimaging evidence. J. Neurosci. 37 (40), 9603–9613.

Boly, M., Smith, R., Vigueras Borrego, G., Pozuelos, J.P., Allaudin, T., Malinowski, P., Tononi, G., 2024. Neural correlates of pure presence. bioRxiv 2024, 2004. 2018.590081.

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

Brown, R., Lau, H., LeDoux, J.E., 2019. Understanding the higher-order approach to consciousness. Trends Cogn. Sci. 23 (9), 754–768.

Butlin, P., Long, R., Elmoznino, E. c., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M.A.K., Schwitzgebel, E., Simon, J., & VanRullen, R. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv preprint arXiv:2308.08708.[Record #3579 is using a reference type undefined in this output style.].

Cabral-Calderin Y, Hechavarria J, Melloni L (2025). Towards a neuroethologica approach to consciousness. https://doi.org/10.31234/osf.io/wyrhu.

Carruthers, 2001. Higher-order theories of consciousness. Stanf. encycl. philos.

Casarotto, S., Comanducci, A., Rosanova, M., Sarasso, S., Fecchio, M., Napolitani, M., Pigorini, A., G. Casali, A., Trimarchi, P.D., Boly, M., Gosseries, O., Bodart, O., F., C., Landi, C., Mariotti, M., Devalle, G., Laureys, S., Tononi, G., Massimini, M., 2016. Stratification of unresponsive patients by an independently validated index of brain complexity. Ann. Neurol. 80 (5), 718–729.

Charles, L., Van Opstal, F., Marti, S., Dehaene, S., 2013. Distinct brain mechanisms for conscious versus subliminal error detection. Neuroimage 73, 80–94.

Chis-Ciure, R., Melloni, L., Northoff, G., 2024. A measure centrality index for systematic empirical comparison of consciousness theories. Neurosci. Biobehav. Rev., 105670

Clark, A., 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci. 36 (3), 181–204.

Cleeremans, A., Achoui, D., Beauny, A., Keuninckx, L., Martin, J.R., Muñoz-Moldes, S., Vuillaume, L., De Heering, A., 2020. Learning to be conscious. Trends Cogn. Sci. 24 (2), 112–123.

Cleeremans, A., Tallon-Baudry, C., 2022. Consciousness matters: phenomenal experience has functional value. Neurosci. Conscious. 2022 (1), niac007.

Cogitate Consortium, Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., Lepauvre, A., Liu, L., Richter, D., Vidal, Y., Niccolò, B., Brown, T., Sripad, P., Armendariz, M., Bendtz, K., Ghafari, T., Hetenyi, D., Jeschke, J., Kozma, C., Mazumder, D.R., Montenegro, S., Seedat, A., Sharafeldin, A., Yang, S., Baillet, S., Chalmers, D.J., Cichy, R.M., Fallon, F., Panagiotaropoulos, T.I., Blumenfeld, H., de Lange, F.P., Devore, S., Jensen, O., Kreiman, G., Luo, H., Boly, M., Dehaene, S., Koch, C., Tononi, G., Pitts, M., Mudrik, L., & Melloni, L. (2023). An adversarial collaboration to critically evaluate theories of consciousness. bioRxiv, 2023.2006. 2023.546249.

Comolatti, R., Grasso, M., & Tononi, G. 2024. Why does time feel the way it does? Towards a principled account of temporal experience. arXiv preprint arXiv: 2412.13198.

Cortese, A., Amano, K., Koizumi, A., Kawato, M., Lau, H., 2016. Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. Nat. Commun. 7 (1), 13669.

Cortese, A., Lau, H., Kawato, M., 2020. Unconscious reinforcement learning of hidden brain states supported by confidence. Nat. Commun. 11 (1), 4429.

Crick, F., Koch, C., 2003. A framework for consciousness. Nat. Neurosci. 6 (2), 119–126.

Csikszentmihalyi, M., Abuhamdeh, S., Nakamura, J., 2005. Flow. Handbook of competence and motivation, pp. 598-608..

D'Angelo, E., Casali, S., 2013. Seeking a unified framework for cerebellar function and dysfunction: from circuit operations to cognition. Front. Neural Circuits 6, 116.

De Martino, B., Fleming, S.M., Garrett, N., Dolan, R.J., 2013. Confidence in value-based choice. Nat. Neurosci. 16 (1), 105–110.

Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., Sablé-Meyer, M., 2022. Symbols and mental programs: a hypothesis about human singularity. Trends Cogn. Sci. 26 (9), 751–766.

Dehaene, S., Changeux, J.P., 1997. A hierarchical neuronal network for planning behavior. Proc. Natl. Acad. Sci. USA 94, 13293–13298.

Dehaene, S., Changeux, J.P., 2005. Ongoing spontaneous activity controls access to consciousness: a neuronal model for inattentional blindness (May). PLOS Biol. 3 (5), e141.
Dehaene, S., Changeux, J.P., Naccache, L., Sackur, J., Sergent, C., 2006. Conscious, preconscious, and subliminal processing: a testable taxonomy (May). Trends Cogn. Sci. 10 (5), 204–211.
Dehaene, S., Kerszberg, M., Changeux, J.P., 1998. A neuronal model of a global workspace in effortful cognitive tasks. Proc. Natl. Acad. Sci. USA 95 (24), 14529–14534.
Dehaene, S., Lau, H., Kouider, S., 2017. What is consciousness, and could machines have it? Science 358 (6362), 486–492.
Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., Pallier, C., 2015. The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. Neuron 88 (1), 2–19.
Dehaene, S., Naccache, L., 2001. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework (Apr). Cognition 79 (1-2), 1–37.
Dehaene, S., Sergent, C., Changeux, J.P., 2003. A neuronal network model linking subjective reports and objective physiological data during conscious perception. Jul 8 Proc. Natl. Acad. Sci. USA 100 (14), 8520–8525.
Demertzi, A., Tagliazucchi, E., Dehaene, S., Deco, G., Barttfeld, P., Raimondo, F., Martial, C., Fernández-Espejo, D., Rohaut, B., Voss, H., 2019. Human consciousness is supported by dynamic complex patterns of brain signal coordination. Sci. Adv. 5 (2), eaat7603.
Dijkstra, N., Warrington, O., Kok, P., Fleming, S.M., 2023. Distinguishing neural correlates of prediction errors on perceptual content and awareness of content 2..
Dijkstra, N., Warrington, O., Kok, P., Fleming, S.M., 2024. Distinguishing neural correlates of prediction errors on perceptual content and detection of content. J. Cogn. Neurosci. 1–16.
Doerig, A., Schurger, A., Herzog, M.H., 2020. Hard criteria for empirical theories of consciousness. Cogn. Neurosci. 1–22.
Doerig, A., Schurger, A., Hess, K., Herzog, M.H., 2019. The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. Conscious. Cogn. 72, 49–59.
Evers, K., Farisco, M., Pennartz, C.M.A., 2024. Assessing the commensurability of theories of consciousness: on the usefulness of common denominators in differentiating, integrating and testing hypotheses. Conscious. Cogn. 119, 103668.
Fahrenfort, J.J., van Leeuwen, J., Olivers, C.N.L., Hogendoorn, H., 2017. Perceptual integration without conscious access. Proc. Natl. Acad. Sci. USA 114 (14), 3744–3749.
Fahrenfort, J.J., Scholte, H.S., Lamme, V.A.F., 2007. Masking disrupts reentrant processing in human visual cortex. J. Cogn. Neurosci. 19 (9), 1488–1497.
Fahrenfort, J.J., Snijders, T.M., Heinen, K., Van Gaal, S., Scholte, H.S., Lamme, V.A.F., 2012. Neuronal integration in visual cortex elevates face category tuning to conscious face perception. Proc. Natl. Acad. Sci. USA 109 (52), 21504–21509.
Feinberg, T.E., 2024. From Sensing to Sentience: How Feeling Emerges from the Brain. MIT Press.
Feinberg, T.E., Mallatt, J., 2016. The Ancient Origins of Consciousness: How the Brain Created Experience. MIT Press.
Feldman, H., Friston, K.J., 2010. Attention, uncertainty, and free-energy. Front. Hum. Neurosci. 4, 215.
Findlay, G., Marshall, W., Albantakis, L., David, I., Mayner, W.G.P., Koch, C., Tononi, G., 2024. Dissociating artificial intelligence from artificial consciousness. arXiv preprint arXiv:2412.04571..
Fleming, S.M., 2020. Awareness as inference in a higher-order state space. Neurosci. Conscious. 2020 (1), niz020.
Fleming, S.M., 2024. Metacognition and confidence: a review and synthesis. Annu. Rev. Psychol. 75, 241–268.
Fleming, S.M., Shea, N., 2024. Quality space computations for consciousness. Trends Cogn. Sci. 28 (10), 896–906.
Fletcher, P.C., Frith, C.D., 2009. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. Nat. Rev. Neurosci. 10 (1), 48–58.
Friston, K.J., Daunizeau, J., Kilner, J., Kiebel, S.J., 2010. Action and behavior: a free-energy formulation. Biol. Cybern. 102, 227–260.
Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., Pezzulo, G., 2015. Active inference and epistemic value. Cogn. Neurosci. 6 (4), 187–214.
van Gaal, S., Lamme, V.A.F., 2012. Unconscious high-level information processing implication for neurobiological theories of consciousness. Neuroscientist 18 (3), 287–301.
van Gaal, S., Ridderinkhof, K.R., Fahrenfort, J.J., Scholte, H.S., Lamme, V.A.F., 2008. Frontal cortex mediates unconsciously triggered inhibitory control. Aug 6 J. Neurosci. 28 (32), 8053–8062.
Gaillard, R., Dehaene, S., Adam, C., Clémenceau, S., Hasboun, D., Baulac, M., Cohen, L., Naccache, L., 2009. Converging intracranial markers of conscious access. PLOS Biol. 7 (3), e1000061.
Gazzaniga, M.S., 1967. The split brain in man. Sci. Am. 217 (2), 24–29.
Gherman, S., Philiastides, M.G., 2018. Human VMPFC encodes early signatures of confidence in perceptual decisions. eLife 7, e38293.
Gigerenzer, G., 1998. Surrogates for theories. Theory Psychol. 8 (2), 195–204.
Godfrey-Smith, P., 2016. Mind, matter, and metabolism. J. Philos. 113 (10), 481–506.
Gregory, R.L., 1970. Intell. eye.
Haarsma, J., Kaltenmaier, A., Fleming, S.M., & Kok, P. (2024). Expectations about presence enhance the influence of content-specific expectations on low-level orientation judgements. *bioRxiv*, 2024.2002. 2022.581334.
Hardstone, R., Zhu, M., Flinker, A., Melloni, L., Devore, S., Friedman, D., Dugan, P., Doyle, W.K., Devinsky, O., He, B.J., 2021. Long-term priors influence visual perception through recruitment of long-range feedback. Nat. Commun. 12 (1), 6288.
Haun, A., Tononi, G., 2019. Why does space feel the way it does? Towards a principled account of spatial experience. Entropy 21 (12), 1160.
He, B.J., 2023. Towards a pluralistic neurobiological understanding of consciousness. Trends Cogn. Sci. 27 (5), 420–432.
Helmholtz, H., 1867/1962. Treatise on Physiological Optics. Dover.
Hohwy, J., Roepstorff, A., Friston, K., 2008. Predictive coding explains binocular rivalry: an epistemological review (Sep). Cognition 108 (3), 687–701.
Hohwy, J., Seth, A.K., 2020. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. Philos. Mind Sci. 1 (II).
IIT Concerned et al. (2025). What makes a theory of consciousness scientific?. Nature neuroscience (In press).
James, W., 1892. The stream of consciousness. Psychology 151–175.
Juan, E., Górska, U., Kozma, C., Papantonatos, C., Bugnon, T., Denis, C., Kremen, V., Worrell, G., Struck, A.F., Bateman, L.M., Merricks, E.M., Blumenfeld, H., Tononi, G., Schevon, C., Boly, M., 2023. Distinct signatures of loss of consciousness in focal impaired awareness versus tonic-clonic seizures. Brain 146 (1), 109–123.
Kapoor, V., Dwarakanath, A., Safavi, S., Werner, J., Besserve, M., Panagiotaropoulos, T. I., Logothetis, N.K., 2022. Decoding internally generated transitions of conscious contents in the prefrontal cortex without subjective reports. Nat. Commun. 13 (1), 1–16.
King, J.R., Gramfort, A., Schurger, A., Naccache, L., Dehaene, S., 2014. Two distinct dynamic modes subtend the detection of unexpected sounds. PLOS One 9 (1), e85791.
King, J.R., Sitt, J.D., Faugeras, F., Rohaut, B., El Karoui, I., Cohen, L., Naccache, L., Dehaene, S., 2013. Information sharing in the brain indexes consciousness in noncommunicative patients. Curr. Biol. 23 (19), 1914–1919.
Kirchberger, L., Mukherjee, S., Schnabel, U.H., van Beest, E.H., Barsegyan, A., Levelt, C. N., Heimel, J.A., Lorteije, J.A.M., van der Togt, C., Self, M.W., 2021. The essential role of recurrent processing for figure-ground perception in mice. Sci. Adv. 7 (27), eabe1833.
Klatzmann, U., Froudist-Walsh, S., Bliss, D.P., Theodoni, P., Mejías, J., Niu, M., Rapan, L., Palomero-Gallagher, N., Sergent, C., Dehaene, S., Wang, X.-J., 2023. A connectome-based model of conscious access in monkey cortex. , 2022.2002.2020.481230. bioRxiv.
Kouider, S., Dehaene, S., 2007. Levels of processing during non-conscious perception: a critical review of visual masking. MAY 29 Philos. Trans. R. Soc. B-Biol. Sci. 362 (1481), 857–875.
Kuhn, R.L., 2024. A landscape of consciousness: Toward a taxonomy of explanations and implications. Prog. Biophys. Mol. Biol.
Lakatos, I., 1978. The Methodology of Scientific Research Programmes: Ed by John Worrall and Gregory Currie. Cambridge University Press.
Lamme, V.A.F., 2006. Towards a true neural stance on consciousness. Sep 22 Trends Cogn. Sci. 10 (11), 494–501.
Lamme, V.A.F. (2010). How neuroscience will change our view on consciousness. Cognitive neuroscience, 1(3), 204-220.
Lamme, V.A.F., 2014. The crack of dawn: perceptual functions and neural mechanisms that mark the transition from unconscious processing to conscious vision. Open MIND. MIND Group, Frankfurt am Main.
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 (1755), 20170344.
Lamme, V.A.F., 2020. Visual functions generating conscious seeing. Front. Psychol. 11, 83.
Lamme, V.A.F., Roelfsema, P.R., 2000. The distinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci. 23 (11), 571–579.
Lamme, V.A.F., Zipser, K., Spekreijse, H., 1998. Figure-ground activity in primary visual cortex is suppressed by anesthesia. Proc. Natl. Acad. Sci. USA 95 (6), 3263–3268.
Lamme, V.A.F., Zipser, K., Spekreijse, H., 2001. Masking interrupts figure-ground signals in V1.. OCT 1 J. Cogn. Neurosci. 14 (7), 1044–1053.
Landman, R., Spekreijse, H., Lamme, V.A.F., 2003. Large capacity storage of integrated objects before change blindness. Vis. Res. 43 (2), 149–164.
de Lange, F.P., Heilbron, M., Kok, P., 2018. How do expectations shape perception? Trends Cogn. Sci. 22 (9), 764–779.
Lau, H., 2019. Consciousness, Metacognition, & Perceptual Reality Monitoring. PsyArXiv.
Lau, H., 2022. In Consciousness We Trust: The Cognitive Neuroscience of Subjective Experience. Oxford University Press.
Lau, H., Michel, M., LeDoux, J.E., Fleming, S.M., 2022. The mnemonic basis of subjective experience. Nat. Rev. Psychol. 1 (8), 479–488.
Lau, H., Rosenthal, D., 2011. Empirical support for higher-order theories of conscious awareness. Trends Cogn. Sci. 15 (8), 365–373.
LeDoux, J.E., 2020a. The Deep History of Ourselves: The Four-billion-year Story of How We Got Conscious Brains. Penguin.
LeDoux, J.E., 2020b. How does the non-conscious become conscious? Curr. Biol. 30 (5), R196–R199.
Lepauvre, A., Melloni, L., 2021. The search for the neural correlate of consciousness: progress and challenges. , 2phimisci.2021.2087.Philos. Mind Sci..
Mack, A., Rock, I., 1998. Inattentional Blindness. MIT Press.
Maniscalco, B., Lau, H., 2012. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Conscious. Cogn. 21 (1), 422–430.
Markov, N.T., Ercsey-Ravasz, M., Van Essen, D.C., Knoblauch, K., Toroczkai, Z., Kennedy, H., 2013. Cortical high-density counterstream architectures. Science 342 (6158), 1238406.

Marti, S., Sackur, J., Sigman, M., Dehaene, S., 2010. Mapping introspection's blind spot: reconstruction of dual-task phenomenology using quantified introspection. Cognition 115 (2), 303–313.
Marti, S., Sigman, M., Dehaene, S., 2012. A shared cortical bottleneck underlying attentional blink and psychological refractory period. Neuroimage 59 (3), 2883–2898.
Maruoka, H., Nakagawa, N., Tsuruno, S., Sakai, S., Yoneda, T., Hosoya, T., 2017. Lattice system of functionally distinct cell types in the neocortex. Science 358 (6363), 610–615.
Mashour, G.A., Roelfsema, P.R., Changeux, J.P., Dehaene, S., 2020. Conscious processing and the global neuronal workspace hypothesis. Neuron 105, 776–798.
Masset, P., Ott, T., Lak, A., Hirokawa, J., Kepecs, A., 2020. Behavior-and modality-general representation of confidence in orbitofrontal cortex. Cell 182 (1), 112–126 e118.
Massimini, M., Ferrarelli, F., Huber, R., Esser, S.K., Singh, H., Tononi, G., 2005. Breakdown of cortical effective connectivity during sleep. Science 309 (5744), 2228–2232.
Mayner, W.G.P., Juel, B.E., Tononi, G., 2024. Intrinsic meaning, perception, and matching. .arXivpreprint arXiv:241221111..
McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748.
Mediano, P.A.M., Rosas, F.E., Bor, D., Seth, A.K., Barrett, A.B., 2022. The strength of weak integrated information theory. Trends Cogn. Sci. 26 (8), 646–655.
Merker, B., 2007. Grounding consciousness: the mesodiencephalon as thalamocortical base. Behav. Brain Sci. 30 (1), 110–120.
Metzinger, T., 2004. Being No One: The Self-model Theory of Subjectivity. mit Press.
Mudrik, L., Deouell, L.Y., 2022. Neuroscientific evidence for processing without awareness. Annu. Rev. Neurosci. 45, 403–423.
Mudrik, L., Hirschhorn, R., Korisky, U., 2024. Taking consciousness for real: Increasing the ecological validity of the study of conscious vs. unconscious processes. Neuron.
Naccache, L., Blandin, E., Dehaene, S., 2002. Unconscious masked priming depends on temporal attention (SEP). Psychol. Sci. 13 (5), 416–424.
Nieder, A., Wagener, L., Rinnert, P., 2020. A neural correlate of sensory consciousness in a corvid bird. Science 369 (6511), 1626–1629.
Northoff, G., Lamme, V.A.F., 2020. Neural signs and mechanisms of consciousness: Is there a potential convergence of theories of consciousness in sight? Neurosci. Biobehav. Rev. 118, 568–587.
Novicky, F., Parr, T., Friston, K., Mirza, M.B., Sajid, N., 2024. Bistable perception, precision and neuromodulation. Cereb. Cortex 34 (1), bhad401.
Odegaard, B., Knight, R.T., Lau, H., 2017. Should a few null findings falsify prefrontal theories of conscious perception? J. Neurosci. 37 (40), 9593–9602.
Panagiotaropoulos, T.I., Deco, G., Kapoor, V., Logothetis, N.K., 2012. Neuronal discharges and gamma oscillations explicitly reflect visual consciousness in the lateral prefrontal cortex. Neuron 74 (5), 924–935.
Parr, T., Corcoran, A.W., Friston, K.J., Hohwy, J., 2019. Perceptual awareness and active inference. Neurosci. Conscious. 2019 (1), niz012.
Paul, L.K., Brown, W.S., Adolphs, R., Tyszka, J.M., Richards, L.J., Mukherjee, P., Sherr, E. H., 2007. Agenesis of the corpus callosum: genetic, developmental and functional aspects of connectivity. Nat. Rev. Neurosci. 8 (4), 287–299.
Phillips, W.A., Bachmann, T., Spratling, M.W., Muckli, L., Petro, L.S., Zolnik, T., 2024. Cellular psychology: relating cognition to context-sensitive pyramidal cells. Trends Cogn. Sci.
Piccinini, G., 2004. Functionalism, computationalism, and mental contents. Can. J. Philos. 34 (3), 375–410.
Piccinini, G., 2020. Neurocognitive Mechanisms: Explaining Biological Cognition. Oxford University Press.
Pigorini, A., Sarasso, S., Proserpio, P., Szymanski, C., Arnulfo, G., Casarotto, S., Fecchio, M., Rosanova, M., Mariotti, M., Russo, G.L., Palva, J.M., Nobili, L., Massimini, M., 2015. Bistability breaks-off deterministic responses to intracortical stimulation during non-REM sleep. Neuroimage 112, 105–113.
Pitts, M., Fennelly, P.D., Martinez, A., Hillyard, S.A., 2014. Gamma band activity and the P3 reflect post-perceptual processes, not visual awareness. Neuroimage 101, 337–350.
Promet, L., Bachmann, T., 2022. A comparative analysis of empirical theories of consciousness. Psychol. Conscious.: Theory, Res., Pract.
Rosenthal, D., 2005. Consciousness and Mind. Clarendon Press.
Rosenthal, D., 2010. How to think about mental qualities. Philos. Issues 20, 368–393.
Sandved-Smith, L., Hesp, C., Mattout, J., Friston, K., Lutz, A., Ramstead, M.J.D., 2021. Towards a computational phenomenology of mental action: modelling meta-awareness and attentional control with deep parametric active inference. Neurosci. Conscious. 2021 (1), niab018.
Sattin, D., Magnani, F.G., Bartesaghi, L., Caputo, M., Fittipaldo, A.V., Cacciatore, M., Picozzi, M., Leonardi, M., 2021. Theoretical models of consciousness: a scoping review. Brain Sci. 11 (5), 535.
Scholte, H.S., Witteveen, S.C., Spekreijse, H., Lamme, V.A.F., 2006. The influence of inattention on the neural correlates of scene segmentation. Brain Res. 1076 (1), 106–115.
Schurger, A., Graziano, M., 2022. Consciousness explained or described? Neurosci. Conscious. 2022 (1), niac001.
Schurger, A., Sarigiannidis, I., Naccache, L., Sitt, J.D., Dehaene, S., 2015. Cortical activity is more stable when sensory stimuli are consciously perceived. Proc. Natl. Acad. Sci. USA 112, E2083–E2092.
Searle, J.R., 2007. Dualism revisited. J. Physiol. 101 (4-6), 169–178.
Searle, J.R., 2017. Biological naturalism. The Blackwell companion to consciousness, pp. 327-336..
Sergent, C., Baillet, S., Dehaene, S., 2005. Timing of the brain events underlying access to consciousness during the attentional blink (Oct). Nat. Neurosci. 8 (10), 1391–1400.
Sergent, C., Wyart, V., Babo-Rebelo, M., Cohen, L., Naccache, L., Tallon-Baudry, C., 2013. Cueing attention after the stimulus is gone can retrospectively trigger conscious perception. Curr. Biol. 23 (2), 150–155.
Seth, A.K., 2013. Interoceptive inference, emotion, and the embodied self. Trends Cogn. Sci. 17, 565–573.
Seth, A.K., 2015. Presence, objecthood, and the phenomenology of predictive perception. Cogn. Neurosci. 6 (2-3), 111–117.
Seth, A.K., 2018. Consciousness: The last 50 years (and the next). Brain Neurosci. Adv. 2, 2398212818816019.
Seth, A., 2021. Being You: A New Science of Consciousness. Penguin.
Seth, A.K., 2024b. Conscious Artificial Intelligence and Biological Naturalism. PsyArXiv.
Seth, A.K. (2024a). Conscious artificial intelligence and biological naturalism. PsyArXiv preprint.
Seth, A.K., Bayne, T., 2022. Theories of consciousness. Nat. Rev. Neurosci. 23 (7), 439–452.
Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., Frith, C.D., 2014. Supra-personal cognitive control and metacognition. Trends Cogn. Sci. 18 (4), 186–193.
Shekhar, M., Rahnev, D., 2018. Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. J. Neurosci. 38 (22), 5078–5087.
Sigman, M., Dehaene, S., 2008. Brain mechanisms of serial and parallel processing during dual-task performance. J. Neurosci. 28 (30), 7585–7598.
Signorelli, C.M., Szczotka, J., Prentner, R., 2021. Explanatory profiles of models of consciousness-towards a systematic classification. Neurosci. Conscious. 2021 (2), niab021.
Sligte, I.G., Vandenbroucke, A.R.E., Scholte, H.S., Lamme, V.A.F., 2010. Detailed sensory memory, sloppy working memory. Front. Psychol. 1, 175.
Solms, M., 2019. The hard problem of consciousness and the free energy principle. Front. Psychol. 9, 2714.
Solomon, S.S., Tang, H., Sussman, E., Kohn, A., 2021. Limited evidence for sensory prediction error responses in visual cortex of macaques and humans. Cereb. Cortex 31 (6), 3136–3152.
Stephan, K.E., Manjaly, Z.M., Mathys, C.D., Weber, L.A.E., Paliwal, S., Gard, T., Tittgemeyer, M., Fleming, S.M., Haker, H., Seth, A.K., 2016. Allostatic self-efficacy: a metacognitive theory of dyshomeostasis-induced fatigue and depression. Front. Hum. Neurosci. 10, 550.
Storm, J.F., Klink, P.C., Aru, J., Senn, W., Goebel, R., Pigorini, A., Avanzini, P., Vanduffel, W., Roelfsema, P.R., Massimini, M., Larkum, M.E., Pennartz, C.M.A., 2024. An integrative, multiscale view on neural theories of consciousness. Neuron 112 (10), 1531–1552.
Super, H., Lamme, V.A.F., 2007. Altered figure-ground perception in monkeys with an extra-striate lesion. Neuropsychologia 45 (14), 3329–3334.
Super, H., Spekreijse, H., Lamme, V.A., 2001. Two distinct modes of sensory processing observed in monkey primary visual cortex (V1) (Mar). Nat. Neurosci. 4 (3), 304–310.
Suzuki, K., Roseboom, W., Schwartzman, D.J., Seth, A.K., 2017. A deep-dream virtual reality platform for studying altered perceptual phenomenology. Sci. Rep. 7 (1), 1–11.
Suzuki, K., Seth, A.K., Schwartzman, D.J., 2023. Modelling phenomenological differences in aetiologically distinct visual hallucinations using deep neural networks. Front. Hum. Neurosci. 17.
Tian, Z., Chen, J., Zhang, C., Min, B., Xu, B., Wang, L., 2024. Mental programming of spatial sequences in working memory in the macaque frontal cortex. Science 385 (6716), eadp6091.
Tononi, G., 2008. Consciousness as integrated information: a provisional manifesto. Biol. Bull. 215 (3), 216–242.
Tononi, G., 2014. Why Scott should stare at a blank wall and reconsider (or, the conscious grid). Shtetl-Optim.: Blog Scott. Aaronson.
Tononi, G., Albantakis, L., Barbosa, L., Boly, M., Cirelli, C., Comolatti, E., Ellia, F., Grasso, M., Haun, A., Hendren, J., Hoel, E., Koch, C., Maier, A., Marshall, W., Massimini, M., Mayner, W., Oizumi, M., Szczotka, J., Tsuchiya, N., Zaeemzadeh, A., in press. Consciousness and pseudo-consciousness: a clash between two paradigms. Nature neuroscience..
Tononi, G., Boly, M., Massimini, M., Koch, C., 2016. Integrated information theory: from consciousness to its physical substrate. Nat. Rev. Neurosci. 17 (7), 450–461.
Tononi, G., Raison, C., 2025. AI, consciousness, and psychiatry (in press.). World Psychiatry.
Tschantz, A., Seth, A.K., Buckley, C.L., 2020. Learning action-oriented models through active inference. PLOS Comput. Biol. 16 (4), e1007805.
Tscshantz, A., Millidge, B., Seth, A.K., Buckley, C.L., 2023. Hybrid predictive coding: inferring, fast and slow. PLOS Comput. Biol. 19 (8), e1011280.
Tsuchiya, N., Wilke, M., Frässle, S., Lamme, V.A.F., 2015. No-report paradigms: extracting the true neural correlates of consciousness. Trends Cogn. Sci. 19 (12), 757–770.
Van Gulick, R., 2004. Higher-order global states (HOGS): an alternative higher-order model. High. -Order Theor. Conscious. 2004, 67–93.
Van Vugt, B., Dagnino, B., Vartak, D., Safaai, H., Panzeri, S., Dehaene, S., Roelfsema, P. R., 2018. The threshold for conscious report: signal loss and response bias in visual and frontal cortex. Science 360 (6388), 537–542.
Wang, L., Mruczek, R.E.B., Arcaro, M.J., Kastner, S., 2015. Probabilistic maps of visual topography in human cortex. Cereb. Cortex 25 (10), 3911–3931.
Ward, L.M., 2011. The thalamic dynamic core theory of conscious experience. Conscious. Cogn. 20 (2), 464–486.
Weiskrantz, L., 1997. Consciousness Lost and Found. Oxford University Press.
Whyte, C.J., Corcoran, A.W., Robinson, J., Smith, R., Moran, R.J., Parr, T., Friston, K.J., Seth, A.K., Hohwy, J., 2024. On the Minimal Theory of Consciousness Implicit in Active Inference. arXiv preprint arXiv:2410.06633..

Whyte, C.J., Hohwy, J., Smith, R., 2022. An active inference model of conscious access: How cognitive action selection reconciles the results of report and no-report paradigms. Curr. Res. Neurobiol. 3, 100036.
Xie, Y., Hu, P., Li, J., Chen, J., Song, W., Wang, X.J., Yang, T., Dehaene, S., Tang, S., Min, B., Wang, L., 2022. Geometry of sequence working memory in macaque prefrontal cortex. Science 375 (6581), 632–639.
Yaron, I., Melloni, L., Pitts, M., Mudrik, L., 2022. The ConTraSt database for analysing and comparing empirical studies of consciousness theories. Nat. Hum. Behav. 6 (4), 593–604.
Zaeemzadeh, A., Tononi, G., 2024. Shannon information and integrated information: message and meaning. arXiv preprint arXiv:2412.10626..