What is neurorepresentationalism?

From neural activity and predictive processing to multi-level representations and consciousness

Pennartz, C.M.A.

DOI
10.1016/j.bbr.2022.113969

Publication date
2022

Document Version
Final published version

Published in
Behavioural Brain Research

Behavioural Brain Research 432 (2022) 113969

Contents lists available at ScienceDirect

Behavioural Brain Research

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

What is neurorepresentationalism? From neural activity and predictive processing to multi-level representations and consciousness

Cyriel M.A. Pennartz [a,b,*]

[a] Swammerdam Institute for Life Sciences, Center for Neuroscience, Faculty of Science, University of Amsterdam, the Netherlands
[b] Research Priority Program Brain and Cognition, University of Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Awareness
Deliberation
Evolution
Goal-directed behavior
Intentionality
Model-based learning
Planning
Predictive coding
Situatedness

ABSTRACT

This review provides an update on Neurorepresentationalism, a theoretical framework that defines conscious experience as multimodal, situational survey and explains its neural basis from brain systems constructing best-guess representations of sensations originating in our environment and body (Pennartz, 2015). It posits that conscious experience is characterized by five essential hallmarks: (i) multimodal richness, (ii) situatedness and immersion, (iii) unity and integration, (iv) dynamics and stability, and (v) intentionality. Consciousness is furthermore proposed to have a biological function, framed by the contrast between reflexes and habits (not requiring consciousness) versus goal-directed, planned behavior (requiring multimodal, situational survey). Conscious experience is therefore understood as a sensorily rich, spatially encompassing representation of body and environment, while we nevertheless have the impression of experiencing external reality directly. Contributions to understanding neural mechanisms underlying consciousness are derived from models for predictive processing, which are trained in an unsupervised manner, do not necessarily require overt action, and have been extended to deep neural networks. Even with predictive processing in place, however, the question remains why this type of neural network activity would give rise to phenomenal experience. Here, I propose to tackle the Hard Problem with the concept of multi-level representations which emergently give rise to multimodal, spatially wide superinferences corresponding to phenomenal experiences. Finally, Neurorepresentationalism is compared to other neural theories of consciousness, and its implications for defining indicators of consciousness in animals, artificial intelligence devices and immobile or unresponsive patients with disorders of consciousness are discussed.

1. Introduction

What is consciousness? An easy and straightforward attempt at a definition follows from the realization that we regain consciousness in the morning, after a good night's sleep, and lose it again when we fall asleep at night. In the morning we wake up, perceive the world around us, realize that it is us who are experiencing this situation, start making decisions and engage in other mental activities such as imagery, planning and memory retrieval. However, terms like 'waking up', 'perceiving' and 'experience' lead us straight into circularity, as their definition requires referral to the same phenomenon captured by 'consciousness'. Therefore, I will defer a definition to a later point in this review, except to emphasize the distinction between consciousness as a *state* and the *content* of conscious experience. In colloquial conversation, expressions like 'he is conscious' or 'he is regaining consciousness' refer to a particular brain state that stands in contrast to nonconscious states such as deep sleep, coma or anesthesia, and the third-person evidence for such expressions arises from behavioral signs and verbal reporting. In addition to brain state descriptions, other characterizations of conscious state can be offered (meta-intentional, phenomenal states, etc.), but these will not be explicitly considered here. Knowledge about being in a state of consciousness tells us little about *what* another person may be experiencing – the subjectively experienced (or phenomenal) content of consciousness. When a subject regaining consciousness opens

Abbreviations: AI, Artificial Intelligence; DoCs, Disorders of consciousness; GDB, Goal-directed Behavior; IoC, Indicator of Consciousness; NCC, neural correlate of consciousness; PP, predictive processing; IIT, Integrated Information Theory; GNWT, Global Neuronal Workspace Theory; AIT, Active Inference Theory.

* Correspondence to: Dept. Cognitive and Systems Neuroscience, Swammerdam Institute for Life Sciences, Faculty of Science, University of Amsterdam, Sciencepark 904, 1098 XH Amsterdam, the Netherlands.
E-mail address: c.m.a.pennartz@uva.nl.

https://doi.org/10.1016/j.bbr.2022.113969
Received 29 October 2021; Received in revised form 18 May 2022; Accepted 20 May 2022
Available online 16 June 2022

his eyes and starts talking, we interpret his body language and verbal signs and can thus draw inferences about the subject's experienced content – which is different from experiencing this content the way the subject does so directly.

This review will primarily target the question how consciously experienced content arises in the brain, although the concept of 'state' will come to pass when scrutinizing neural mechanisms for generating conscious experience. I will first describe what can be considered *'hallmarks'* of conscious experience – phenomenal properties that are core characteristics or inalienable features, at least of consciousness in healthy human subjects. This will lead us to a more precise *definition* of conscious experience, including its delineation from related cognitive abilities. Over the past forty years a combination of empirical and theoretical neuroscience, psychology, philosophy and computational neuroscience led me to formulate a theory dubbed Neurorepresentationalism, and the main tenets of this theory will be concisely summarized. Briefly, the philosophical position of Representationalism usually (but not necessarily) attempts to reconcile the existence of sensory qualities (e.g. the redness of a tomato) in conscious experience with materialism – the notion that brain matter constitutes an ultimate ground for conscious sensations [2].

These qualities traditionally pose a problem for many materialist theories of mind in the sense that one can ask where they are localized in physical space [2]. For instance, when we experience an illusion such as Kanizsa's triangle, it would be incorrect to state that closed triangular contours are physically on the paper in front of us. But it is equally incorrect to infer that the contours are literally present inside our brain (which contains neurons, glial cells, blood vessels etc.). So where is the quality of contour shape? Representationalism solves this dilemma by regarding sensory qualities as being components of representations, which can be veridical or unveridical. The contour is therefore a represented property of the represented object, which in this case is illusory.

Representationalism, as known from philosophy, is not a brain-mind theory in the sense of a comprehensive framework explaining how representations might arise from neural activity patterns [3]. Even though it has often been placed in the school of reductive, or eliminativist, materialism, it need not be reductive per se [2,4–6]. Indeed, the neurorepresentationalist view proposed here defends a non-reductive variant of representationalism, as it acknowledges the reality of qualitative aspects of phenomenal experience (as being unequal to properties of underlying neural substrates such as action potentials, molecules, cell membranes etc.). The adoption of the term 'neurorepresentationalism' to describe the theory is warranted nonetheless because it properly reflects that conscious experience can be understood as implying a particular kind of representation of the world and body. The term can thus be contrasted with opposing views, such as Direct (or Naive) Realism or Integrated Information Theory IIT (see below).

This general representationalist stance has been in search of potential neural substrates that may provide plausible mechanisms for generating the kind of representations we are conscious of. Neurorepresentationalism aims to fill the gap between the philosophical notion of (conscious) representation and its mechanistic underpinnings in the brain. I will briefly outline what Neurorepresentationalism is about and indicate how conscious representations differ from nonconscious information processing. Two cornerstones of the theory will be highlighted in particular: the *function* of consciousness in the light of biological evolution [1] (cf. [7]), and the principles by which basic computational operations of *predictive processing* (PP, see below) are postulated to be expandable to enable conscious representation. In passing, I will pay attention to the different *modes* of consciousness that can be distinguished: perception (conscious sensation, externally driven via sensor activations), imagery (internally driven experience, under considerable cognitive control) and dreaming (also internally driven, but with less cognitive control)[1].

Subsequently, we will consider how this theory offers an approach to tackle the Hard Problem, which has been set apart from 'Easy Problems' charting the neural mechanisms underlying tractable cognitive abilities such as working memory, attention or motor decisions. Even if these abilities can be well explained with conventional neural mechanisms, Chalmers has argued that we remain 'stuck' with the Hard Problem of explaining why all of these mechanisms would give rise to subjective experience [8]. Differences between Neurorepresentationalism and other neuroscientific theories of consciousness will be pointed out, focussing on Global Neuronal Workspace [9–11], Information Integration Theory [12–15] and Active Inference theories [16–19]. Finally, I will briefly review some of the implications of Neurorepresentationalism for our thinking about consciousness in patients suffering from disorders of consciousness (DoCs), animals and Artificial Intelligence (AI) devices.

2. Hallmarks of conscious experience

Which aspects or properties of conscious experience can be argued to constitute hallmarks of it? A description of hallmarks is considered significant because it will help us to circumscribe what we mean with terms like 'conscious experience'. Previously I have argued [1] that five hallmarks can be distinguished, constituting core properties of conscious experience in healthy human beings. That is, in healthy situations these core properties will always be present, and thus form an inalienable set of aspects. Without them, conscious experience would become a radically different phenomenon from what we commonly regard as such. I will refrain here from presenting a long list of other, potential properties; some of these will be considered below and can be better characterized as auxiliary properties, as they appear not to be inextricably linked to conscious experience in every modality or circumstance.

If we wake up in the early morning, we find ourselves in a particular situation. For example, we are lying in bed, see the sunlight shimmering through the curtains, hear birds singing and feel bedsheets sliding across our skin. This situation exemplifies the *multimodal richness* of conscious experience: it is endowed with sensations in several, qualitatively distinct modalities (like vision, audition and touch; Table 1). Within a

Table 1
Neurorepresentationalism considers five properties of phenomenal experience in healthy individuals to constitute inalienable hallmarks of consciousness.

Hallmark of conscious experience	Explanation
Multimodal richness	Conscious experience is qualitative in nature, i.e., it is characterized by sensations in multiple distinct modalities (vision, audition, somatosensation, smell, taste, vestibular sense). These main modalities can be partitioned into submodalities (e.g., for vision: color, texture, shape, motion, etc.).
Situatedness and immersion	In a conscious state we find ourselves situated in a space that is usually characterized by certain objects in the foreground and other stimuli in the background. Our body is experienced as immersed in the situation, occupying a central position relative to the surroundings.
Unity and integration	Consciousness is not made up of different elemental experiences, but is unified or integrated in that we have only one single experience at any given time. Our senses work together to enable the construction of an undivided, multimodal, spatially encompassing representation.
Dynamics and stability	Conscious experience is continuously updated following changes in the external environment and our body. Despite this dynamic aspect and ubiquitous movement of the head, eyes and other body parts, stationary objects in the environment are experienced as stable.
Intentionality	The property that a carrier substrate of consciousness can generate signals that are interpreted as, and refer to, something other than itself ('aboutness'). The brain's ability to interpret its own neural activity patterns not only pertains to ambiguous stimuli, illusions or hallucinations, but is considered a general and fundamental hallmark of consciousness.

single modality, such as vision, we experience a richness of submodalities like color, texture, shape and (visual) motion occurring in a spatial arrangement, experienced as happening before our eyes. This multimodal richness is considered a first hallmark of conscious experience. Philosophically, it is related to the notion of qualia, although this concept has come to be strongly connotated with qualitative 'standalone' or non-representational properties of objects, whereas multimodal richness is meant to be fully inclusive of the integration of multiple sensory qualities into represented scenes we perceive as objects against a background.

The example of waking up also appeals to the notion of space, but rather than relying on the Newtonian concept of an empty and abstract space, this second hallmark of conscious experience can be characterized as the property of being *situated*: in a conscious state we find ourselves in a situation typically filled with objects and other stimuli, discernible from the background and in a spatiotemporal context. Bodily sensations also partake in this situatedness, as we find our own body *immersed* in the situation. This means that there is no phenomenal distance between the subject and its environment: we are not 'looking at' our surroundings as if looking at a painting or other distal representation, but are situated in the middle of our subjectively experienced world. If we are looking at a painting in a museum, our conscious experience is not fully defined by the visual qualities of the painting, but also by the fact we assume a certain posture and position in space, as specified by proprioceptive, vestibular, tactile and other modalities.

In our daily life we need not pay attention to all sensory modalities at the same time: attention (in several of its incarnations, such as top-down and bottom-up attention) is considered a function separate from basic conscious experience (cf. [20]), and serves to amplify or intensify certain streams of sensory processing at the expense of others. Thus, multimodal richness characterizes conscious experience not because all kinds of modality are experienced simultaneously and continuously in the foreground, in the spotlight of attention; instead, many are contributing less conspicuously in the background. That they nevertheless contribute to conscious experience is underwritten by the realization that our conscious experience focused on the sunlight through the window of our bedroom would be quite different if we would not be lying in bed but standing upright, elsewhere in the room. To remove a potential source of confusion, multimodal richness does not equate with the popular notion of (Bayesian) cue integration, which refers to the brain's more basic ability to integrate sensory evidence on a stimulus and its properties from multiple sensory sources, relying on the notion of Bayesian inference [21–24]. For conscious experience, it is equally important to recognize the separation between, and distinctness of, modalities, a problem that has its historical roots in the Labeled-Lines hypothesis and the related problem of Modality Identification [1,25].

Next to the hallmarks of multimodal richness and situatedness-with-immersion, conscious experience is fundamentally characterized by *unity*. Visually, for instance, we do not experience two images originating from our eyes at once, but have an integrated, stereoscopic sensation. This unity is also apparent in the merging of sensations from different modalities into one integrated experience: when another person speaks to us, our brains merge sound information with visual information about lip movements, resulting in a crossmodal integration as long as the time lags between audition and vision remain limited [26]. Because 'integration' has been figuring as a key concept also in some other consciousness theories (e.g. IIT; [13]), I will not go into this hallmark here at length, but rather acknowledge that its inclusion in a set of hallmarks is appropriate: in conscious experience, different sensory input streams are unified into a single, situational experience. However, the hallmark of unity does not imply that the neural mechanisms underlying it would be unitary as well. Most likely, there are different neural mechanisms at work in our brain to mediate different aspects of integration (monocular integration for stereoscopic vision, multisensory integration, feature binding, etcetera). Only the result of all of these operations is experienced as being integrated or unitary.

Some theories of consciousness have also proposed perspectivalness as a property of conscious experience [27–30]. Indeed, visual experience is bound to a first-person perspective or point of view, depending on the orientation of our eyes and head relative to their surroundings. However, when we consider conscious experience as made up by sensations in many modalities, the notion of a singular perspective becomes less strong. While visual and auditory sensations might be argued to be set in a perspective originating from an 'egocenter' in the subject's brain [27], the existence and location of such a center (a 'cyclopean eye' located frontally in the brain and along the midline) are much less obvious for modalities such as somatosensation, taste and the vestibular sense. If anything, the focal point of sensation seems to be centered on the average location of the sensors being stimulated, such as the tongue or oral cavity in the case of gustation. Because of this failure in generalizing perspectivalness to a broad range of senses, and the accompanying difficulties in denominating a common, physical 'egocenter' across all modalities, I refrain here from classifying first-person perspective as a necessary hallmark of conscious experience – as opposed to unity or integration, which is considered here to constitute the third hallmark.

Fourth, conscious experience is characterized by a combination of *dynamics* and *stability*. 'Dynamics' refers to the phenomenon that our experience is continuously updated following changes in body posture, head orientation, eye positions, etcetera, as well as changes in the external environment surrounding us. In contrast, 'stability' refers to the steady positioning of objects in the world around us, as it appears to us in the face of the multifarious movements our body produces, affecting a panoply of sensors in our body, whether or not contributing to conscious experience. We perceive a vase with flowers as having a stable position on the table that is close to us, even if our eyes and head are moving, and even if we walk around that table. Apparently, neural systems contributing to conscious experience have the ability to take into account, and correct for, body and sensor motion while they represent external objects as being invariant and stable. And yet, simultaneously our experience is dynamically and smoothly adjusted as changes in the external environment, our visual viewpoint, body position and orientation are incorporated in it.

A fifth hallmark is constituted by the concept of *intentionality*, in other words, the brain's ability to interpret neural activity patterns as something else than themselves – something that is moreover represented as being outside the brain. Phenomenologically, this interpretability is illustrated by ambiguous pictures such as the Necker Cube (as an example of perceptual ambiguity) or Jastrow's duck-rabbit (exemplifying a cognitive or semantic ambiguity given the same sensory input). However, intentionality stretches far beyond such individual cases, including the multi-interpretability inherent to illusions, the non-veridical nature of hallucinations (in the face of a subject feeling that his non-veridical experience is 'real'), and Gestalt compositions from which objects can suddenly 'pop out' and recognized semantically (e.g., a Dalmatian dog). In representationalist terms, the illusory feature is also referred to as a non-actual intentional content, whereas a real triangle in front of our eyes will have actual intentional content [2]. Arguably, intentionality is such a widely encompassing, fundamental hallmark of consciousness that a theory failing to address it cannot be classified as a complete theory of consciousness. Importantly, intentionality is equally at stake in unambiguous and ambiguous percepts. This implies that the intentional, world-modeling machinery of the brain is constantly at work for as long as we are awake or dreaming, even when we receive sensory inputs that do not rival for different interpretations.

The five hallmarks listed above constitute essential core properties of conscious experience in healthy human beings, but it is of note that pathological cases may present deviations from this listing. Consciousness continues to persist in patients with brain damage affecting specific modalities such vision, and for instance blindsight type II [31] results in basic sensations of 'sensing the presence' of something that cannot be seen. Thus, blindsight type II arguably presents us with an exception to the core property of situatedness and immersion, even though the patient will experience situatedness and spatiality by way of other, intact modalities (proprioception, vestibular sense, sound, etc.).

This diminished form of (visual) consciousness is rare and should not distract us from the phenomenology reported by healthy people. For the sake of completeness, it may be added that the notion of 'self' (and self-awareness) is not irrelevant to the present discussion, but can be regarded as a more highly developed form of consciousness (or meta-cognition) than the concept of consciousness as sensorily rich experience, presented here.

3. Neurorepresentationalism: biological function and definition of consciousness

The hallmarks raised above characterize conscious experience as a multimodally rich, integrated phenomenon, having the subject in the midst of a spatiotemporal situation that contains stably positioned objects, whilst at the same time accommodating dynamic bodily and externally induced changes. Herein intentionality acts as an overarching concept, enabling us to perceive and semantically interpret the myriad of sensory inputs in a specified and quasi-stable manner. Yet, before attempting to define consciousness more specifically, we must consider its potential relationships to various kinds of body movement. Here, it is crucial to distinguish fast and automated actions from more slowly prepared and deliberate actions. Reflexes serve as examples of fast, automated actions par excellence, as their execution does not generally require an involvement of brain systems involved in consciousness (Fig. 1). For instance, the knee-jerk reflex is mediated by a reflex arc running from stretch receptors in the patellar tendon, via dorsal root sensory neurons and motor neurons in the spinal cord, towards antagonistic flexor and extensor muscles in the thigh. Although higher brain centers can exert modulatory control over this reflex, it is usually executed automatically, as illustrated by the fact that this reflex is utilized almost continuously during regular walking, without the subject being conscious of it. We tend to become conscious of the stimuli and sensory consequences of reflexes *after* they have been executed – such as becoming aware of the fly that approaches our face after reflexively closing our eyelids– but this leaves the argument of reflexes operating without awareness intact. In adults, the acquisition of novel stimulus-response associations generally takes place with some level of awareness of both stimuli and motor responses, but when numerous repetitions lead to automatization of behavior (also known as habit formation [32]), people become much less aware of these simple stimulus-response behaviors, making place for percepts or imagery of other items that draw our interest and attention. Thus, both reflexes and habitual stimulus-response behaviors are generally not or weakly coupled to conscious experience.

This weak coupling stands in contrast to goal-directed behavior (GDB), which is defined as planned behavior based on the subject sustaining a representation of its goal before and during execution of the action pattern [33–36]. Although the concept of GDB has been operationalized by the encoding of action-outcome relationships and of knowledge of the action being causally required to obtain a desired outcome [36], it is closely akin to the concept of model-based learning [34,37]. In contrast to model-free learning, which corresponds to classical Reinforcement Learning (i.e., with a tracking and caching of reward value of stimuli and actions, and thereby tightly connected to habit formation), model-based learning is defined by the subject acquiring a causal model of its environment, including the causal relationships between sensory-specific stimuli and rewards, thus going beyond the memorization of the reward value of stimuli and actions alone. Because neural systems for GDB allow for complex planning of subsequent actions in time, lined up towards reaching an end goal beyond the subject's immediate temporal horizon, they have a greater temporal depth than reflexes or simple habits, and enable the subject to deliberate and improvise based on the subject's current situation.

This distinction, viz. between (i) reflexes and habits, and (ii) planned, goal-directed behavior, leads us to delineate a biological *function* of consciousness: conscious experience provides the subject with a *best-guess survey* (or perceptual 'summary') of its current situation, including its own body parameters relative to its environment [1]. From the need to generate a comprehensive survey, it follows that it must leverage all modalities relevant to subsequent complex decision-making, and must be spatially encompassing in order to permit inspection of all environmental items that may affect subsequent actions. The deeper temporal horizon associated with planned behavior confronts the subject with a potential problem: which environmental or body parameters, events and objects become important for future action sequences is often unknown from one's current vantage point. Hence, from a teleological-evolutionary viewpoint it is understandable that conscious experience is spatially 'wide' and encompasses not only the object of immediate attention, but also items making up the contextual background, at least in the form of 'summary statistics' [38].

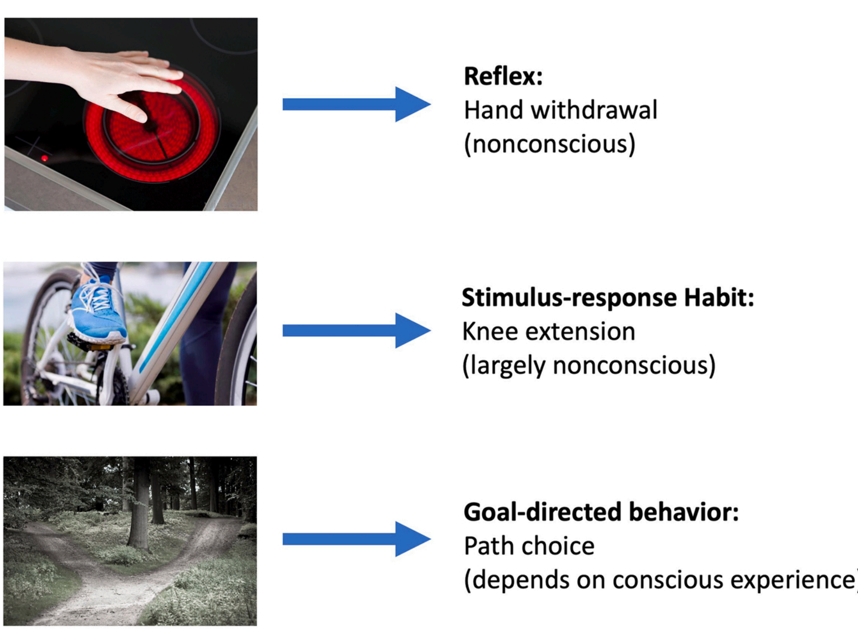

Fig. 1. Three types of behavior with different degrees of association with conscious experience. Upper panel: for executing a withdrawal reflex when the hand is exposed to excessive heat, no conscious processing is postulated to be necessary; conscious processing occurs mostly after withdrawal and is not needed for reflex execution. Middle panel: after prolonged training on bicycle riding, execution of knee extension or flexion movements becomes habitual and occurs largely outside awareness. Both reflexes and habits can be under voluntary control, but this is not postulated to be required for movement execution. Bottom panel: when a subject, facing a novel Y-maze junction, has to make a deliberate choice, goal-directed behavior comes into play. It is based on goal representation, knowledge of action-outcome relationships and action causality. In animals and humans, this planned behavior is postulated to depend on conscious experience, defined as rich multimodal survey of the subject's environmental situation and body state.

Computationally, providing a near-continuous and dynamic update of one's current multimodal, situational survey is heavier than is required for the automated execution of reflexes and habits, which can be readily activated from procedural memory [39]. That this is associated with longer reaction times than is the case for reflexes and habits – which renders consciousness particularly useful on a medium-to-fast time scale - is not considered a critical disadvantage, because the computations involving internal deliberations and planning also take considerable time (cf.[40]), and the long-term consequences in terms of organismal survival and reproduction - of GDB relative to reflexes and habits - weigh heavily as well.

Several important qualifications are in place when defining the function of conscious experience as done here. First, ascribing a function to consciousness in a teleological, evolutionary sense – while acknowledging the existence of qualitative properties of phenomenal experience – must not be confused with the philosophical position of (classic) functionalism. As originally proposed by Armstrong [41] and Lewis [42]; (see also [43]), functionalism posits that what we call 'mind' acts as a causal intermediate (or "occupant of a causal role" [42]) between input into a system and (behavioral) output generated from that system – a view akin to behaviorism. Classic functionalism rejects the concept of mental states as subjective feelings or qualitative experiences, but views them as states that are triggered by inputs and exert particular causal functions in our behavior. In contrast, Neurorepresentationalism puts subjective feelings or qualitative experiences at the center stage of consciousness, while recognizing that these can *simultaneously* subserve – but are different from – functions such as planning and execution of goal-directed behavior. In short, Neurorepresentationalism proposes that the subject benefits from its qualitatively rich, situational survey because this survey is informative about the state of the world and body in view of possibilities for subsequent action.

Second, conscious experience does not *equate* with planned decision-making or GDB. We keep on having conscious experiences when disengaging from planned behaviors, such as during daydreaming or simply staring out of the window. Such action-disengaged experience may serve to allow more free, associative or counterfactual deliberations, or may serve to anticipate on improvised behavior potentially arising in case of surprising stimuli, although this is more speculative. The key point here is that, even though we often associate the term 'conscious' with 'decision-making', 'thinking' and 'deliberation', the core function of consciousness is to provide a multimodal, situational survey that subserves subsequent decision-making [35]. This function is logically associated with a definition of conscious experience (which was burgeoning already): it is the multimodally rich, dynamic survey of the subject's current situation, including his own body and functionally earmarked for planned behavioral and cognitive actions in the future.

A third qualification is that planning and deliberation do not depend on having an update on one's current situation alone. Episodic knowledge of the immediate past, the semantic meaning of items in the environment, knowledge of their practical use, positive or negative reinforcement value, etcetera, are required for model-based planning as well [35,37]. Thus, it is mandatory that knowledge can be rapidly retrieved from declarative memory when facing a particular situation in which a complex, far-reaching decision is required. Previously it was proposed that systems for perception and declarative memory engage in bilateral interactions, with perceptual systems triggering semantic and episodic recall of information associated with the object of interest, whereas declarative memory recall supplements and enriches percepts with meaning, value and episodic context [35,44–46].

One may object to this definition of consciousness that the actual work that needs to be done providing decision-making systems in the brain is done by neurons, not separately by some sort of high-level process dubbed 'consciousness'. This reflects a misunderstanding of neurorepresentational theory, arising from either dualistic or eliminative-materialistic background assumptions. Whereas classic dualism maintains a conceptual and causal distinction between a (conscious) mind and the brain, the stance adopted here is that the 'hard work' done by the relevant neuronal networks actually *corresponds* to conscious experience – not as a causal result of neuronal operations that *precede* consciousness in time, but corresponds to it immediately, with conscious experience being an emergent and supervenient phenomenon – without exerting any kind of top-down influence (cf. [47]). In contrast to eliminative materialism or related forms of strong reductionism [43,48–51], Neurorepresentationalism acknowledges phenomenal experience as real and existing, on an equal ontological footing with the electrophysiological mechanisms ultimately underlying it (see also below). Thus, consciousness is not an epiphenomenon relative to the functional operations carried out by neurons transforming sensory inputs into motor outputs. Rather, at least for higher vertebrates, it is postulated that consciousness *is* a main, high-level function of the underlying neuronal networks. In other words, the job to be done by the neuronal networks is to realize phenomenal experience – not to merely act as occupant of a role causally transforming sensory input into behavioral output, but to *generate* conscious experience to the benefit of the subject's subsequent action repertoire. That these systems also fulfill other functions during wakefulness or sleep – such as memory consolidation – does not contradict that one of the main functions of these networks is to sustain consciousness. In fact, paradigms like predicting processing suggest that perception, learning and memory at the cortical level are inextricably intertwined (see below). Conscious experience helps the organism survive and reproduce, owing to the grand multimodal overview it provides for subsequent complex planning and actions.

4. Neural mechanisms underlying conscious experience

If we accept the definition of conscious experience as a multimodally rich, situational survey, how does it arise in relation to brain activity? Various schools in the philosophy of mind have dealt with this issue in very different ways. According to direct realism, for instance, it makes no sense to say that we would be perceiving, or looking at, representations, or 'sense impressions' [52,53] (cf. [35]). Direct realism holds that we do not see the world through a 'veil' of perception [54], but perceive physical objects directly. Akin to this, the sensorimotor contingency theory of O'Regan and Noë [55] and related enactivist accounts posit that consciousness exists by virtue of our bodily interactions with the world, conceiving the brain as a device for coupling sensory input to behavioral output. Qualitative phenomena we experience would arise from the rules by which motor actions govern changes in our sensory apparatus; seeing color is thus explained from knowing and mastering the pattern of sensory changes that will occur when one moves the eyes towards an object given the pattern of impinging sunlight. Despite the attention this line of thinking draws towards the long-neglected role of embodiment in consciousness, the account faces serious objections. To summarize these, it has been argued [1,35] that (i) conscious experience persists in the absence of overt motor activity (e.g. during paralysis, locked-in syndrome, rapid eye movement (REM) sleep and imagery) and (ii) qualitative experience can occur in a non-motor related manner (e.g. synesthesia), (iii) sensorimotor contingencies can be exerted through real-world interactions in the (very likely) absence of consciousness (e.g. in epileptic automatisms), and (iv) conscious experience can pertain to non-veridical objects or object properties (e.g. in illusions and hallucinations) the subject cannot physically interact with, because they are non-existent. One way of coping with these objections is to attempt to 'internalize' action in that, for instance, mental acts of motor planning, imagery and top-down attention – relying on one's life history of sensorimotor learning - are taking place to enable conscious experience [17,55,56]. Space does not permit us to review all arguments raised against this 'internalized' variant of sensorimotor contingency theory, except to mention the phenomenal feelings of amelia patients in limbs they never possessed (i.e., amelia is the condition of being born without arms and legs; [57]) and to recall empirical arguments for a dissociation between top-down (or endogenous) attention and consciousness [1,20].

A crucial argument against Direct Realism has yet to be mentioned: no matter how perception exactly relates to brain activity, it is not feasible to elicit conscious percepts of the environment while bypassing peripheral sensors and connected afferent nerves (except by artificial means such as transcranial magnetic stimulation, TMS). Thus, one must account for the hard-wired delays in transmitting signals from the peripheral sensors to the brain, even though we have the feeling that we perceive events in the world, and our interactions with it, as happening 'directly'. Moreover, the generally accepted validity of the labeled-lines hypothesis, attributed to Müller [58], has misled some students of the brain to think that sensory modalities are identified peripherally (i.e. by the nature of the corresponding sensors and their connected nerve fibers). The labeled-lines hypothesis does correctly posit that the nature of sensory modalities is associated with the nature of the peripheral sensor and its nerve fibers reaching the central nervous system, but defers the issue of *modality identification* to the brain. Briefly, adopting the brain's viewpoint, its receptive cortical areas face the problem that they receive uniformly shaped action potentials via all afferent cranial nerves and ascending spinal cord pathways [25]. Given that the brain has no intrinsic anatomic knowledge of where its sensory afferents come from, it is tasked with drawing inferences about what causes the patterns of sensory input to arrive in the brain the way they do [25,59] – a conclusion already implicit in Helmholtz' work [60]. This provides another, compelling motivation to explain conscious experience in terms of neural representations, with an important further qualification.

Proponents of enactivist and sensorimotor accounts have correctly pointed out that conscious experience cannot consist of 'pictures in the brain' the subject would be looking at (e.g. [55]). Not only would this view mislead us into a kind of dualism in which we are resorting to an internal homunculus or spectator, it also belies our perceptual experience as pertaining directly to objects and events occurring outside our brains, and predominantly outside our body. However, (neuro)representationalism is by no means committed to postulating 'pictures in the brain'. Rather, the key distinction between the 'vehicle' and 'content' of representation (cf. [61]) leads to the clarification that the 'vehicles' (neural coding mechanisms, or carriers of information) are located inside the brain (specifically, the lower and higher sensory corticothalamic areas involved in generating a multimodal, environmental survey), whereas the 'content' is experienced as happening outside the brain. It seems as if brain systems for conscious representation thus have the capacity to 'project' coded content into our body and/or the external world, but a more accurate account holds that there is no such projection (in the sense of information being transported from a site of origin to an external destination): the perception of our body and the external world are the result of the representational activity of the brain. Thus, perceiving things as being in the world is a consequence of the intentionality inherent to our representational apparatus, situating, for instance, vegetables that we see not inside the brain, but outside. Similarly, illusory objects or object properties arise from representational mechanisms in the brain attempting to interpret ambiguous information, but are not literally present as images-in-the-brain. They are perceived outside our body and are represented as objects located in our external environment. Importantly, perceptual content resulting from internal modeling activity is delimited by the sensory evidence on offer, meaning that we do not see a full three-dimensional model of a car we pass by on the street, but just the in-depth representation of the surface of the car.

The consequence is that we are not experiencing our external environment directly, but experience a 'virtual reality' (cf. [29,62]). Nonetheless, this experience feels real and direct to us because, at least in healthy people, the 3-dimensional, 'holographic' representation forged by the brain does in fact allow us to navigate through and act in the world efficiently, resulting in satisfaction of our basic homeostatic needs for food, drinks, sex, shelter, etcetera. That this 'virtual reality' simulacrum can go awry is illustrated by brain disorders such as Anton's and Charles Bonnet syndrome as well as schizophrenia. In Anton's syndrome [63], for instance, patients suffering from blindness due to occipital cortical damage sustain visual hallucinations while denying that they are blind. Here, the misrepresentations synthesized by a partially functioning system can strongly hamper effective navigation and goal-directed action to the extent that the life of a patient is endangered when acting purely on his own.

A computational mechanism offering a primordial basis for the inferential construction of conscious representations is predictive processing (PP; understood as a wider concept than predictive coding; [64,65], cf. [66,67]). As a theoretical framework, predictive processing can be traced back to Helmholtz' work [68] which already characterized the brain as continuously generating an internal model of the environment. The model constantly generates predictions (or representations, inferences, hypotheses) on the causes of sensory input, and particular units in the model serve to compare these predictions with actual sensory input, resulting in the computation of prediction errors that are used to draw better inferences and guide learning within the same model [65,69]. In Rao and Ballard's classic model [69], PP is implemented by a 'low' and 'high' layer of neurons (symbolizing primary visual cortex and a higher visual area), in which the high layer generates predictions about the sensory inputs arriving in the low layer. When sensory input enters this low layer, neural circuits compute the difference between this input and the high-level prediction (alternatively labeled 'representation'), yielding an error signal that is subsequently transmitted to the high layer to infer a better prediction, and to induce synaptic modifications by which the network learns to improve its predictions. These and similar configurations have successfully simulated physiological properties of visual cortical neurons such as end-stopping in V1 [69], orientation selectivity [70] and mismatch negativity [71–73]. More recently, Dora et al. [74] expanded the prior two-layered architecture to four or more layers and scaled up neuron and synapse counts (Fig. 2), showing that higher model layers show increased stimulus selectivity, in agreement with electrophysiological recordings from high-level visual areas such as inferotemporal cortex [75]. Moreover, representational sparseness increased when ascending the model hierarchy, with a strong dependence on the strength of regularization factors (which generally result in a suppression of neural activity). In this model, the updating of neuronal activity values and synaptic weights followed neurobiological principles, as illustrated by a locally operating, Hebbian learning rule including a gating factor. Another recent development in PP has been to design a multisensory predictive architecture in which visual and tactile representations interact to generate cross-modal predictions. Real-world robots navigating a space with obstacles have been shown to achieve better place recognition using this MultiPredNet architecture as compared to state-of-the-art AI tools (i.e., variational autoencoders) [76]. These and other efforts to modify and expand PP models underscore their functional versatility whilst operating in the domain of neurobiologically plausible computation.

Despite the advances demonstrated by computational models of PP, some fundamental questions on its relevance for consciousness remain. Although physiological evidence for neural coding of prediction errors and predictive representations is mounting [72,77–80], it is largely unknown how PP may be implemented in neural circuitry. Various computational schemes in line with Fig. 2 have been proposed (e.g. [72,78]), but it has been argued, for instance, that - contradictory to most PP schemes - feedforward projections from superficial layers in low visual areas to higher areas transmit more information than merely errors (e.g. [81–84]. Alternative configurations have been proposed, for instance placing the computation of errors in higher areas and that of predictions in lower ones [85]). Alternatively, each visual cortical area – whether low or high, but with its own purview or specialization in spatiotemporal feature processing – can be conceptualized to compute both errors and predictions, and pass these on to other cortical areas, bidirectionally for both types of quantity (cf. [77]). A further question on the neural implementation of PP is how brain systems may arbitrate between

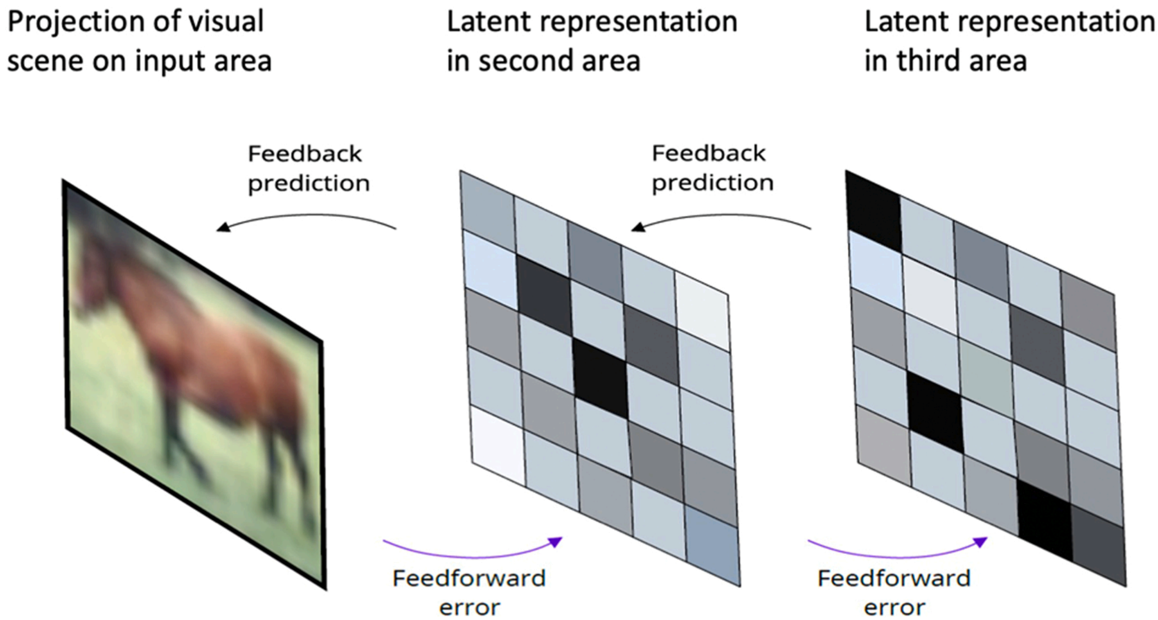

Fig. 2. Construction of latent perceptual representations by predictive processing in scalable, multi-layer networks. On the left, a visual scene including a horse is processed by the retina, the result of which is projected onto an input area of a modeled visual cortical hierarchy. The transmission of signals from this input area to a second-level area depends on the comparison between the actual sensory input and a prediction emitted by this second-level area (cf. [69]). The subtraction operation results in an error (or surprise) signal, which is fed forward to the second area, higher in the cortical hierarchy. This area produces a representation that is considered 'latent' because an explicit depiction of the horse is no longer visible at this level. Nonetheless, in a trained network the representation will be passed in a top-down (feedback or recurrent) direction and will result in a prediction and input-image reconstruction (i.e., auto-encoding) once the feedback synapses to the input area have been passed. Output from the second area is transmitted as error in a feed-forward direction to a third area, which in turn sends predictions as feedback to the second level. Errors are utilized to improve both within-trial inference and across-trial learning. The architecture is scalable and can be expanded with more layers [74]. The output of higher layers can be used to classify different groups of stimuli, serving as a basis for semantic categorization and semantic memory (cf. [1]). In Neurorepresentationalism, predictive processing is considered a useful basis for developing computational approaches to consciousness, but no claim is made that the model as presented here produces conscious experience; essential properties of multimodal, situational survey are still lacking.
Figure adapted from: [147].

different representations pertaining to the same scene or object, such as in the case of ambiguous images, illusions or binocular rivalry. Conceivably, corticothalamic systems may generate competitive representations that rival with one another for dominance before a representational equilibrium and 'choice' is established. One possibility for implementing such a competition is that neural systems coding a particular representation-error loop possess attractor properties that may 'win out' over other, less powerful attractors coding alternative hypotheses ('winner-take all' principle; [25,86,87]). Another possibility holds that the output of the corticothalamic systems involved is collectively transmitted to other areas, such as frontal areas and/or the basal ganglia, for selection of a perceptually dominant best-guess representation. More empirical and computational studies will be required to elucidate this issue of competing representations.

5. Multi-level representations and the hard problem

An even more fundamental question on PP, however, is why neural activity constituting this type of coding would give rise to consciousness in the first place. This directly confronts us with the Explanatory Gap [88] or the Hard Problem [8]. The Hard Problem revolves around the question why any sensory, motor, memory and other cognitive processing in the brain should have to be accompanied by consciousness, while it is seems conceivable that all of this processing may as well proceed without consciousness or subjectivity. In addition, an alternative problem (the Real Problem of consciousness) has been phrased: how can the various properties of *specific* conscious experiences be explained from physical processes in the brain, sensory organs and other body parts? Thus, the Real Problem does not attempt to explain why and how there is consciousness in the universe in the first place, but seeks to explain why a particular experience is the way it is in terms of underlying brain-body processes [89]. Below, we will examine briefly what Neurorepresentationalism has to say about both problems.

A key issue to consider is why purely numerical representations (e.g., firing-rate values) coded in lower or higher areas of PP models would empower the overall network to experience any qualitative, non-numerical properties. This is, inter alia, a problem not only for PP, but for all computational models making any claim about consciousness. Secondly, predictions are thought to be generated by many brain structures and multi-area systems – not only in the realm of corticothalamic processing, but also in structures where lesions do not cause any clear-cut impairments of consciousness (e.g., cerebellum; suprachiasmatic nucleus and other hypothalamic and mesencephalic nuclei [1,90–92]).

Somehow, it has remained tempting to think that the result of modality-specific computations must be 'read out' by some higher brain area to render it conscious. Yet, this idea is argued to be misguided because there is no single, homunculoid brain structure that is equipped to do the 'reading', 'interpretation' or 'translation' into a conscious percept. This can be inferred because any brain area, interconnected with others in a serially connected configuration, consists of neurons that generate firing patterns in just same way as the foregoing, afferent structures do, and a single neuron will not have the representational power to code anything as complex and rich as a conscious percept. An important clue to get past this stage of 'serial thinking' is provided by the structure and function of the cerebellum. Structurally, the cerebellum does not possess the multi-area hierarchical structure as has been identified for several sensory modalities in the neocortex [1]. Functionally, the cerebellum is understood to subserve the gain control and plastic adjustment of fast-operating sensorimotor loops with limited parametric scope, such as the vestibulo-ocular reflex [93]. Both the fast time scale and relative simplicity of the sensorimotor processes regulated by the cerebellum differ from the function I have proposed above for conscious brain systems: to subserve complex decision-making and planning on a slower, intermediate-to-fast time scale by way of providing a multimodal, *situational survey* of the subject's own situation.

The conclusion from this cerebellar-neocortical comparison can be phrased as follows. Instead of serially adding more and more way stations attempting to explain how the 'stage' or 'domain' of consciousness may be reached according to a 'horizontal' (serial) conceptualization, it is more productive to think of consciousness as arising emergently in a 'vertical' (multi-level) conceptualization. That is, the kinds of predictions the cerebellum may mediate are 'small' and spatiotemporally restricted in the sense that they pertain to a limited range of sensorimotor parameters, whereas conscious representations are 'big' in the sense that they code many features simultaneously: in 3-dimensional space, in a multiplicity of (sub)modalities and cast in a temporally deep framework for subsequent use in planning and GDB (cf. [94]). Herein the 'vertical' conceptualization somewhat resembles David Marr's notion of multiple levels of nervous system function [95](the level of hardware implementation; the level of algorithm and representation, and the level of computational theory, defining the goals and strategy of the model; note, however, that Marr was not concerned with consciousness as a subject of study).

More concretely, multi-level representations underlying conscious experience have been proposed to be constructed bottom-up by the level of: (i) single neurons, having the capacity to respond to single features; (ii) ensembles of neurons, forming small, within-area local networks capable of pattern coding within a single submodality (e.g. shape); (iii) unimodal metanetworks, which combine the hypotheses from lower-order ensembles into representations of objects considered within a single modality (e.g. all visual features making up a visual object) and (iv) multimodal metanetworks, integrating the information coded by unimodal metanetworks into multisensory object representations (Fig. 3)[1,96]. Different ensembles and metanetworks covering different locations of the environment and body must simultaneously collaborate to provide a spatial, situational survey in which both peripheral and central portions are registered. Given this spatial and multimodal integration, phenomenal experience is postulated to arise at this highest level of representation [1]. How integrative processes synthesizing high-level representations exactly operate is a subject for empirical, computational and theoretical investigation, although recent neural-network modeling studies and the empirical support for cue integration and cue separation may provide indications how to proceed [21,66,74,97–100].

To avoid confusion, the multi-level concept deployed here (Fig. 3) is a different one than the anatomical notion of 'low' and 'high' levels in the cortex, such as in the visual system. For instance, Jackendoff [101] argued that conscious experience would predominantly correlate with neural activity in intermediate levels of the (anatomical) visual hierarchy. In contrast, Neurorepresentationalism deploys its multi-level concept such that *both* low (e.g., V1) and high (e.g., inferotemporal) cortical areas can contribute to visual consciousness. For instance, inferotemporal areas of the non-human primate brain, containing neurons with very wide receptive fields, likely contribute position-invariant shape information to a perceptual representation, whereas V1 neurons, having small receptive fields, may contribute local visual details at a high resolution. Thus, conscious experience is considered the (phenomenological) result of a process of 'superinference' – a gigantic 'supercomputation' comprising both low- and high-level representations, which simultaneously integrate across different modalities. Similarly, intermediate areas such as V2, V4, MT/V5 contribute to this

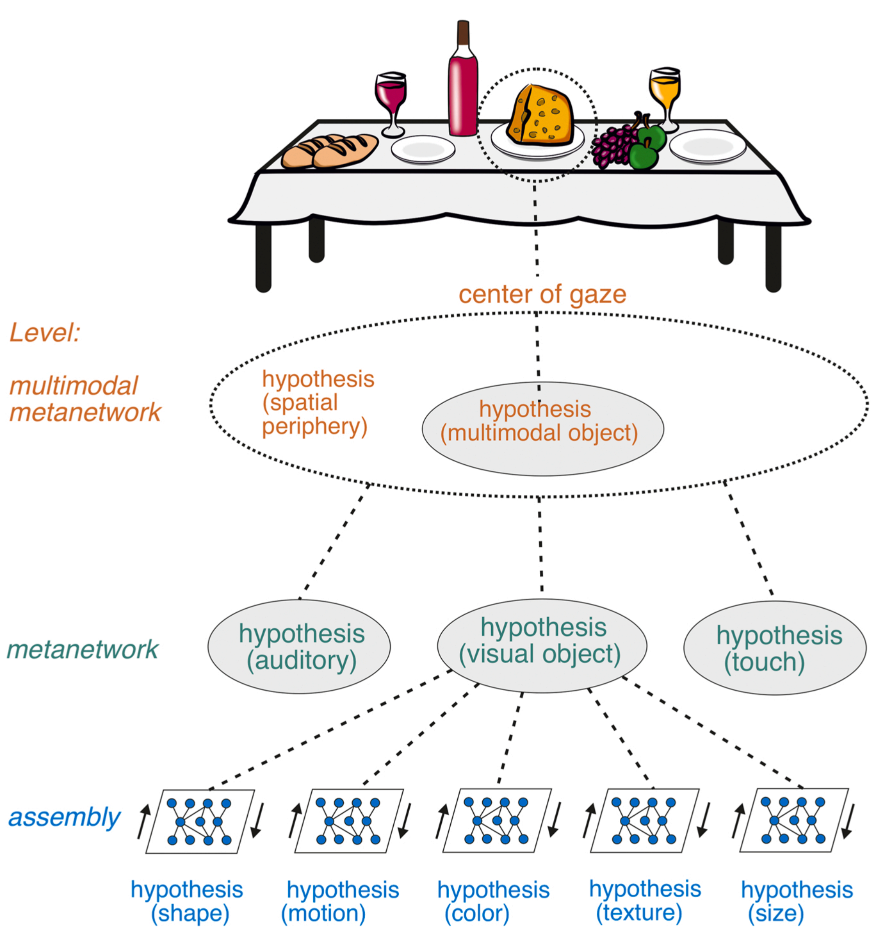

Fig. 3. Functional organization of different levels of representation in the construction of conscious experience. At the lowest levels of representation, single neurons are connected into local cell assemblies coding single-feature hypotheses by way of predicting-coding operations (e.g., on shape or motion in a particular subregion of the visual field; up- and downward arrows symbolize feedforward and feedback communication). These assembly-level hypotheses are subsumed under a larger (metanetwork-level) hypothesis on a visual object to which these individual features are attributed. These unimodal, metanetwork-level hypotheses are subsumed under a single sensory modality (e.g., vision), but are integrated into object representations specified in various relevant modalities (e.g. vision, audition, touch) at the multimodal metanetwork level. Simultaneously, information about other objects and background in the visual periphery is integrated at this level. Phenomenal experience is postulated to correspond to this highest level. Representations unfold simultaneously and emergently at different levels, and no causal-chain interactions take place between levels. Importantly, the multi-level representational scheme presented here is different from the notion of 'low' versus 'high' levels within a modality-specific hierarchy, such as V1 versus inferotemporal cortex in the visual cortical system; neither is it necessarily equivalent to the notions of aggregate levels or aggregate sizes in the central nervous system, because it may be implemented differently in e.g. invertebrate nervous systems or artificial devices.
Adapted from [1].

process with area-specific coding of visual attributes.

The multi-level concept implies that conscious representations arise *emergently* from low-level neuronal configurations. This emergence is already envisioned to take place at the level of ensembles or small, local networks, as illustrated by the Hopfield model of auto-associative memory [102]. Thus, there is no serial-causal relationship in which consciousness temporally *follows upon* neural activity. Representations at multiple levels arise simultaneously, although it remains the case that high-level, conscious representation depends on low-level neuronal activity (whereas the opposite is not true: neurons can exist without high-level, conscious representation). This type of relationship has been characterized as one of *emergent correspondence* [1]. This concept implies that both the micro-level of neurons, transmitters and spikes, and the level of phenomenal experience are existing and real, which characterizes Neurorepresentationalism as a non-reductionistic form of materialism.

Importantly, the highest, phenomenal level of multimodal, situational survey is devoid of, and has no need to invoke, a 'spectator' or 'self' who does the watching. Phenomenal experience just 'is'. The self and the related concept of self-awareness arise as a form of metacognition serving to explain the way we associate our diverse, subjective experiences with our body, our senses, our actions and our external, social world [1] (cf.[43,103]). A crucial aspect of the concept of emergent correspondence is that it avoids any form of mental top-down causation. As indicated by Kim [47], a type of emergentism where mental activity exerts a causal influence on physical brain activity would run into the problem that the physical activity would also be governed by other factors in the physical domain, forcing one to suppose a dual causality, leading to overdetermination and a violation of the causal closure of the physical domain. Following Fig. 3, consciously experienced sensations manifested at a high representational level automatically correspond to low-level processes, obviating the need to suppose a 'mental influencing' of neuronal operations.

To further clarify the concept of multi-level representation, it may be useful to invoke some bad and good metaphors. First, a metaphor like the distinction between hardware ('low level') and software ('high level') would be misleading, because software comes down to the set of programmed instructions needed for computer operations – instructions which are by themselves devoid of experiential content. A better metaphor is given by a major newspaper like the New York Times, which can be said to exist across several different levels (a physical edition of paper and ink; a digital version stored on electromagnetic hard disks; the group of editors writing the contents of articles, and the newspaper considered at a conceptual, cultural level, regardless of the persons involved). A caveat accompanying this metaphor is that it presupposes the presence of an independent reader which, as noted, is absent within the conscious brain. A better metaphor may therefore be Gödel's proof of his first incompleteness theorem [104], which is marked by information coded both at the 'low' levels of numbers and the 'higher' level of typographic symbols. This higher level symbolizes that at least one statement in the logical system under scrutiny remains undecidable – which remains true even in the absence of an external observer [105].

The multi-level concept of consciousness offers a useful approach to tackle the Hard Problem – which is not to say that it easily 'reveals' a tangible, visible solution to it. If we confront ourselves with the stark gap between action potentials, neurotransmitter and neurons with phospholipid membranes on the one hand, and subjective experience on the other hand, seeing a 'direct connection' between these seems an insurmountable problem. According to Neurorepresentationalism it is indeed impossible to *visualize* this transitioning from neurons to conscious experience, as the intermediate levels of representation are not accessible to subjective experience. A potential critique of this approach holds a reasoning like the following: 'I cannot imagine that seeing a red apple is actually based on the firing patterns of all those millions of neurons in my brain – for me, these two phenomena remain different and stand apart. The Hard Problem remains'.

This counterargument precisely reflects the crux of the Hard Problem: it originates from the limitations of our own imagination, as our mental imagery has been thoroughly shaped by our previous phenomenal experiences, without us having access to lower levels of representation. In our autobiographical history, our neural systems supporting imagination have been trained on the basis of prior conscious experience, and so we never learned to 'look under the hood' of our imagination, to see the connection or transitioning from neurons, via larger and larger networks, to consciousness [1]. As compared to some major theories in physics – such as quantum mechanics or the Big Bang theory – this situation is not particularly exceptional, because we also cannot literally imagine clouds of uncertainty pertaining to the position or momentum of an elementary particle, or a Big Bang from which not only matter, but also space and time originated. Returning to Chalmers' original argument on the Hard Problem, he referred to gamma oscillations in cortical systems, which around 1995 were still regarded as a potential neural correlate of consciousness. Chalmers pointed out that these oscillations fail to clarify why they should give rise to conscious experience. According to Neurorepresentationalism, however, it is impossible to see (or imagine) a direct connection between gamma oscillations and consciousness in the first place, because gamma oscillations are low-level neural phenomena, and if they would be essential to consciousness (as is probably not the case [1,106]), their correspondence to phenomenal experience could only be understood by abstract reasoning along intermediate levels of representation.

Despite this approach to tackle the Hard Problem, important questions on multi-level representations remain to be answered. What, for instance, constitutes a neural correlate of a multi-level representation coded in corticothalamic systems? What would be the smallest, minimal neural substrate capable of sustaining a simple, yet unitary conscious experience? To begin with the latter question, this can be referred to as the 'consciousness-in-a-bottle' problem (cf. [107]). Let us assume that an isolated, but otherwise intact, neural circuit is constructed in which V1 provides low-level, spatially detailed information to a high-level visual area such as MT/V5, causally required for conscious motion vision [108]. As is the case in vivo, the isolated circuit includes a projection from MT back to V1, but neither V1 nor MT has connections to any other cortical areas. Would this reduced V1-MT loop be able to sustain conscious motion vision? Neurorepresentationalism denies this possibility on the ground that this isolated circuit cannot solve the Modality Identification (MI) problem [25]: the circuit has no way of discriminating the processed information as pertaining to motion vision as opposed to other visual submodalities (e.g. color, shape, texture) or other main modalities. Therefore this system will not be capable of coding for distinct qualitative properties, the basis of which is provided by a multimodal topology in which each (sub)modality occupies a unique position – partly constituted by feature detectors that are unique to a submodality and partly by the correlational and predictive relationships existing between (sub)modalities [1,25]. Thus, the minimal neural substrate for sustaining a unitary conscious experience is predicted to comprise more than one submodality, and experience will grow qualitatively richer and more differentiated once more (sub)modalities are included. Anatomically, this implies that not only bottom-up and top-down projections in sensory cortical hierarchies are deemed necessary for conscious experience, but also extensive *lateral connectivity* in the cortex and thalamus, that is, between brain areas that are each traditionally viewed as belonging to one particular sensory modality (Fig. 4). Evidence for such cross-modal connections is mounting [21, 100,109–113].

Identifying 'the' neural correlate of consciousness (NCC) under this conceptualization is arguably a difficult but also simplified idea, because the conscious percept of e.g. visual motion will be co-dependent on representational activity in many connected cortical areas, the neurons of which should at least be active at baseline level (for a more detailed account involving firing rate- versus firing phase in relational coding, see [1,25]). However, it may be reasonably assumed that the postulated

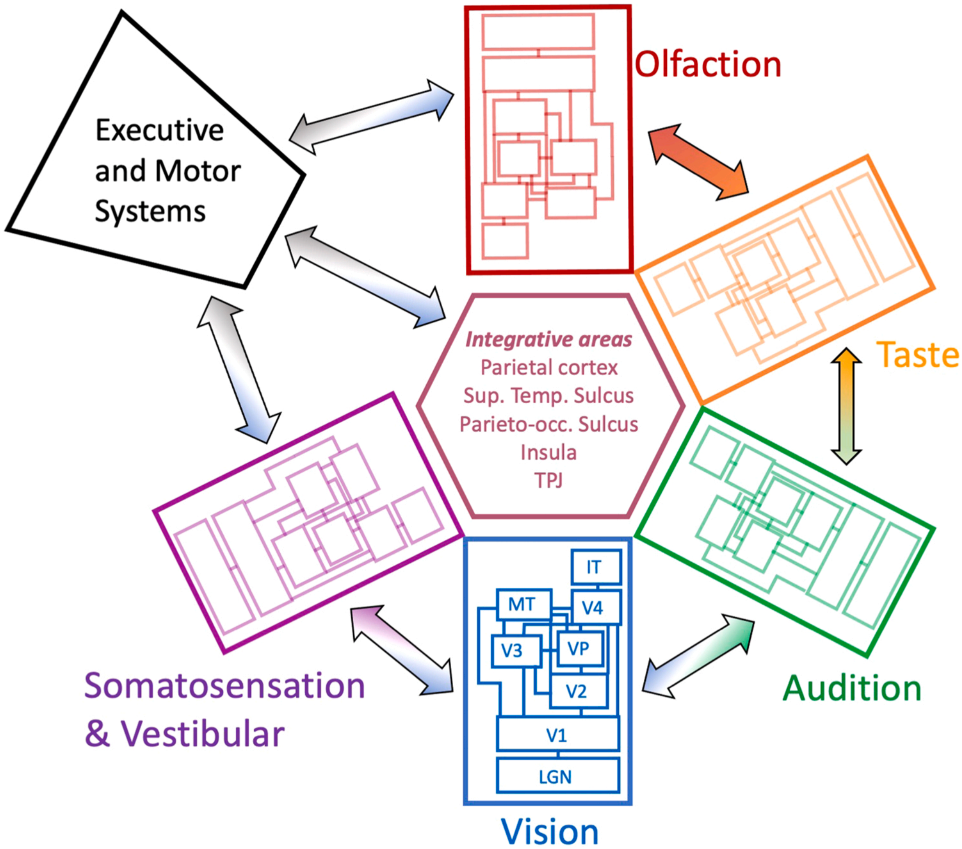

Fig. 4. Interconnected modality-specific cortical hierarchies, higher integrative areas and motor-related systems in frontal cortex. The visual cortical hierarchy (blue, bottom) is rendered in simplified form (based on [148]; LGN, lateral geniculate nucleus; VP, ventral posterior cortex; MT, middle temporal area or V5; IT, inferotemporal cortex). Evidence has been raised for hierarchical processing in other modalities as well (e.g. for audition: [149]; somatosensation: [150]; olfaction: [151]; see also [1]). This hierarchical structure is symbolized by semitransparent copies of the visual hierarchy. Integrative, multimodal areas include the superior temporal sulcus (Sup. Temp. Sulcus), the parieto-occipital sulcus and temporo-parietal junction (TPJ). These areas form a subset of all cortical areas implicated in multisensory processing (see e.g., [21,152, 153]). Lateral connectivity between modalities is symbolized by bicolored, bidirectional arrows. Altogether, the interconnected unimodal hierarchies and higher integrative areas form the multimodal sensory network postulated to form the core substrate for conscious experience in mammalian brains. Executive and motor systems in frontal cortical areas are informed by the multimodal sensory network to guide decisions and actions, but are not considered essential for consciousness. In return, they provide efference copy signals and other feedback to shape predictions about sensory consequences of actions. In this sense, the motor system functions as an additional modality.

multimodal topology is actively coded in corticothalamic systems, not by more basal, subcortical systems involved in maintaining basic vital functions such as breathing, blood pressure and osmotic control (mediated by e.g. hypothalamic and brainstem nuclei) or by the cerebellum or other structures where inflicted damage does not notably hamper consciousness (e.g. prefrontal cortex [1,114]). A topology of cortical systems, each characterized by a distinct hierarchical structure within a modality, but also by specific connections with other modalities, may be conceived to be implemented primarily by sensory and higher associative cortical areas, but in addition efference copy and/or predictive signals from motor systems will be integrated in these interdigitating representational hierarchies: these are necessary for the stabilization of our conscious 'world view' in the face of eye-, head- and body movements [115,116]. Even though this framework restricts the neural substrate of consciousness mainly to the sensory cortices and their intimately connected, integrative areas (e.g. parietal cortex; Fig. 4), it is clear that isolating a precisely anatomically localizable 'NCC' would become problematic in practice. If this is the fate of any simplified NCC concept, so be it. Theoretically, identifying *causal mechanisms* underlying consciousness is much more interesting than identifying a (more loosely defined) correlate.

Finally, it is worthwhile considering what Neurorepresentationalism has to say about the Real Problem of consciousness relative to the Hard Problem [89]. If we are able to advance on the Real Problem, this can be argued to help us forward in tackling the Hard Problem as well. The Hard Problem may be thought of as a more ambitious project that the Real Problem, but under a functional-teleological viewpoint, the two problems are arguably closely related. Because Neurorepresentationalism attributes a clear function to conscious experience, it offers to approach the Real Problem by explaining the various properties of conscious experience in terms of its biological-psychological functionality for the subject, which then logically connects to physical brain-body processes clarifying how this functionality may be realized. Neurorepresentationalism recognizes that conscious perception is deeply marked by qualitative properties such as smell, color, depth, shape, but also functionally characterizes these properties as reflecting *the best possible representations* of the corresponding properties of perceived objects for the subject planning to act in his world.

For instance, if one accepts that a (healthy) conscious experience of an apple is characterized by both a specific view of the apple as well as the recognition of the identity of the object (belonging to the category of 'apple'), then we can next acknowledge the functionality of both properties: to identify the apple as a target for foraging behavior, we need to know what the object is (i.e., an apple, as something edible and tasty), but also to have a specific rendering of how the object is positioned relative to our head, arms and hands (to be able to grab the object efficiently). From this dual functionality, one may derive computational models bridging high-level representational properties to low-level hardware, such as neurons and plastic synapses. Recent work in our lab has indeed confirmed that a single, multi-layer predictive-coding model can realize both aspects (i.e., specific image and invariant object representation) within the same network.

Now, the Hard Problem becomes approachable from the same functional viewpoint: regardless of precise contents and consciousness properties, consciousness 'is there' because it is *functional* to have a multimodal survey of our situation, rather than having a non-conscious, purely quantitative 'blob of information' that would be much less useful in helping us to navigate the world. In other words, if an organism lacks consciousness, it is still able to process information to the extent that individual stimuli can lead to reflexes or habitual actions, but a functionality to represent a stimulus meaningfully *within* its environmental and multisensory context in an immediately graspable form would be lacking, and this will strongly hamper planned, goal-directed behaviors. If the appropriate representational machinery is present in the wakeful brain, it follows that this active machinery will do its hypothesis-generating job at multiple levels of representation, regardless of the precise external or internal inputs into the system, so that some kind of conscious experience will be always generated in that state, regardless of its contents. In this sense, Neurorepresentationalism is reminiscent of other, teleological positions in the philosophy of mind [7]. In sum, the question of why and how there is consciousness in the universe in the first place, can be answered by saying that it is deeply functional to have it, whereas the alternative of having no consciousness at all would severely hamper complex planned behaviors, at least in biological creatures.

6. Comparison to other neuroscientific theories

How does Neurorepresentationalism compare to other neuroscience-based theories of consciousness? Space is lacking to address this comparison fully, so it will be limited to the main differences with three other theories: (i) Global Neuronal Workspace Theory (GNWT) [9–11], (ii) Integrated Information Theory (IIT; [12-15]) and (iii) Active Inference Theory (AIT; [16–19]). According to GNWT, primary and secondary sensory areas feed their information into a highly connected, central neural domain, from where it can be broadcast to reach other functional modules such as for working memory, emotional evaluation, motor decisions and attention. This central domain forms a 'workspace'- an informational hub or marketplace - but in order for information to reach it, sensory signals (at first being 'preconscious') must pass a threshold of ignition that unleashes the broadcasting of activity – considered equivalent to reaching consciousness. When following the distinction between 'phenomenal' and 'access' consciousness [107](but see [1]), the workspace mechanism fits the notion of 'access' consciousness [117], as it is not particularly concerned with how phenomenal content arises bus focuses on the logistics of information distribution across neural systems operating on and using this content (e.g. for memorization, motor decisions etc.). In line with Dennett's deflationary perspective on phenomenal content [118], GNWT thus focuses on what happens with the information once (phenomenal) consciousness has been realized (cf. [117]). This view markedly deviates from Neurorepresentationalism, which does acknowledge qualitative, multimodal properties as a hallmark of conscious experience, and does address how phenomenal content arises from the interplay of sensory inputs, predictive representations and efference copy. Previously, I have argued that computational network models of GNWT, based on reinforcement learning principles, have been successful in simulating cognitive processes such as working memory and in solving cognitive paradigms such as the Stroop task [119], but are insufficient when it comes to credibly simulate conscious processing, that is, insufficient in explaining why the network model would yield consciousness as opposed to automated, non-conscious processing. Thus, this argumentation characterizes GNWT models as underconstrained [1]. Recently, Whyte and Smith [120] combined PP principles with GNWT, which may eventually bring GNWT closer to an explanation of phenomenal awareness.

In contrast to GNWT, Integrated Information theory (IIT) does focus on explaining phenomenal consciousness and does so by postulating that consciousness equates with the cause-effect structure (neural, or other) having a maximum of integrated cause-effect power within a larger system. IIT is founded on the premise that conscious experience is both differentiated (with respect to other, alternative experiences) and integrated (marking the unity of conscious experience as already referred to above). According to IIT, conscious experience is irreducible to non-interdependent components [121]. Mathematically, Integrated Information is captured by Φ, which measures the conscious system's irreducibility with respect to the partitioning of system elements (for further details, see [12–15]).

IIT deserves praise for its fundamental and audacious attempt to establish a quantitative basis for physical substrates of consciousness, but is also subject to various criticisms (e.g. [1,25,122,123]). One main line of critique holds that IIT is underconstrained, as integrated information can reach considerable magnitudes in systems not considered to assume conscious states at all, such as cyclones, DVD players, or Vandermonde matrices marked by both differentiation and integration (also living human brains clinically labeled as 'nonconscious', such as in coma or anesthesia, should be considered by IIT to possess a certain degree of consciousness). IIT is defended against this critique by arguing that (partial) panpsychism is a tenable position [124]. This viewpoint implies that we may have been thinking too conservatively about which organisms or non-living objects in nature may possess a considerable degree of consciousness. In return, however, this argument can be countered by the objection that IIT's position leaves us no criteria to test which systems are conscious or not (except by measuring Φ itself, which is practically impossible); options for cross-examination and cross-validation are lacking. This makes IIT essentially untestable [125].

A further objection speaking against IIT is the lack of a 'world-matching' principle, which should apply at least to (conscious) perception. In healthy people, perception is usually coupled to efficient navigation through, and interactions with the external world, indicating that perceptual content reflects objects and properties of our environment in such a way that it can be used to satisfy evolutionarily drives for survival and reproduction. Because – in contrast to Neurorepresentationalism – IIT assumes that consciousness does not exert any function, percepts can be literally about anything, no matter how remote they are from our daily-life reality. Related to the lack of world- or reality-matching, integrated information is also devoid of intentionality, defined above as an overarching hallmark of conscious experience. Finally, IIT maintains that not only active, but also inactive neurons can contribute to conscious experience, which is assumed because these neurons help to specify what is *not* part of one's current experience. However, inactive neurons are unable to transfer information about their state to other members of the larger network, and may therefore as well be eliminated from the network: the other neurons will remain unaffected by their removal anyway (the difference is not 'noted'). This leaves IIT with a functional contradiction about the (un)importance of inactive neurons.

How does Neurorepresentationalism hold up against the criticisms raised against IIT? The functional and architectural requirements of Neurorepresentationalism are more constrained and more strongly defined than for IIT, as Neurorepresentationalism relies on interlinked sensory hierarchies computing low- and high-level predictive representations and accompanying errors, matching the specific bottom-up and top-down flows of cortical information (cf. [66,69,74,78]). Combined with the requirements for cross-modal predictive activity, inclusion of efference copy signals and upscaling to superinference levels (Figs. 3 and 4), this configuration is argued to be too specific to be found throughout nature, be it living of non-living (but see below for AI devices). Thus, the required computations and architectures preclude panpsychism in Neurorepresentationalism, whilst not excluding a possible implementation in artificial computing hardware. At the same time, representations carry the signature of 'best-guess' hypotheses, which implies that they are *about* something, fulfilling the basic hallmark of intentionality because hypotheses by definition have an interpretandum, or object on which the hypothesizing is centered. In practice, this means that conscious brain systems are constantly trying to infer what is 'out there' in the world (or the body) to explain the inputs impacting on our sense organs, and this inferential activity is internally construed to produce reality-matching, or in other words, producing a world model as accurately as possible. Thus, where IIT proposes Integrated Information or Φ as core concept, the key principles in Neurorepresentationalism are given by its computational rules for learning and inferring latent, predictive representations, which emergently give rise to consciousness once they are sufficiently scaled up in both spatial and multimodal dimensions.

Based on similar PP principles as deployed by Neurorepresentationalism, Active Inference has been proposed as a framework for understanding action-based sensory processing and perception, ramifying into consciousness research (e.g., [16–19,116]). Different from Neurorepresentationalism, Active Inference postulates that overt actions are essential for sensory inputs to reach consciousness; there is no 'seeing' without 'looking'. Actions such as eye movements are accompanied by predictions on their sensory consequences, and the novel sensory feedback elicited via action serves to reduce uncertainty in one's internal world model, thereby making predictions more accurate. Given the above-mentioned critique that consciousness persists in the absence of overt motor movement, Active Inference takes on board the notion that actions can also be generated covertly, that is internally, by way of motor imagery (fictive movements) or directing endogenous attention [126]. Such covert actions bring Active Inference closer to Neurorepresentationalism, which does acknowledge that motor and attentional systems can influence conscious experience (including imagery), but that this influence is not necessary per se for experience to exist.

Despite this kinship, at least some aspects of conscious experience appear to be hard to reconcile with a strong, mandatory requirement for motor action (e.g., synesthesia; phantom limb sensations in amelia; see above and [35]). Based on empirical findings, Neurorepresentationalism maintains a distinction between conscious experience on the one hand and (overt or covert) motor actions and attention on the other hand, as conscious experience is maintained in their absence [1,20]. Within the Active Inference framework, attention is associated with the expected precision applying to a particular sensory modality [127] and this obligatory one-to-one coupling of attention to expected precision is not found in Neurorepresentationalism. According to Neurorepresentationalism, (endogenous) attention is a high-level, cognitive factor having a more loosely defined relationship with conscious sensation: at least some conscious content persists in the absence of directed attention (such as information about the background or periphery of a perceived visual scene). Attention is understood here to intensify the processing of information about the object under focal scrutiny, but is thereby not a necessary prerequisite for consciousness (cf. [20]).

Upon further scrutiny, other differences between Neurorepresentationalism and Active Inference become apparent: the two theories make contrasting predictions about the importance of baseline firing activity in early sensory cortices, functioning to maintain multimodal topologies (Neurorepresentationalism) versus to code residual errors (AIT). The technicalities of the arguments will not be reviewed in detail here. What is worth highlighting, though, is that Neurorepresentationalism proposes the concept of multi-level representations (and our subjective inability to imagine consciousness arising by emergent correspondence) as a way of tackling the Hard Problem, whereas in Active Inference it is yet unclear how (or why) action-based PP as such might give rise to a subjective phenomenology. In conclusion, despite their common dependence on PP principles, significant differences between Active Inference and Neurorepresentationalism can be noted, and to a great extent these can be addressed by neuroscientific experiments and clinical observations.

7. Implications for animals, artificial intelligence and patients with disorders of consciousness

Which implications does a neurorepresentationalist view have for our thinking about consciousness in animals, patients with DoCs and intelligent machines? To begin with animals, to which ethical considerations of well-being and bioindustry are highly relevant, we have previously proposed a number of 'Indicators of Consciousness' (IoCs) which have been derived from Neurorepresentationalism and Intrinsic Consciousness Theory [128]. The key idea is that it is extremely difficult, if not impossible, to define hard-threshold criteria about a possible subjectivity in animals and that it is more fruitful to delineate less definite IoCs which are linked to conscious experience, understood as multimodal, situational survey. When taken together, a given animal species can score positively or negatively (or inconclusively) on each of these IoCs, of which we have proposed six, yielding an overall estimate (along a graded scale) of how likely the species is to have consciousness. This approach is somewhat comparable to the Glasgow Coma Scale applied to patients with a compromised or uncertain consciousness status [129].

The first IoC is derived from the architectural and functional constraints in sustaining a neural structure for multimodal, spatially encompassing representations. In vertebrates, these constraints naturally link to having a neocortex or similar structure – reciprocally connected to a thalamic system - with a physiology resembling the human brain (displaying, for instance, deep sleep, REM and awake states as characterized by desynchronized EEG versus up- and down state activity of single neurons, etc.). This IoC is relatively difficult to apply to highly developed organisms with a central nervous system very different from vertebrates (e.g., in the octopus). Therefore, other IoCs must be deployed in addition.

A second IoC considers whether or not the organism is capable of goal-directed behavior, planning and model-based learning (or neural correlates thereof; cf. [34,37,40,130]). If so, this again does not prove the presence of consciousness on its own, but – as postulated above - we may assume that these capacities greatly benefit from some degree of multimodal, situational survey, and therefore they can be regarded as an 'echo' or behavioral trait suggestive of consciousness as an enabling mental ability. Leveraging on the situational and spatial aspects of conscious experience, the display of sophisticated visuospatial behaviors can be taken as a third IoC, or as a sign pointing to conscious experience, as a function enabling these behaviors. For instance, rats are normally capable of finding their way in complex maze environments, but decorticated subjects get easily trapped in mazes and even single alleys [131]. Similarly, the susceptibility to illusions and ambiguous stimuli, the presence of human-like psychometrics and metacognitive judgments in perceptual tasks, as well as behavioral manifestations of episodic memory have been proposed as a fourth, fifth and sixth IoC, respectively (cf. [132]). When integrating the scores on these IoCs for rodents and birds, it turns out that the overall scores clearly argue in favor of at least some degree of consciousness in the taxonomic order of rodentia and the class of birds [133].

In principle, these and similar IoCs can be extrapolated to computers and robots, however with the caveat that computers can be preprogrammed to exhibit particular goal-directed, visuospatial, planning, psychometric (etcetera) behaviors, and can nowadays rely on chip arrays that are many orders of magnitude faster than the speed at which neurons operate. In principle, AI devices could leverage this computing speed to generate complex behaviors like highly developed vertebrate species display, but without a necessity to appeal to consciousness. One way to deal with this computational inequality is to evaluate AI consciousness not based on a quick, Turing-type of test (which might be solved by a preprogrammed and nonconscious system), but to study the AI's ethology on the long run, paying particular attention to improvising behaviors in unfamiliar situations requiring novel planning [133].

Two examples may serve to clarify this stance on AI consciousness. First, the museum of Emerging Science and Innovation in Tokyo displays Honda's robot Asimo [134], which is able to stage a technically advanced show of skills, including kicking a soccer ball into a goal placed in the museum's arena. This could be taken as a goal-directed behavior, yet from repeated daily shows it becomes clear that the robot's behavior is highly stereotyped (despite variations in soccer ball position relative to the goal, prompting adaptive sensorimotor adjustments), and can be mediated by pre-programmed instructions without having to appeal to conscious representation. Thus, this situation calls for more tests in addition to using goal-directed behavior as an IoC for animals which, for instance, will take the ability to improvise in novel situations into account, because such improvising behaviors cannot be pre-programmed. This is not to say that improvising in novel situations would be useless in studying animals, but rather that robots should be arguably subjected to more stringent scrutiny because of the option of pre-programming.

Second, we consider the example of a group of Orcas collectively hunting for seals in an artic environment [135]. The display of complex visuospatial behaviors by which an Orca attempts to capture a seal that is trying to defend its position on an ice shelf – including its tail flips that generates waves in synchronized coordination with fellow Orcas – has been argued to provide a positive IoC because it necessitates precisely the kind of multimodally rich, situational survey outlined above [133]. In principle, also this kind of behavior could be mimicked by AI-driven robots, although an exact emulation will be hard to achieve in case improvisations coping with novel defense strategies of the seal would be required. When sufficiently fast and powerful, serially programmed computers or deep-learning networks may eventually be able to mimic even these complex improvising behaviors and meet other IoCs. This would then force us to admit that AI systems can meet all IoCs used to assess consciousness in animals, without having certainty that these systems are actually conscious in the same sense as meant for humans and other animal species. Apart from this caveat (that must, as a matter of course, accompany inferences from third-person evidence about consciousness in other beings than oneself), Neurorepresentationalism makes no assumption that would strictly prohibit consciousness from arising in AI systems, but it should be recalled that this permissiveness to multiple realizability does not entail an endorsement of (classic) functionalism. Classic functionalism conforms to reductive and eliminative materialism, whereas Neurorepresentationalism does not aspire to reduce or eliminate qualitatively rich, phenomenal experience.

Immobile patients with DoCs, however, confront us with a very different puzzle: they may largely lack the motor capacities to display overt behaviors, leaving clinicians with the limited option to estimate their consciousness by way of neural measures of the anatomical and physiological integrity of their brain. In addition to anatomical measures relying on e.g. MRI or DTI (diffusion tensor imaging) scans, physiological measures can be used, such as burst-suppression activity in the spontaneous EEG [136], and EEG or EMG reactivity to sensory stimuli and verbal commands [137]. Inspired by IIT, a perturbational complexity index (PCI) has been introduced, reflecting the spatiotemporal complexity of the EEG wave pattern elicited by a TMS pulse applied locally to the cortex [138,139]. Rather than measuring the complexity or integrated information of subjective experience per se, PCI reflects the ability of excited cortical populations to propagate and reverberate wave activity through the recurrently connected network of the cortex (potentially involving subcortical feedback loops as well) – an ability that can be hampered by up- and down states such as during slow-wave sleep (cf. [140]). The neurorepresentational perspective entails a different type of measure for DoC patients, which is more akin to the approach to instruct patients to generate yes/no answers via mental imagery [141]. Even though DoC patients may be completely immobilized or locked-in, their conscious brains may still be susceptible to illusions and ambiguous pictures, and the neural substrates for generating wakefulness-like psychometric responses, metacognitive judgment, visuospatial behaviors, planning for GDB, and episodic memory may still be fully or partially intact. Assuming subjects are responsive to instructions from a clinician, they may be asked to engage in mental planning, episodic memory retrieval, etcetera, while recording brain activity changes via fMRI, scalp EEG or stereo-EEG (sEEG). Herein the range of mental activities can in principle be expanded beyond those studied in [141] because of their denomination as IoCs. For instance, presentations of illusions may be contrasted with unambiguous stimuli to leverage on the different fMRI laminar activation patterns identified in healthy persons (cf. [142]). With sufficient neural resolution – such as achieved in ensemble recordings and population decoding - representational measures of consciousness may be derived, meaning that not only the state, but also aspects of the content of conscious experience can be decoded. This overall approach to DoC patients has already been brought closer to clinical practice by the development of a brain-computer interface via which nearly locked-in amyotrophic lateral sclerosis (ALS) patients can communicate by operating a computer typing program [143](cf. [144,145]). Also the deployment of speech-decoding brain-computer interfaces connected to sensorimotor cortical motor areas is promising in this respect [146].

8. Conclusion

Here we have reviewed the main tenets of a neuroscientific theory of consciousness – Neurorepresentationalism – which, as I have argued, provide a viable avenue for unraveling the neural basis of consciousness. Reasoning from five hallmarks of subjective experience – (i) multimodal richness, (ii) situatedness and immersion, (iii) unity, (iv) dynamics combined with stability, and (v) intentionality - Neurorepresentationalism defines conscious experience as a multimodal, situational survey of our situation, with our body immersed in it, subserving planning of goal-directed behaviors and complex decision-making. On the one hand, the predictive-processing type of architecture for computing perceptual inferences and errors offers a quantitative basis for the framework and makes this basis testable. The extension of a single, unimodal sensory hierarchy to a large-scale, multimodal topological structure provides the basis for building high-level representations, but it is also fair to say that this 'superinferential' aspect of the theory [96], as laid out across different representational levels (Fig. 3), needs further elaboration in theory, computational modeling and experiment. Yet, the connotated concept of emergent correspondence between low-level, neuronal and high-level, multimodal representation offers a useful handle on the Hard Problem of consciousness.

As compared to a few other neural theories of consciousness, several strong or weaker dissimilarities are noted: with its focus on executive aspects of cognitive processing, GNWT is a very different type of theory than Neurorepresentationalism. Relative to IIT, major differences are apparent in the approach to neural mechanisms as well, with Neurorepresentationalism putting stronger functional requirements for conscious experience into place, with reality-matching constraints in perception. While PP principles form a shared basis with Active Inference Theory, this framework nonetheless diverges from Neurorepresentationalism, which does not attribute an essential role for motor behavior in consciousness. Neuroreprentationalism provides handles for assessing consciousness in animals, intelligent computers or robots, as well as unresponsive patients: it suggests a way to translate the first-person hallmarks of conscious experience into behaviorally observable or brain-based derivatives, such as the physiology of the waking-sleeping brain, goal-directed and visuospatial behaviors and susceptibility to illusions. Additional IOCs appear to be required to estimate consciousness in AI devices, which may be enabled by fast computing capacities to carry out conscious-like behaviors in a preprogrammed fashion. In addition to established physiological brain markers to assess DoCs, Neuroreprentationalism suggests the development and testing of representational measures of consciousness, which focus on the decoding of content of neural information as being specific to the awake, conscious state.

Acknowledgments

I would like thank the organizers and participants of the MESEC seminar held on Corsica in 2021 (Mediterranean Seminar of Consciousness), in particular Mathis Bassler, Ulysse Klatzmann, Kwangjun Lee, Esteban Munoz, Julie Hemmerlin and many others for helpful discussions. This project has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3).

Competing interest statement

The author has no competing interests to declare.

References

[1] C.M.A. Pennartz, The Brain's Representational Power – On Consciousness and the Integration of Modalities, MIT Press, Cambridge, Mass, 2015.
[2] W. Lycan, Representational theories of consciousness, in: E.N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Stanford, CA, USA, 2019.
[3] F. Dretske, Experience as representation, Philos. Issues 13 (2003) 67–82.
[4] T. Crane, The intentional structure of consciousness, in: Q. Smith, A. Jokic (Eds.), Consciousness: New Philosophical Perspectives, Oxford University Press, Oxford, UK, 2003.
[5] S. Shoemaker, Phenomenal character, Nous 28 (1) (1994) 21–38.
[6] D.J. Chalmers, The representational character of experience, in: B. Leiter (Ed.), The Future for Philosophy, Oxford University Press, Oxford, 2004.
[7] R.G. Millikan, Language, Thought, and Other Biological Categories, The MIT Press, Cambridge, MA, 1984.
[8] D.J. Chalmers, Facing up to the problem of consciousness, J. Conscious. Stud. 2 (1995) 200–219.
[9] S. Dehaene, J.P. Changeux, Experimental and theoretical approaches to conscious processing, Neuron 70 (2) (2011) 200–227.
[10] B.J. Baars, The conscious access hypothesis: origins and recent evidence, Trends Cogn. Sci. 6 (1) (2002) 47–52.
[11] B.J. Baars, Conscious contents provide the nervous system with coherent, global information, in: R.J. Davidson, G.E. Schwartz, D. Shapiro (Eds.), Consciousness and Self-Regulation, Plenum Press, New York, 1983.
[12] C. Koch, et al., Neural correlates of consciousness: progress and problems, Nat. Rev. Neurosci. 17 (5) (2016) 307–321.
[13] G. Tononi, et al., Integrated information theory: from consciousness to its physical substrate, Nat. Rev. Neurosci. 17 (7) (2016) 450–461.
[14] G. Tononi, An information integration theory of consciousness, BMC Neurosci. 5 (1) (2004) 42.
[15] A. Haun, G. Tononi, Why does space feel the way it does? Towards a principled account of spatial experience, Entropy 21 (12) (2019) 1160.
[16] J. Hohwy, A. Seth, Predictive processing as a systematic basis for identifying the neural correlates of consciousness, Philos. Mind Sci. (1(II)) (2020).
[17] T. Parr, et al., Perceptual awareness and active inference, Neurosci. Conscious. 2019 (1) (2019).
[18] J. Hohwy, The Predictive Mind, Oxford University Press, 2013.
[19] J.A. Hobson, K.J. Friston, Waking and dreaming consciousness: neurobiological and functional considerations, Prog. Neurobiol. 98 (1) (2012) 82–98.
[20] C. Koch, N. Tsuchiya, Attention and consciousness: two distinct brain processes, Trends Cogn. Sci. 11 (1) (2007) 16–22.
[21] G.T. Meijer, et al., The circuit architecture of cortical multisensory processing: distinct functions jointly operating within a common anatomical network, Prog. Neurobiol. 174 (2019) 1–15.
[22] C.R. Fetsch, G.C. DeAngelis, D.E. Angelaki, Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons, Nat. Rev. Neurosci. 14 (6) (2013) 429–442.
[23] S. Deneve, A. Pouget, Bayesian multisensory integration and cross-modal spatial links, J. Physiol. 98 (1–3) (2004) 249–258.
[24] M.O. Ernst, M.S. Banks, Humans integrate visual and haptic information in a statistically optimal fashion, Nature 415 (6870) (2002) 429–433.
[25] C.M.A. Pennartz, Identification and integration of sensory modalities: neural basis and relation to consciousness, Conscious. Cogn. 18 (3) (2009) 718–739.
[26] W. Fujisaki, et al., Recalibration of audiovisual simultaneity, Nat. Neurosci. 7 (2004) 773–778.
[27] B. Merker, K. Williford, D. Rudrauf, The integrated information theory of consciousness: a case of mistaken identity, Behav. Brain Sci. (2021) 1–72.
[28] T. Nagel, What is it like to be a bat? Philos. Rev. 83 (1974), 453-50.
[29] S.M. Lehar, The World in Your Head: A Gestalt View of the Mechanism of Conscious Experience, Psychology Press, 2003.
[30] M. Velmans, Is human information processing conscious? Behav. Brain Sci. 14 (4) (1991) 651–669.
[31] A. Sahraie, et al., Consciousness of the first order in blindsight, Proc. Natl. Acad. Sci. USA 107 (49) (2010) 21217–21222.
[32] T.W. Robbins, R.M. Costa, Habits, Curr. Biol. 27 (22) (2017). R1200-R1206.
[33] L. Weiskrantz, The problem of animal consciousness in relation to neuropsychology, Behav. Brain Res. 71 (1–2) (1995) 171–175.
[34] S.I. Rusu, C.M.A. Pennartz, Learning, memory and consolidation mechanisms for behavioral control in hierarchically organized cortico-basal ganglia systems, Hippocampus 30 (1) (2020) 73–98.
[35] C.M.A. Pennartz, Consciousness, representation, action: the importance of being goal-directed, Trends Cogn. Sci. 22 (2) (2018) 137–153.
[36] A. Dickinson, Associative learning and animal cognition, Philos. Trans. R. Soc. Lond. B Biol. Sci. 367 (1603) (2012) 2733–2742.
[37] N.D. Daw, Y. Niv, P. Dayan, Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control, Nat. Neurosci. 8 (12) (2005) 1704–1711.
[38] M.A. Cohen, D.C. Dennett, N. Kanwisher, What is the bandwidth of perceptual experience? Trends Cogn. Sci. 20 (5) (2016) 324–335.
[39] G. Pezzulo, et al., Internally generated sequences in learning and executing goal-directed behavior, Trends Cogn. Sci. 18 (12) (2014) 647–657.
[40] A.D. Redish, Vicarious trial and error, Nat. Rev. Neurosci. 17 (3) (2016) 147–159.
[41] D.M. Armstrong, A Materialist Theory of the Mind, Humanities Press, New York, 1968.
[42] D. Lewis, Psychophysical and theoretical identifications, Australas. J. Philos. 50 (3) (1972) 249–258.
[43] D.C. Dennett, Consciousness Explained, Little, Brown, Boston, 1991.
[44] J.R. Manns, L.R. Squire, Perceptual learning, awareness, and the hippocampus, Hippocampus 11 (6) (2001) 776–782.
[45] N. Suthana, I. Fried, Percepts to recollections: insights from single neuron recordings in the human brain, Trends Cogn. Sci. 16 (8) (2012) 427–436.
[46] M.A. Peterson, J.T. Enns, The edge complex: implicit memory for figure assignment in shape perception, Percept. Psychophys. 67 (2005) 727–740.
[47] J. Kim, Making sense of emergence, Philos. Stud. 95 (1999) 3–36.
[48] P.M. Churchland, The Engine of Reason, the Seat of the Soul, MIT Press, Cambridge, Mass., USA, 1995.
[49] D.C. Dennett, Sweet dreams. Philosophical Obstacles to a Science of Consciousness, MIT Press, Cambridge, MA, 2005.
[50] R. Rorty, Mind-brain identity, privacy and categories, in: D. Rosenthal (Ed.), Materialism and the Mind-body Problem, Prentice Hall, Englewood Cliffs, N.J, 1971, pp. 174–199.
[51] P.S. Churchland, T.J. Sejnowski, The Computational Brain, MIT Press, Cambridge, Mass., USA, 1992.
[52] J.R. Searle, Seeing Things as They Are, Oxford University Press, Oxford, 2015.
[53] J.L. Austin, Sense and Sensibilia, Oxford University Press, Oxford, 1962.
[54.] T. Reid, Essays on the intellectual powers of man (reprinted Critical Edition), Edinburgh edition of Reid Thomas, K. Haakonssen, D.R. Brookes (Eds.), Vol. 3, 1785, Edinburgh University Press, Edinburgh, 2002.
[55] J.K. O'Regan, A. Noë, A sensorimotor account of vision and visual consciousness, Behav. Brain Sci. 24 (2001) 939–973.
[56] H. Brown, K. Friston, S. Bestmann, Active inference, attention, and motor preparation, Front. Psychol. 2 (2011) 218.
[57] P. Brugger, et al., Beyond re-membering: phantom sensations of congenitally absent limbs, Proc. Natl. Acad. Sci. USA 97 (11) (2000) 6167–6172.
[58] J. Müller, Handbuch der Physiologie des Menschen für Vorlesungen, J. Hölscher, Coblenz, 1838.
[59] E.G. Boring, A History of Experimental Psychology, second ed. Appleton Century Crofts, New York, 1950.
[60] H. von Helmholtz. Handbuch der physiologischen Optik, Leopold Voss, Leipzig, 1866.
[61] G. O'Brien, J. Opie, A connectionist theory of phenomenal experience, Behav. Brain Sci. 22 (1999) 127–148.
[62] R.L. Gregory, Perceptions as hypotheses, Philos. Trans. R. Soc. Lond. B Biol. Sci. 290 (1038) (1980) 181–197.
[63] G. Anton, Ueber die Selbstwahrnehmung der Herderkrankungen des Gehirns durch den Kranken bei Rindenblindheit und Rindentaubheit, Arch. Psychiatr. Nervenkrankh. 32 (1) (1899) 86–127.
[64] M.V. Srinivasan, S.B. Laughlin, A. Dubs, Predictive coding: a fresh view of inhibition in the retina, Proc. R. Soc. Lond. B Biol. Sci. 216 (1205) (1982) 427–459.
[65] P. Dayan, et al., The Helmholtz machine, Neural Comput. 7 (5) (1995) 889–904.
[66] T.S. Lee, D. Mumford, Hierarchical Bayesian inference in the visual cortex, J. Opt. Soc. Am. A Opt. Image Sci. Vis. 20 (7) (2003) 1434–1448.
[67] A. Clark, Whatever next? Predictive brains, situated agents, and the future of cognitive science, Behav. Brain Sci. 36 (3) (2013) 181–204.
[68] H. Helmholtz von, in: J.P.C. Southall (Ed.), Handbuch der physiologischen optik Vol. 3, Dover, Dover, 1962.
[69] R.P. Rao, D.H. Ballard, Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects, Nat. Neurosci. 2 (1) (1999) 79–87.
[70] M.W. Spratling, Predictive coding as a model of response properties in cortical area V1, J. Neurosci. 30 (9) (2010) 3531–3543.
[71] C. Wacongne, J.P. Changeux, S. Dehaene, A neuronal model of predictive coding accounting for the mismatch negativity, J. Neurosci. 32 (11) (2012) 3665–3678.
[72] C.M.A. Pennartz, et al., Towards a unified view on pathways and functions of neural recurrent processing, Trends Neurosci. 42 (9) (2019) 589–603.
[73] K. Friston, A theory of cortical responses, Philos. Trans. R. Soc. Lond. B Biol. Sci. 360 (1456) (2005) 815–836.
[74] S. Dora, S.M. Bohte, C.M.A. Pennartz, Deep gated hebbian predictive coding accounts for emergence of complex neural response properties along the visual cortical hierarchy, Front. Comput. Neurosci. 15 (2021), 666131.
[75] T. Sato, G. Uchida, M. Tanifuji, Cortical columnar organization is reconsidered in inferior temporal cortex, Cereb. Cortex 19 (8) (2009) 1870–1888.
[76] M.J. Pearson, et al., Multimodal representation learning for place recognition using deep Hebbian predictive coding, Front. Robot. AI 8 (2021), https://doi.org/10.3389/frobt.2021.732023.
[77] G.B. Keller, T.D. Mrsic-Flogel, Predictive processing: a canonical cortical computation, Neuron 100 (2) (2018) 424–435.
[78] A.M. Bastos, et al., Canonical microcircuits for predictive coding, Neuron 76 (4) (2012) 695–711.
[79] C.A. Bosman, et al., Attentional stimulus selection through selective synchronization between monkey visual areas, Neuron 75 (5) (2012) 875–888.
[80] T. Egner, J.M. Monti, C. Summerfield, Expectation and surprise determine neural population responses in the ventral visual stream, J. Neurosci. 30 (49) (2010) 16601–16608.

[81] P.M. Goltstein, et al., In vivo two-photon Ca2+ imaging reveals selective reward effects on stimulus-specific assemblies in mouse visual cortex, J. Neurosci. 33 (28) (2013) 11540–11555.
[82] J.S. Montijn, U. Olcese, C.M. Pennartz, Visual stimulus detection correlates with the consistency of temporal sequences within stereotyped events of V1 neuronal population activity, J. Neurosci. 36 (33) (2016) 8624–8640.
[83] L.F. Rossi, K.D. Harris, M. Carandini, Spatial connectivity matches direction selectivity in visual cortex, Nature 588 (7839) (2020) 648–652.
[84] M.W. Self, et al., Distinct roles of the cortical layers of area V1 in figure-ground segregation, Curr. Biol. 23 (21) (2013) 2121–2129.
[85] D.J. Heeger, Theory of cortical function, Proc. Natl. Acad. Sci. USA 114 (8) (2017) 1773–1782.
[86] S. Marx, et al., Competition with and without priority control: linking rivalry to attention through winner-take-all networks with memory, Ann. N. Y. Acad. Sci. 1339 (1) (2015) 138–153.
[87] A.R. Koene, A model for perceptual averaging and stochastic bistable behavior and the role of voluntary control, Neural Comput. 18 (12) (2006) 3069–3096.
[88] J. Levine, Materialism and qualia: the explanatory gap, Pac. Philos. Q. 64 (1983) 354–361.
[89] A. Seth, Illuminating consciousness, New Sci. 251 (3350) (2021) 44–48.
[90] G. Holmes, The cerebellum of man, Brain 62 (1939) 1–30.
[91] A.A. Sokolov, R.C. Miall, R.B. Ivry, The cerebellum: adaptive prediction for movement and cognition, Trends Cogn. Sci. 21 (5) (2017) 313–332.
[92] W. Schultz, P. Dayan, P.R. Montague, A neural substrate of prediction and reward, Science 275 (5306) (1997) 1593–1599.
[93] H. Rambold, et al., Partial ablations of the flocculus and ventral paraflocculus in monkeys cause linked deficits in smooth pursuit eye movements and adaptive modification of the VOR, J. Neurophysiol. 87 (2) (2002) 912–924.
[94] K.J. Friston, W. Wiese, J.A. Hobson, Sentience and the origins of consciousness: from Cartesian duality to Markovian monism, Entropy 22 (5) (2020) 516.
[95] D. Marr, Vision, WH Freeman, New York, 1982.
[96] U. Olcese, M.N. Oude Lohuis, C.M.A. Pennartz, Sensory processing across conscious and nonconscious brain states: from single neurons to distributed networks for inferential representation, Front. Syst. Neurosci. 12 (2018) 49.
[97] L. Aitchison, et al., Synaptic plasticity as Bayesian inference, Nat. Neurosci. 24 (4) (2021) 565–571.
[98] A. Pouget, et al., Probabilistic brains: knowns and unknowns, Nat. Neurosci. 16 (9) (2013) 1170–1178.
[99] L. Aitchison, M. Lengyel, With or without you: predictive coding and Bayesian inference in the brain, Curr. Opin. Neurobiol. 46 (2017) 219–227.
[100] M.N.O. Lohuis, et al., Multisensory task demands temporally extend the causal requirement for visual cortex in perception, Nat. Commun. 13 (1) (2022) 2864.
[101] R. Jackendoff, Consciousness and the Computational Mind, MIT Press, Cambridge, MA, 1987.
[102] J.J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proc. Natl. Acad. Sci. USA 79 (8) (1982) 2554–2558.
[103] T. Metzinger, Empirical perspectives from the self-model theory of subjectivity: a brief summary with examples, Prog. Brain Res. 168 (2008) 215–245.
[104] K. Gödel, Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I, Mon. Math.Phys. 38 (1) (1931) 173–198.
[105] D.R. Hofstadter, Gödel, Escher, Bach, an eternal golden braid, Basic Books, New York, 1979.
[106] F.H. Lopes da Silva, Contribution to a neurophysiology of consciousness, Adv. Clin. Neurophysiol. Suppl. 57 (2004) 645–655.
[107] N. Block, Two neural correlates of consciousness, Trends Cogn. Sci. 9 (2) (2005) 46–52.
[108] J. Zihl, D. von Cramon, N. Mai, Selective disturbance of movement vision after bilateral brain damage, Brain 106 (Pt 2) (1983) 313–340.
[109] G.T. Meijer, et al., Neural correlates of multisensory detection behavior: comparison of primary and higher-order visual cortex, Cell Rep. 31 (6) (2020), 107636.
[110] L.A. Ibrahim, et al., Cross-modality sharpening of visual cortical processing through layer-1-mediated inhibition and disinhibition, Neuron 89 (5) (2016) 1031–1045.
[111] R.J. Morrill, A.R. Hasenstaub, Visual information present in infragranular layers of mouse auditory cortex, J. Neurosci. 38 (11) (2018) 2854–2862.
[112] C. Kayser, N.K. Logothetis, S. Panzeri, Visual enhancement of the information representation in auditory cortex, Curr. Biol. 20 (1) (2010) 19–24.
[113] N. van Atteveldt, et al., Multisensory integration: flexible use of general operations, Neuron 81 (6) (2014) 1240–1253.
[114] M. Boly, et al., 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) (2017) 9603–9613.
[115] M. Leinweber, et al., A sensorimotor circuit in mouse cortex for visual flow predictions, Neuron 96 (5) (2017) 1204.
[116] K. Friston, Am I self-conscious? (Or does self-organization entail self-consciousness?), Front. Psychol. 9 (2018) 579.
[117] G.A. Mashour, et al., Conscious processing and the global neuronal workspace hypothesis, Neuron 105 (5) (2020) 776–798.
[118] D.C. Dennett, Facing up to the hard question of consciousness, Philos. Trans. R. Soc. B Biol. Sci. 2018 (373) (1755) 20170342.
[119] S. Dehaene, M. Kerszberg, J.P. Changeux, A neuronal model of a global workspace in effortful cognitive tasks, Proc. Natl. Acad. Sci. USA 95 (24) (1998) 14529–14534.
[120] C.J. Whyte, R. Smith, The predictive global neuronal workspace: a formal active inference model of visual consciousness, Prog. Neurobiol. 199 (2021), 101918.
[121] M. Oizumi, L. Albantakis, G. Tononi, From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0, PLoS Comput. Biol. 10 (5) (2014), e1003588.
[122] S. Aaronson, Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander) https://www.scottaaronson.com/blog/?p=1799. (2014).
[123] T. Bayne, On the axiomatic foundations of the integrated information theory of consciousness, Neurosci. Conscious. 2018 (1) (2018) niy007.
[124] G. Tononi, C. Koch, Consciousness: here, there and everywhere? Philos. Trans. R. Soc. Lond. B Biol. Sci. 370 (2015) 1668.
[125] C.M.A. Pennartz, What is exactly the problem with panpsychism? (Commentary on Merker et al.), Behav. Brain Sci. 45 (2022) e57.
[126] M.B. Mirza, et al., Introducing a Bayesian model of selective attention based on active inference, Sci. Rep. 9 (1) (2019) 13915.
[127] J. Limanowski, K. Friston, Active inference under visuo-proprioceptive conflict: Simulation and empirical results, Sci. Rep. 10 (1) (2020) 4010.
[128] M. Farisco, S. Laureys, K. Evers, The intrinsic activity of the brain and its relation to levels and disorders of consciousness, Mind Matter 15 (2) (2017) 197–219.
[129] G. Teasdale, et al., The Glasgow Coma Scale at 40 years: standing the test of time, Lancet Neurol. 13 (8) (2014) 844–854.
[130] B.W. Balleine, J.P. O'Doherty, Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action, Neuropsychopharmacology 35 (1) (2010) 48–69.
[131] I.Q. Whishaw, T. Schallert, B. Kolb, An analysis of feeding and sensorimotor abilities of rats after decortication, J. Comp. Physiol. Psychol. 95 (1) (1981) 85–103.
[132] E. Dere, et al., The case for episodic memory in animals, Neurosci. Biobehav. Rev. 30 (8) (2006) 1206–1224.
[133] C.M.A. Pennartz, M. Farisco, K. Evers, Indicators and criteria of consciousness in animals and intelligent machines: an inside-out approach, Front. Syst. Neurosci. 13 (2019) 25.
[134] M. Hirose, K. Ogawa, Honda humanoid robots development, Philos. Trans. Math. Phys. Eng. Sci. 2007 (365) (1850) 11–19.
[135] R.L. Pitman, J.W. Durban, Cooperative hunting behavior, prey selectivity and prey handling by pack ice killer whales (Orcinus orca), type B, in Antarctic Peninsula waters, Mar. Mamm. Sci. 28 (1) (2012) 16–36.
[136] G.B. Young, The EEG in coma, J. Clin. Neurophysiol. 17 (5) (2000) 473–485.
[137] J.T. Giacino, et al., Comprehensive systematic review update summary: disorders of consciousness: report of the guideline development, dissemination, and implementation subcommittee of the American Academy of Neurology; the American Congress of Rehabilitation Medicine; and the National Institute on Disability, Independent Living, and Rehabilitation Research, Neurology 91 (10) (2018) 461–470.
[138] M. Massimini, et al., Breakdown of cortical effective connectivity during sleep, Science 309 (5744) (2005) 2228–2232.
[139] A.G. Casali, et al., A theoretically based index of consciousness independent of sensory processing and behavior, Sci. Transl. Med. 5 (198) (2013) 198ra105.
[140] M. Steriade, Grouping of brain rhythms in corticothalamic systems, Neuroscience 137 (4) (2006) 1087–1106.
[141] A.M. Owen, et al., Detecting awareness in the vegetative state, Science 313 (5792) (2006) 1402.
[142] L. Muckli, et al., Contextual feedback to superficial layers of V1, Curr. Biol. 25 (20) (2015) 2690–2695.
[143] M.J. Vansteensel, et al., Fully implanted brain-computer interface in a locked-in patient with ALS, N. Engl. J. Med. 375 (21) (2016) 2060–2066.
[144] B. Jarosiewicz, et al., Virtual typing by people with tetraplegia using a self-calibrating intracortical brain-computer interface, Sci. Transl. Med. 7 (313) (2015) 313ra179.
[145] B. Wodlinger, et al., Ten-dimensional anthropomorphic arm control in a human brain-machine interface: difficulties, solutions, and limitations, J. Neural Eng. 12 (1) (2015), 016011.
[146] D.A. Moses, et al., Neuroprosthesis for decoding speech in a paralyzed person with anarthria, N. Engl. J. Med. 385 (3) (2021) 217–227.
[147] S. Dora, C. Pennartz, S. Bohte, A deep predictive coding network for inferring hierarchical causes underlying sensory inputs, in: Proceedings of the International Conference on Artificial Neural Networks, Springer, 2018, pp. 457–467.
[148] D.J. Felleman, D.C. Van Essen, Distributed hierarchical processing in the primate cerebral cortex, Cereb. Cortex 1 (1) (1991) 1–47.
[149] C.C. Lee, J.A. Winer, Convergence of thalamic and cortical pathways in cat auditory cortex, Hear. Res. 274 (1–2) (2011) 85–94.
[150] S. Hsiao, Central mechanisms of tactile shape perception, Curr. Opin. Neurobiol. 18 (4) (2008) 418–424.
[151] C. Zelano, N. Sobel, Humans as an animal model for systems-level organization of olfaction, Neuron 48 (3) (2005) 431–454.
[152] G.A. Calvert, Crossmodal processing in the human brain: insights from functional neuroimaging studies, Cereb. Cortex 11 (12) (2001) 1110–1123.
[153] B.E. Stein, T.R. Stanford, Multisensory integration: current issues from the perspective of the single neuron, Nat. Rev. Neurosci. 9 (4) (2008) 255–266.