Neuroscience Research 201 (2024) 31–38

Contents lists available at ScienceDirect

Neuroscience Research

journal homepage: www.sciencedirect.com/journal/neuroscience-research

Review Article

The cognitive reality monitoring network and theories of consciousness

Aurelio Cortese [a,*], Mitsuo Kawato [a,b,**]

[a] Computational Neuroscience Labs, ATR Institute International, Kyoto 619-0228, Japan
[b] XNef, Kyoto 619-0288, Japan

ARTICLE INFO

Keywords:
Computational theory of consciousness
Metacognition
Cognitive reality monitoring network
Marr's levels
Representations

ABSTRACT

Theories of consciousness abound. However, it is difficult to arbitrate reliably among competing theories because they target different levels of neural and cognitive processing or anatomical loci, and only some were developed with computational models in mind. In particular, theories of consciousness need to fully address the three levels of understanding of the brain proposed by David Marr: computational theory, algorithms and hardware. Most major theories refer to only one or two levels, often indirectly. The cognitive reality monitoring network (CRMN) model is derived from computational theories of mixture-of-experts architecture, hierarchical reinforcement learning and generative/inference computing modules, addressing all three levels of understanding. A central feature of the CRMN is the mapping of a gating network onto the prefrontal cortex, making it a prime coding circuit involved in monitoring the accuracy of one's mental states and distinguishing them from external reality. Because the CRMN builds on the hierarchical and layer structure of the cerebral cortex, it may connect research and findings across species, further enabling concrete computational models of consciousness with new, explicitly testable hypotheses. In sum, we discuss how the CRMN model can help further our understanding of the nature and function of consciousness.

1. Introduction

The scientific study and understanding of consciousness is simultaneously a young research field and a long-standing philosophical ambition. There are many theories of consciousness [Seth & Bayne provide an excellent overview (Seth and Bayne, 2022)], which differ from one another at the level of definitions, the focus on global versus local scales, and the underlying neural circuits. Recent work has sought to define mechanisms and consensus for studying consciousness in biological and artificial systems (Butlin et al., 2023).

When we compare consciousness studies with other cognitive neuroscience fields, e.g. motor control, visual perception, memory, attention, and perceptual learning, we hit upon three significant differences in research.

The first is related to the definition of consciousness. Providing a clear definition of higher cognitive functions such as attention could be as difficult as for consciousness, yet there is consensus on attention (Lindsay, 2020; Petersen and Posner, 2012), and we can run animal experiments on attention with neural recordings and causal methods (Norman et al., 2021; Noudoost and Moore, 2011). It has proven challenging to run animal experiments on consciousness with the same level of rigour as with other cognitive functions. This difficulty primarily arises because consciousness has been classically described in relation to human subjective experience, and it is therefore hard to analyse in objective experiments applicable to other animals. We do not possess common experimental paradigms directly targeting consciousness in humans and experimental animals (but see (Birch et al., 2022)).

The second difficulty may come from failing to explicitly address David Marr's first level (Marr, 1982), the computational theory. What is consciousness for, how can its computation be made possible, and what are its inputs and outputs? To some extent, we can answer these questions for most brain functions, including motor control, visual perception, memory, attention, and perceptual learning. Although some theories suggest specific computations (see next section, *Definitions and theories of consciousness*), we are still missing a computational-level understanding of consciousness.

Finally, the third aspect concerns the computational models of David Marr's second and third levels. What are the algorithms, why could they be efficient with the computation aimed, and how are algorithms implemented in neural substrates? Some theories hint at possible

* Corresponding author.
** Corresponding author at: Computational Neuroscience Labs, ATR Institute International, Kyoto 619-0228, Japan
E-mail addresses: cortese.aurelio@gmail.com (A. Cortese), kawato@atr.jp (M. Kawato).

https://doi.org/10.1016/j.neures.2024.01.007
Received 28 December 2023; Accepted 17 January 2024
Available online 3 February 2024

algorithms or neural circuits as hardware. Global neural ignition across the brain (Dehaene et al., 1998; Mashour et al., 2020), local recurrrent computations in sensory circuits (Lamme, 2006, 2010), and self-monitoring of internal representations (Lau et al., 2022; Lau and Rosenthal, 2011), but remain relatively vague because they have no explicit link to the first level of understanding; the computational theory. Integrated Information Theory (IIT) instead has the form of a mathematical theory (Oizumi et al., 2014; Tononi, 2004) but is far removed from David Marr's three levels of understanding. IIT aims to achieve something similar to the Hamiltonian physics mechanics by the first principle. If it is valid, that will be transformative, but there have also been severe doubts about IIT, including invalid axioms, ad-hoc interpretations and changing assumptions (Bayne, 2018; Doerig et al., 2019; Lau, 2023; Morch, 2019).

The cognitive reality monitoring network (CRMN) is different from these previous theories of consciousness in its origins. It is a computational model of neural processing mapping to a hierarchical-modular reinforcement learning architecture (Kawato and Cortese, 2021). Thus, the computational theory of CRMN is explicit with respect to David Marr's first level: to maximise ecological fitness through reinforcement learning (i.e., maximising some expected reward signal[1]). The CRMN algorithm is a hierarchical and modular reinforcement learning algorithm with multiple paired generative and inference models. CRMN solves sensory and motor control, perceptual processing, objective maximisation, and metacognition as subordinate functions. At the hardware level, CRMN allocates paired generative and inference models to the layer structure of the cerebral cortex, reward prediction errors to the basal ganglia, and gains switching and cognitive reality monitoring to the PFC. A unique feature of CRMN is that consciousness was not the main objective of computational modelling. Metacognition and consciousness result from gating in hierarchical-modular reinforcement learning. That is, they both describe an inherent process of the model. As such, there is an explicit function for consciousness and metacognition in behavioural terms. In this sense, CRMN sits at the opposite extreme of IIT.

In the following sections, we will examine the current landscape of consciousness theories and how they map onto different levels of ontological representation. We will analyse the CRMN and its building blocks, providing an overview of its computations and how they can help our understanding of consciousness by linking it to other major consciousness theories. Within the CRMN framework, we will define the prerequisites for consciousness, such as a neural system with hierarchy, modularity, and information compression with representations chained over time. For an agent 'acting' in a dynamic, complex, non-stationary environment, consciousness as a coherent evolution in time of low-dimensional representations can provide an intrinsic tool for fast adaptive behaviour. Finally, we will discuss how the CRMN can be integrated into consciousness studies and help bridge the thorny question of how we can study theories of consciousness in animal models.

2. Definitions and theories of consciousness

What is consciousness, and how do we define it? Typically, we equate consciousness with subjective experience. That is, a system (human, animal, or otherwise artificial) has a conscious experience when there is "something it is like" for the system to be the subject of that experience (Butlin et al., 2023; Nagel, 1974).

The four main theories of consciousness (Michel et al., 2018) are the global neuronal workspace theory (GNW), local recurrent processing theory (RPT), higher-order theory (HOT), and integrated information theory (IIT). There are dozens more in practice, even though not all are as well known (Seth and Bayne, 2022). Some theories are widely accepted and seen as strong candidates to explain consciousness by the broader neuroscientific community (e.g., GNW, LR), while other theories are better known by the general public (e.g., IIT) and seen as more controversial in the field (Fleming et al., 2023; Lau, 2023; Michel et al., 2018).

We briefly highlight the key features of the four main theories. Table 1 refers to these four theories and the CRMN, introducing the specifics for each of Marr's three levels.

- GNW: global neuronal workspace (globalist view). Equates consciousness with conscious access, broadcasting information across cortical areas (Dehaene et al., 1998; Mashour et al., 2020). For the GNW theory, the hallmark of consciousness is a "global ignition phenomenon, the sudden, coherent, and exclusive activation of a subset of workspace neurons coding for the current conscious content" (quoted from Mashour et al., 2020). For consciousness to arise, the theory proposes that neural information processing has to travel from sensory areas to prefrontal cortices and reverberate across the brain's neural networks.
- RPT: recurrent processing theory (localist view). First-order representations generated through local recurrency are sufficient for consciousness (Lamme, 2010). Recurrent connections within and between sensory processing areas form the basis of conscious experience. Feedforward sweeps in recurrent networks will not be sufficient for consciousness, although they may be necessary.
- HOT: higher-order theory of consciousness (meta-representations). More than mere first-order representations are required for consciousness to arise, for a first-order sensory representation may happen nonconsciously and determine behavioural responses but is

Table 1
CRMN, mainstream theories of consciousness and how they map onto Marr's three levels of understanding. CRMN: Cognitive reality monitoring network, GNW: Global neuronal workspace theory, RPT: Recurrent processing theory, HOT: Higher-order theory, IIT: Integrated information theory. * Note that while both GNW and HOT have an entry in the computational theory field, neither strictly describes the computational theory.

	Computational theory	Representation & algorithm	Hardware implementation
CRMN	Maximise ecological fitness by reward-guided sensory-motor learning. For fast learning from small samples, dimension reduction is used through a divide-and-conquer strategy with hierarchy and modularity.	Fine to coarse representation hierarchy connected by generative/ inference model pairs. Hierarchical and modular reinforcement learning algorithm based on representation hierarchy and modularity.	Generative and inference models are implemented as feedback and feedforward connections within cortical layer circuits. Basal ganglia as a reinforcement learning hub, and PFC for selection and switching of modules.
GNW	Solving the frame problem, i.e., *	Coherent activation of workspace neurons: broadcasting of information across cortical areas	Connections between sensory areas to prefrontal cortices
RPT	-	First-order representations through local recurrency	Recurrent connections within local networks
HOT	distinguish reliable from unreliable mental representations for belief formation and future behaviour *	Second-order or meta representation in PFC and first order representations in the lower areas are matched	Lower areas and PFC, and neural connections between them
IIT	-	-	-

[1] By 'expected reward' here we do not mean simply a rewarding outcome. Rather, it is intended as the basis function of reinforcement learning, i.e. computational prediction function, with reward being the outcome conditional on the objective.

not enough for phenomenal conscious experience. A second-order representation, or 'meta-representation', is necessary. Consciousness requires inner awareness from monitoring first-order states (Brown et al., 2019; Lau et al., 2022; Lau and Rosenthal, 2011). Several versions of the theory differ based on the nature of these second-order representations and the connection between first and second-order representations.

- IIT: integrated information theory (theoretical-mathematical view). IIT identifies consciousness with a causal structure. Based on three axioms, IIT proposes that we can analytically calculate a measure (of integrated information) for any system from its causal structure. This measure directly reflects the 'amount' of consciousness in the system. IIT purports that recurrent systems are always conscious, while feedforward systems are never (Oizumi et al., 2014; Tononi, 2004).

These summary definitions indicate that these theories do not necessarily address or formalise an overarching computational problem. Instead, they are concerned with explaining the nature of consciousness or the minimal set of neural underpinnings. Thus, each theory takes a different approach or targets different levels of Marr's three levels of understanding of the brain (Table 1). To be fair, several researchers mention interesting characteristics of the computational aspects of GNW and HOT. In particular, Baars & Shanahan have argued that a global workspace architecture solves the frame problem in artificial intelligence (Shanahan and Baars, 2005). In the same way, for HOT, some researchers propose a computational-level goal: distinguish reliable from unreliable mental representations for belief formation and future behaviour (Gershman, 2019; Lau, 2019).

Some theories offer insight into the neural hardware (sensory vs associative or frontal areas of the brain, whole brain networks), others into the algorithms (recurrence). Because of these differences and a need for standard definitions and computational building blocks, arbitrating among theories or falsifying predictions can be difficult and problematic. The difficulty of arbitrating between theories lies clear in recent adversarial collaborations (Cogitate Consortium et al., 2023). While these efforts are great for the field, they face inherent problems when the theories tested need a robust common play level (in particular, neural substrate and computations/representations).

Therefore, how can we bridge theories of consciousness and start validating or falsifying specific claims? One way we are advocating here is to use a framework that maps onto Marr's three levels of analysis. Thus, to make within each theory explicit links with the computational theory (what computations underpin consciousness), the algorithms (what are the representations, what processes transform these representations, e.g., feedback for generative models, feedforward for inference models, prediction errors), and the neural hardware (e.g., cortical layers, circuits, regions).

3. The computational objective and prerequisites of consciousness

What is consciousness' function for survival? What is the input/output of conscious processes? What is the mechanism for its usefulness? These questions directly relate to Marr's levels of understanding in that they lead us to think about the underlying computational theory.

We first suggest that consciousness is critical to creating relational content. That is to say, mental or neural content that is compositional and relational, as in an episode linking together different streams of information (e.g., context, actions, time, perception, etc.), combining sensory dimensions, cognitive processes and levels of abstraction. At its core, and as advocated before (Bengio, 2017; Cortese et al., 2019), consciousness could generate a low- or even uni-dimensional data point from very high-dimensional inputs. This way, over time, the brain can create a time series of low-dimensional data points (snapshots), amalgamating different sources of information into one coherent dimension.

One thing that results from this view is the actionability of such a construct. In short, a time series of low-dimensional representations can support complex behaviours without needing nonlinear transformations of high-dimensional data. However, to perform efficiently, and in light of recent work, the brain likely operates offline and online systems to manipulate the low-dimensional time series.

The online system can respond in real time to sensory information and the behavioural context in a changing, dynamic world. Sensorimotor, attention and executive function networks could support such an online system (Vidaurre et al., 2017). Instead, the offline system can simulate behavioural trajectories (in the past or future) for learning, optimal control, and adaptive strategies. Replay could be the neuro-computational mechanism enabling precisely this, generating simulations and creating connections between events, goals and contexts. Forward and backward replay has been shown across species, including pioneering work in mice and later in humans (Gupta et al., 2010; Schapiro et al., 2018; Wang et al., 2020). The default mode network may play an essential function in replay (Kaefer et al., 2022). Off-policy reinforcement-learning simulations could depend on the interplay between the default mode network and PFC. VmPFC is probably the gating network for this offline system, and DMN is the main computing body.

For the online and offline systems to interact efficiently, they would need a neural and representational subspace with greatly reduced information content (with very few dimensions). Consciousness could be the key. In short, consciousness, by creating low-dimensional time series for the coherent evolution in time of internal representations, may thus generate the abstract subspace for offline and online systems to cooperate smoothly. Note that extreme dimension reduction can be achieved through symbolic representations, but the CRMN is agnostic. In the CRMN, low dimensional representations can be symbolic or analogue in nature, probably depending on the content [i.e., perception probably is not symbolic (Beck, 2019), but other conscious contents may be].

At this stage, we can now summarise consciousness' prerequisites. Consciousness depends on hierarchy, modularity, and information compression (in the extreme, uni-dimensional representations). In addition, a conscious agent needs to be acting in a dynamic, complex, non-stationary environment, which leads to creating time series (episodes).

4. The architecture and computations of the Cognitive Reality Monitoring Network (CRMN)

With this background, approach and goals in mind, the CRMN can provide a new opportunity to study and understand consciousness. Contrary to most theories of consciousness, the CRMN was not developed as a primary model of consciousness or to study consciousness per se. Instead, the CRMN stems from separate lines of research in artificial intelligence and computational neuroscience (Kawato and Cortese, 2021). It was developed as a model to explain some fundamental features of neural computations in the brain, integrating reinforcement learning theory and information abstraction (Fig. 1). However, metacognition and consciousness are naturally accounted for by the CRMN and play a central role in its computations, as we will discuss further below.

The CRMN is a mixture-of-experts architecture (Haruno et al., 2001; Jacobs et al., 1991; Sugimoto et al., 2012). Each expert (i.e., functional modules, the computation unit of the CRMN) is a generative-inference (forward-inverse) model conjugate pair (Kawato et al., 1987; Kawato and Cortese, 2021). The forward model is a generative model of the rawer representation (the computation flow goes from higher to lower areas). In contrast, the inverse model computes an analytical, one-shot estimation of higher-order representations (the computation flow goes from lower to higher areas). The inverse model computation is fast, as it can be operated in a single forward sweep because it just has to approximate ground truth information about latent variables. The forward model computation is slow, and recurrence is needed to converge

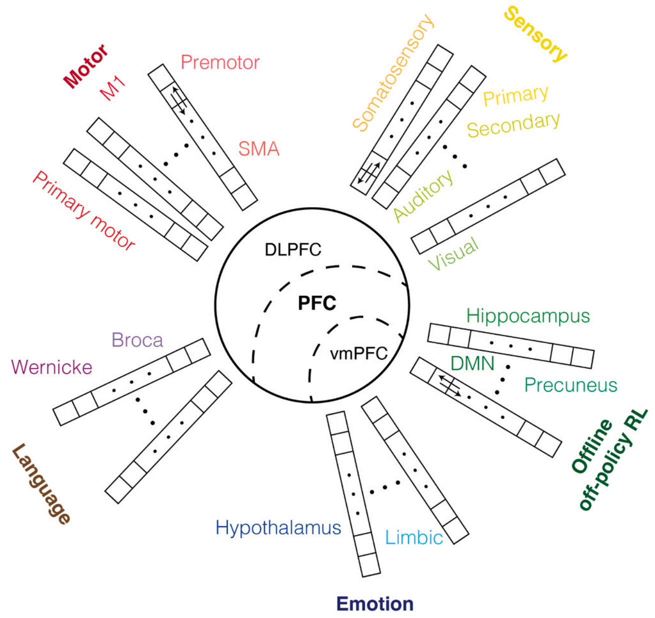

Fig. 1. Neural substrates and conceptual architecture of the neural CRMN. Various online and offline module computations related to consciousness map onto the brain's separate functional and anatomical regions. The central sphere represents the broader PFC, with dotted lines delineating the DLPFC and the vmPFC. Vector lines with squares and dots represent each hierarchy within the CRMN. For instance, in the Sensory cortex, the somatosensory hierarchy represents the multiple levels of abstraction in computation from the lowest to the highest, closest to the PFC. Parallel hierarchies represent the idea of modularity. Parallel arrows (upward and downward) between two boxes in some spokes represent inference/generative computations. Consciousness results from computations across all modules and hierarchies gated through the PFC.

to a stable solution because it has to generate a complex, multidimensional signal. This generative-inference model computation is related to predictive coding (Rao and Ballard, 1999), a mechanism proposed to explain neural activity's logic in the hierarchical visual stream. In the predictive coding model, a prediction error between the generative model output and lower-level representation is filtered by the inference model and sent back to the higher level, which is the same algorithm as that of forward-inverse computation (Kawato et al., 1993).

Note here that we are not advocating a static, fixed view of the brain segregated into many small neural modules independent of each other. Rather, these forward-inverse pairs are *functional* modules that overlap at the neural level. They are flexible enough to recruit new and forego old units, with connections changing based on experience and computational/behavioural demands via synaptic update mechanisms.

The basic representations in the CRMN are the states (k_{ik}), single representations of sensory (or mnemonic) information, and the content of a single module. State representations change across modules, increasing in abstraction as one moves from modules closer to the sensors (early sensory cortices) up to higher sensory cortices, associative cortices, and the prefrontal cortex.

A fundamental computation of the CRMN is the mismatch between the forward and inverse conjugate pairs. Modules compute this mismatch at any given time. In our original formulation of the CRMN model, we termed this a sensory prediction error in the sensory cortex and a motor prediction error in the motor cortex (Kawato and Cortese, 2021). However, this error can also be a memory prediction error arising from internal memory signals or a reward prediction error in the basal ganglia. The different prediction errors are propagated across modules towards the higher levels of the cortical hierarchy (i.e., the frontal part of the brain) to a gating network that computes general 'cognitive' prediction errors (Kawato and Cortese, 2021). We called these general error signals cognitive prediction errors because they incorporate sensory, motor, memory, and reward prediction errors. Feedback neural connections backpropagate error signals to lower modules, signalling matches or mismatches, and thus, how computations should be updated accordingly. Skip connections, popular in artificial neural networks (Emin Orhan and Pitkow, 2017; He et al., 2015), are present in the brain, too, in the form of direct pathways from sensory areas to PFC (Zikopoulos and Barbas, 2007). Direct connections can propagate mismatch signals to the relevant modules or gating network more rapidly.

The CRMN incorporates hierarchical and modular reinforcement learning. Modules replicate along the cortical hierarchy, and abstraction increases as one moves from sensory areas to associative and prefrontal cortices (Kawato and Cortese, 2021). Early works have demonstrated that the basal ganglia, hosting a range of reinforcement learning processes, are connected to cortical areas through parallel and hierarchical loops (Draganski et al., 2008; Graybiel et al., 1994; Haruno and Kawato, 2006; Tanaka et al., 2004). Basal ganglia connectivity patterns reflect anatomical gradients across neocortical regions (Choi et al., 2018; Jarbo and Verstynen, 2015). Thus, it is hierarchical because different loops carry information at different levels of abstraction and complexity, often in a nested manner. Information is recombined and reduced to the few maximally relevant dimensions for the ongoing task/goal through selection, transformation and abstraction across cortical areas. The system is also modular because different loops and cortical subregions map onto anatomically or functionally defined computational units (e.g., the forward-inverse models' pair)—for instance, within the visual areas, different subregions process colour, motion, or shapes. Note that most reinforcement learning processes should be unconscious. In reinforcement learning, consciousness should be necessary only when the problem is vast and involves many modalities and hierarchies.

Finally, the CRMN maps onto GAN models [generative adversarial networks (Goodfellow et al., 2014)] and their extension to theories of consciousness (Gershman, 2019; Lau, 2019; Lau et al., 2022). GANs are a class of artificial neural networks composed of two sub-networks harbouring opposing goals. A 'discriminator' network that aims to tell the difference between real and fake/synthetic data. Instead, a 'generator' network aims to generate the synthetic data (e.g., images) that can trick the discriminator into believing it is real. Based on this framework, Lau and Gershman (Gershman, 2019; Lau, 2019) independently proposed that the brain and specifically the prefrontal cortex may operate as a discriminator, i.e., continuously comparing and discriminating endogenously generated neural activity against neural activity triggered by external stimuli. In this view, consciousness emerges if the prefrontal discriminator judges the first-order representation as "real". These ideas are tightly related to the perceptual reality monitoring theory (Simons et al., 2017). Conscious perception occurs if there is a relevant high-order representation with the content that a particular first-order perceptual representation is a reliable reflection of the external world. In terms of consciousness, this means the gating network harbours a re-representation in frontal areas—within the gating network, of a maximally abstract representation of a sensory, motor or memory input.

5. How consciousness and metacognition map to the CRMN

While prediction errors per se do not directly relate to consciousness, it is the computation by the CRMN gating network on the prediction errors that should determine the content of consciousness. The gating network computes a responsibility signal λ_{ik} as a softmax of the cognitive prediction error for each module and hierarchy (Kawato and Cortese, 2021). Responsibility signal λ_{ik} represents the local strength of the internal evidence (i.e., the matching of prediction and incoming information) and directly reflects the uncertainty around these representations. The responsibility signal priors ($\widehat{\lambda}_{ik}$) are also locally implemented. Priors influence the computation of responsibility signals λ_{ik} by providing a scaling factor. They arise from long timescales, such as genetics and evolution, and short timescales, such as learning and adaptation. Metacognition reflects the responsibility signals within modules' hierarchies, thus potentially giving rise to multiple signals because responsibility signals are multi-dimensional. A key prediction of this formalisation is that one can have simultaneous metacognition of different contents, i.e., metacognition is expressed as a multi-dimensional vector of responsibility signals. In other words, metacognition is not a serial process but is parallel and multiple metacognitive processes can operate simultaneously. If so, a behavioural prediction might be that subjects should be as metacognitively efficient (and as fast) for judging their confidence in multiple decisions (e.g., about several stimulus features) compared to a single decision.

In the CRMN, consciousness instead is directly related to the entropy (S) computed from responsibility signals *across* all modules' hierarchies, thus giving rise to a single, unitary construct (Fig. 1). This construct is the content of consciousness—i.e., qualia (Kanai and Tsuchiya, 2012), which is now computationally and mathematically defined. Computed across all modules and hierarchies, the entropy indexes whether the agent is conscious and possibly the level of consciousness. Low entropy will reflect that a given module, or a minimal set of modules, has high responsibility signals, and the agent will be conscious of the relevant modules' content. Consciousness constantly evolves within a low-dimensional (unitary) content space, one dimension at a time. Conscious content is a weighted summation of a very small number of active modules with large responsibility signals.

Binocular rivalry is a good example supporting the single content view of consciousness –in binocular rivalry, two different stimuli are projected onto the two eyes separately. The content of consciousness alternates from one stimulus to the other but seldom results in an actual merging of the two stimuli (Tong et al., 2006). Others have made a strong case for the unitary content of consciousness in related tasks (Kapoor et al., 2022).

The case for metacognition is more challenging to test because we have to demonstrate that one has simultaneous metacognitive access to two or more information streams. The brain has related but separated metacognitive mechanisms for different domains, such as memory and visual perception (McCurdy et al., 2013; Morales et al., 2018). Thus, there is evidence for domain-general and domain-specific processes. The CRMN can naturally accommodate and expand these findings thanks to its hierarchical and modular architecture. Responsibility signals within a cortical hierarchy will give rise to a confidence/metacognition signal. This confidence maps to a particular domain. For, confidence about colour perception might differ from confidence about motion perception. At a more general level, we can have confidence/metacognition about perception, separated from confidence/metacognition about memory. The mapping here happens at a higher hierarchy level within the gating network. Finally, the CRMN posits a full, domain-general metacognitive signal, encompassing all lower, more specific, levels. A second prediction by the CRMN is that these multiple metacognitive signals should be correlated yet distinct. More specific metacognitive signals may exist in areas related to that domain's computations. In contrast, the more general metacognitive signal centres on the gating network in the PFC.

Future work should test and dissect these predictions; we highlight some ways forward in the section 'future experiments'. One may wonder how the CRMN model accounts for sensory processing that seldom reaches consciousness, such as the dorsal stream in visual processing (Goodale and Milner, 1992). While metacognition can be about local representations and responsibility signals, it is still computed at the highest levels of the hierarchy –top layers and within the CRMN. Furthermore, the brain's anatomy and cortical connections strongly constrain the CRMN. Because dorsal visual areas provide inputs to occipito-parietal areas and then directly to motor control regions, processing in the dorsal stream can bypass the CRMN (e.g., PFC areas) and thus consciousness. This does not mean we cannot be conscious of, e.g. motion information, but it can generate meaningful behaviour without going through conscious states. Instead, the ventral stream is better connected to visual and semantic memory areas, planning and offline simulation (Milner, 2012), such as the hippocampus and, ultimately, the prefrontal cortex, with most processing thus directly reaching the CRMN.

We can make several predictions concerning consciousness with the above definitions of the CRMN variables.

(1) consciousness is associated with low entropy S. The consequence is a brain state in which specific cognitive modules are active, sensory information is re-represented at multiple levels of abstraction, and cortical-subcortical loops are active—such definition maps to the GNW theory and the HOT of consciousness. In short, we have information broadcast across selected (low entropy) hierarchies/modules and re-representation of states from lower to higher levels. However, a global broadcast is unnecessary for CRMN to display consciousness and solve complex RL problems efficiently. Only when the PFC selects a specific module and hierarchy and bidirectional communication is secured between the PFC and the selected one is sufficient in CRMN. This is one of the main differences between GNW and CRMN.

(2) metacognition is distributed across cognitive modules. That is, metacognition is not unitary (unlike consciousness), and one can be metacognitively aware of more than one cognitive module/content at a time. In addition, this also means there is (theoretically) an entire hierarchy of metacognitive processes and representations.

(3) metacognition can exist even in the absence of consciousness (meta-d' > d') (Charles et al., 2017, 2013; Cortese et al., 2020). This statement might sound surprising or controversial, given that metacognition intimately relates to or directly reflects conscious experience. There is also evidence that metacognition is "consciousness-selective", such that metacognitive inefficiency derives from unconscious sensory activity (Michel, 2022). Theoretical work with signal detection theory can accommodate these conflicting findings. If metacognition builds on decision-congruent evidence alone in a context of unequal variance, it is possible to obtain meta-d' > d' (Miyoshi and Lau, 2020). The CRMN assigns different computational entities, responsibility signals and their entropy to metacognition and consciousness separately.

6. Connecting findings in consciousness science across species

The absence of a firm reliance on the human subjective experience to explain consciousness and its explicit development over Marr's three levels of understanding makes the CRMN model more flexible in its application and comparison across species. While most theories of consciousness are concerned about (or were developed around) the human subjective experience, other animals, too, presumably have conscious experiences of some form or degree (Barron and Klein, 2016; Frith, 2019; Gutfreund, 2017). This statement may sound obvious to many, but because animals cannot unequivocally report their conscious experiences, the quality or nature of their conscious experiences remains a mystery. Thus, it has proven challenging to carry out consciousness research with animal models [but see (Boly et al., 2013) and (Birch et al., 2022) for fruitful overviews].

The primary reason is the absence of self-reports. While no-report paradigms (Tsuchiya et al., 2015) have partly addressed this issue with tasks that can be equally administered to humans and other animals (Hesse and Tsao, 2020; Kapoor et al., 2022), some have argued that despite their promises, no-report paradigms can be as confounded as standard report experiments (Block, 2019; Overgaard and Fazekas, 2016), and in the end it still depends on how we interpret findings (Panagiotaropoulos et al., 2020).

A second contentious point is that we often equate intelligence with consciousness. But that is a misappropriation of what consciousness is—pure, simple experience, the feeling of what it is like to be something. Of course, consciousness probably has meaningful functions for intelligence (Bengio, 2017; Cortese et al., 2019; Goyal and Bengio, 2020; Kawato and Cortese, 2021), and the two are (highly) correlated, but this need not necessarily be a fundamental requirement.

The CRMN can be helpful in this context. In particular, as highlighted in the previous section, the CRMN describes a specific set of computations. Some, such as the forward-inverse model pairs, are well established (Haruno et al., 2001; Kawato et al., 1987; Wolpert and Kawato, 1998); others, such as cognitive prediction errors, responsibility signals or concurrent metacognitive signals, will need to be explored in future research. Animal models will bring crucial insight. We will next discuss possible experimental paradigms to disentangle some of the issues.

7. Future experiments

Considering the CRMN and the importance of its constituting parts to consciousness, how can we experimentally calculate the model's variables, especially responsibility signals and entropy, from empirical data (e.g., electrophysiology or neuroimaging)? There are three obstacles to answering this question: in general, we do not have direct access to the generative and inverse model pairs in the brain (except perhaps in specific cases, i.e. in the visual stream); we do not know the exact boundaries of the functional units/modules of the CRMN in the brain; and we do not yet possess a proper description of the function governing the abstraction process (e.g., from early sensory areas to associative and prefrontal cortices). That said, we believe it will be possible to overcome these difficulties by combining computational model-based analyses with neuroimaging. One can compute responsibility signals and entropy via computational models applied to multi-dimensional tasks / behavioural responses. Below, we propose a few ways forward with a programme of research experiments.

(1) Behavioural experiments.

Use ambiguous stimuli created as a combination of weakly coherent features, such as colour, motion, and size, or unimodal (one feature only). Participants view the ambiguous stimuli and make detection/discrimination choices, confidence reports, and/or free visual awareness reports. In the CRMN, the spontaneous emergence of visual awareness for coherent multi-modal stimuli would require further computation of a compressed multi-dimensional representation for small entropy cases. It should thus differ from an unimodal detection task, which does not require entropy. However, to be clear, CRMN does not oppose unconscious cross-modal binding (Scott et al., 2018). The CRMN nevertheless predicts the necessity of conscious cross-modal binding if RL problems are complicated and require abstract and low-dimensional RL states binding multiple modalities.

(2) fMRI experiments.

From the above behavioural experiments, we can decode PFC multivoxel patterns to predict the above three factors (discrimination, detection, and visual awareness). We expect differences between unimodal detection and mixed conscious visual awareness. Computational simulations of multi-modality RL tasks by CRMN (in the simplest case, a mixture-of-experts RL) can calculate trial-by-trial responsibility signals and entropy, which we can then decode from fMRI multi-voxel patterns.

(3) Decoded neurofeedback experiments.

Building on a previous study (Knotts et al., 2019), we try to increase and decrease entropy (or decoded confidence) in PFC and examine the decrease and increase of false alarm events. Unlike the previous study, which simultaneously targeted PFC and V1, we expect this more targeted approach (neurofeedback of PFC signals alone) to lead to apparent changes in false alarms. In addition, based on the hypothesis that there may be a common metacognitive basis for detection choices and confidence reports, manipulating decoded confidence could change the internal threshold for detection.

Finally, we could manipulate information transmission (Amano et al., 2016; Shibata et al., 2011) between V1 and PFC regarding detecting a given visual feature (e.g., motion). Using multivariate pattern analysis, we can measure the degree of coupling between two regions, conditional on a specific informational content (representation or likelihood thereof). We may increase false alarms and decrease the detection threshold in perceptual decision-making tasks by manipulating information transmission. This novel experiment mixes decoded neurofeedback and functional connectivity neurofeedback (Cortese et al., 2021; Megumi et al., 2015; Shibata et al., 2011; Watanabe et al., 2017) in its methodology, which will be interesting in and by itself.

8. Conclusion and outstanding questions

In this paper, we have briefly introduced the main theories of consciousness, highlighting how their relative distance from a complete account, according to David Marr's three levels of understanding (Table 1), may limit their explanatory power. We have sought to introduce arguments in favour of a more holistic approach that involves all three levels, from the computational theory (i.e., the computational objective of consciousness), hardware (the neural underpinnings), and software (the representations and algorithms). In doing so, we have discussed how the CRMN, a computational model of sensory, cognitive and learning processes in the brain, can formally account for consciousness and metacognition while simultaneously fulfilling Marr's three levels of understanding.

Our description and discussion are only partial and leave several questions unanswered. Here, we highlight a few we deem particularly significant.

Is the PFC necessary for consciousness? How can we solve the existing debate around PFC function in consciousness? Previous work has seen two main stances, with some theories ascribing a central role to PFC (HOT, GNW), while others less so (LR, IIT). The CRMN makes a strong prediction that the PFC is critical for consciousness by being the site of the gating network where responsibility signals converge, and computations over internal distributions take place.

Can prediction errors directly cause consciousness? C.f. Jun Tani's work on surprise signals and consciousness (Tani, 2016). The CRMN postulates that a significant mismatch between forward and inverse models will result in a large prediction error, leading to small responsibility signals and, thus, high entropy, meaning no conscious access. However, a large prediction error and a specific behavioural demand could lead the brain to focus resources on those modules, resulting in converging updates that will quickly minimise the error and thus entropy, quickly resulting in a conscious percept.

Can we find direct evidence for high-level *re*-representations (in the PFC) of low-level sensory coding? How would these representations appear? Fleming and Lau discussed this problem to some extent (Fleming, 2020; Lau et al., 2022). The fact that PFC neurons code sensory information is now well known, as consistently reported in electrophysiology experiments (Kapoor et al., 2022; Mante et al., 2013) as well as in humans with neuroimaging (Cortese et al., 2016; Jung et al., 2018; Weilnhammer et al., 2021). However, how these representations relate to lower-level sensory representations in sensory cortices remains unknown. Multivariate pattern analyses linking the representation content in different areas, combined with dimensionality analyses, might help answer this question.

Does the CRMN generalise to other biological organisms that display complex behaviours but have a vastly different neural system? Octopuses, for instance, have distributed neural systems, which are much less hierarchical than human or primate brains (Gray, 1970; Young, 1971). Yet, they appear to have many functions seen in other complex animals, such as two stages of sleep and even dreams (Pophale et al., 2023). While the CRMN builds on solid assumptions about the hierarchy of computational modules, the formulation of the model should be general enough to apply readily to numerous neural architectures. Future simulation work could establish the nature of these architectures and the requirements for consciousness (in computational terms, i.e., responsibility signals and entropy).

Finally, what role does the cerebellum play? Contrary to the classical view of the cerebellum as solely related to motor control, more than a decade of work has now shown its implication in a variety of cognitive processes, from attention states to reinforcement learning and social behaviours (Brissenden et al., 2021; Kawato et al., 2021; Pu et al., 2020; Sendhilnathan et al., 2020), and in particular to sequencing (Tedesco et al., 2011). Further, the cerebellum has been highlighted as a unique computing core for dimensionality control and modular reinforcement learning (Hoang et al., 2023a, 2023b, 2020; Rotondo et al., 2023).

To conclude, we have discussed how the CRMN can bridge computational theory and consciousness research, linking to Marr's three levels of understanding. In so doing, we provided what we hope is a new view on the computational objective of consciousness, its neural underpinnings and future avenues of investigation.

CRediT authorship contribution statement

Cortese Aurelio: Conceptualization, Funding acquisition, Project administration, Visualization, Writing – original draft, Writing – review & editing. **Kawato Mitsuo:** Conceptualization, Funding acquisition, Project administration, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

None.

Acknowledgements

A.C. and M.K. are supported by grants from the Japan Science and Technology Agency (ERATO Ikegaya brain-AI fusion grant number JPJMER1801), JSPS KAKENHI (grant number JP22H05156), and the Agency for Technology, Labour and Innovation (ATLA, grant number JP004596).

References

Amano, K., Shibata, K., Kawato, M., Sasaki, Y., Watanabe, T., 2016. Learning to associate orientation with color in early visual areas by associative decoded fMRI neurofeedback. Curr. Biol. 26, 1861–1866.

Barron, A.B., Klein, C., 2016. What insects can tell us about the origins of consciousness. Proc. Natl. Acad. Sci. U. S. A. 113, 4900–4908.

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

Beck, J., 2019. Perception is Analog: The Argument from Weber's Law. J. Philos. 116, 319–349.

Bengio, Y., 2017, The Consciousness Prior. arXiv [cs.LG].

Birch, J., Broom, D.M., Browning, H., Crump, A., Ginsburg, S., Halina, M., Harrison, D., Jablonka, E., Lee, A.Y., Kammerer, F., Klein, C., Lamme, V., Michel, M., Wemelsfelder, F., Zacks, O., 2022. How should we study animal consciousness scientifically? J. Conscious. Stud.

Block, N., 2019. What is wrong with the no-report paradigm and how to fix it. Trends Cogn. Sci. 23, 1003–1013.

Boly, M., Seth, A., Wilke, M., Ingmundson, P., Baars, B., Laureys, S., Edelman, D., Tsuchiya, N., 2013. Consciousness in humans and non-human animals: recent advances and future directions. Front. Psychol. 4 https://doi.org/10.3389/fpsyg.2013.00625.

Brissenden, J.A., Tobyne, S.M., Halko, M.A., Somers, D.C., 2021. Stimulus-Specific Visual Working Memory Representations in Human Cerebellar Lobule VIIb/VIIIa. J. Neurosci. 41, 1033–1045.

Brown, R., Lau, H., LeDoux, J.E., 2019. Understanding the Higher-Order Approach to Consciousness. Trends Cogn. Sci. 23, 754–768.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M.A.K., Schwitzgebel, E., Simon, J., VanRullen, R., 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv [cs.AI].

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

Charles, L., Gaillard, R., Amado, I., Krebs, M.-O., Bendjemaa, N., Dehaene, S., 2017. Conscious and unconscious performance monitoring: Evidence from patients with schizophrenia. Neuroimage 144, 153–163.

Choi, E., Drayna, G.K., Badre, D., 2018. Evidence for a Functional Hierarchy of Association. Netw. J. Cogn. Neurosci. 1–17.

Cogitate Consortium , Ferrante, O. , Gorska-Klimowska, U. , Henin, S. , Hirschhorn, R. , Khalaf, A. , Lepauvre, A. , Liu, L. , Richter, D. , Vidal, Y. , Bonacchi, N. , Brown, T. , Sripad, P. , Armendariz, M. , Bendtz, K. , Ghafari, T. , Hetenyi, D. , Jeschke, J. , Kozma, C. , Mazumder, D.R. , Montenegro, S. , Seedat, A. , Sharafeldin, A. , Yang, S. , Baillet, S. , Chalmers, D.J. , Cichy, R.M. , Fallon, F. , Panagiotaropoulos, T.I. , Blumenfeld, H. , de Lange, F.P. , Devore, S. , Jensen, O. , Kreiman, G. , Luo, H. , Boly, M. , Dehaene, S. , Koch, C. , Tononi, G. , Pitts, M. , Mudrik, L. , Melloni, L. , 2023, An adversarial collaboration to critically evaluate theories of consciousness. bioRxiv. https://doi.org/10.1101/2023.06.23.546249.

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

Cortese, A., De Martino, B., Kawato, M., 2019. The neural and cognitive architecture for learning from a small sample. Curr. Opin. Neurobiol. 55, 133–141.

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

Cortese, A., Tanaka, S.C., Amano, K., Koizumi, A., Lau, H., Sasaki, Y., Shibata, K., Taschereau-Dumouchel, V., Watanabe, T., Kawato, M., 2021. The DecNef collection, fMRI data from closed-loop decoded neurofeedback experiments. Sci. Data 8, 65.

Dehaene, S., Kerszberg, M., Changeux, J., 1998. A neuronal model of a global workspace in effortful cognitive tasks. P Natl. Acad. Sci. Usa 95, 14529–14534.

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

Draganski, B., Kherif, F., Klöppel, S., Cook, P., Alexander, D., Parker, G., Deichmann, R., Ashburner, J., Frackowiak, R., 2008. Evidence for Segregated and Integrative Connectivity Patterns in the Human Basal Ganglia. J. Neurosci. 28, 7143–7152.

Emin Orhan, A., Pitkow, X., 2017, Skip Connections Eliminate Singularities. arXiv [cs. NE].

Fleming, S.M., 2020. Awareness as inference in a higher-order state space. Neurosci. Conscious 2020, niz020.

Fleming, S.M., Frith, C., Goodale, M., Lau, H., LeDoux, J.E., Lee, A.L.F., Michel, M., Owen, A., Peters, M.A.K., Slagter, H.A., 2023. The Integrated Information Theory of Consciousness as Pseudoscience. psyArXiv. https://doi.org/10.31234/osf.io/zsr78.

Frith, C.D., 2019. The neural basis of consciousness. Psychol. Med. 1–13.

Gershman, S.J., 2019. The Generative Adversarial Brain. Front Artif. Intell. 2, 18.

Goodale, M.A., Milner, A.D., 1992. Separate visual pathways for perception and action. Trends Neurosci. 15, 20–25.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Nets, in: Advances in Neural Information Processing Systems 27. Presented at the Neural Information Processing, Curran Associates, Inc.

Goyal, A., Bengio, Y., 2020, Inductive Biases for Deep Learning of Higher-Level Cognition. arXiv [cs.LG].

Gray, E.G., 1970. The fine structure of the vertical lobe of octopus brain. Philos. Trans. R. Soc. Lond. B Biol. Sci. 258, 379–394.

Graybiel, A.M., Aosaki, T., Flaherty, A.W., Kimura, M., 1994. The Basal Ganglia and Adaptive Motor Control. Science 265.

Gupta, A., Meer, M., Touretzky, D., Redish, A., 2010. Hippocampal Replay Is Not a Simple Function of Experience. Neuron 65, 695–705.

Gutfreund, Y., 2017. The Neuroethological Paradox of Animal Consciousness. Trends Neurosci. 40, 196–199.

Haruno, M., Kawato, M., 2006. Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. Neural Netw. 19, 1242–1254.

Haruno, M., Wolpert, D.M., Kawato, M., 2001. Mosaic model for sensorimotor learning and control. Neural Comput. 13, 2201–2220.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. arXiv [cs. CV.

Hesse, J.K., Tsao, D.Y., 2020. A new no-report paradigm reveals that face cells encode both consciously perceived and suppressed stimuli. Elife 9, e58360.

Hoang, H., Tsutsumi, S., Matsuzaki, M., Kano, M., Toyama, K., Kitamura, K., Kawato, M., 2023b. Predictive reward-prediction errors of climbing fiber inputs integrate modular reinforcement learning with supervised learning. bioRxiv. https://doi.org/10.1101/2023.03.13.532374.

Hoang, H., Lang, E.J., Hirata, Y., Tokuda, I.T., Aihara, K., Toyama, K., Kawato, M., Schweighofer, N., 2020. Electrical coupling controls dimensionality and chaotic firing of inferior olive neurons. PLoS Comput. Biol. 16, e1008075.

Hoang, H., Tsutsumi, S., Matsuzaki, M., Kano, M., Kawato, M., Kitamura, K., Toyama, K., 2023a. Dynamic organization of cerebellar climbing fiber response and synchrony in multiple functional components reduces dimensions for reinforcement learning. Elife 12. https://doi.org/10.7554/eLife.86340.

Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E., 1991. Adaptive Mixtures of Local Experts. Neural Comput. 3, 79–87.

Jarbo, K., Verstynen, T., 2015. Converging structural and functional connectivity of orbitofrontal, dorsolateral prefrontal, and posterior parietal cortex in the human striatum. J. Neurosci. 35, 3865–3878.

Jung, Y., Larsen, B., Walther, D.B., 2018. Modality-independent coding of scene categories in prefrontal cortex. J. Neurosci. https://doi.org/10.1523/JNEUROSCI.0272-18.2018.

Kaefer, K., Stella, F., McNaughton, B.L., Battaglia, F.P., 2022. Replay, the default mode network and the cascaded memory systems model. Nat. Rev. Neurosci. https://doi.org/10.1038/s41583-022-00620-6.

Kanai, R., Tsuchiya, N., 2012. Qualia. Curr. Biol. 22, R392–R396.

Kapoor, V., Dwarakanath, A., Safavi, S., Werner, J., Besserve, M., Panagiotaropoulos, T. I., Logothetis, N.K., 2022. Decoding internally generated transitions of conscious contents in the prefrontal cortex without subjective reports. Nat. Commun. 13, 1535.

Kawato, M., Cortese, A., 2021. From internal models toward metacognitive AI. Biol. Cybern. 115, 415–430.

Kawato, M., Furukawa, K., Suzuki, R., 1987. A hierarchical neural-network model for control and learning of voluntary movement. Biol. Cybern. 57, 169–185.

Kawato, M., Hayakawa, H., Inui, T., 1993. A forward-inverse optics model of reciprocal connections between visual cortical areas. Netw.: Comput. Neural Syst. 4, 415–422.

Kawato, M., Ohmae, S., Hoang, H., Sanger, T., 2021. 50 Years Since the Marr, Ito, and Albus Models of the Cerebellum. Neuroscience 462, 151–174.

Knotts, J.D., Cortese, A., Taschereau-Dumouchel, V., Kawato, M., Lau, H., 2019, Multivoxel patterns for perceptual confidence are associated with false color detection. bioRxiv. https://doi.org/10.1101/735084.

Lamme, V., 2006. Towards a true neural stance on consciousness. Trends Cogn. Sci. 10, 494–501.

Lamme, V.A.F., 2010. How neuroscience will change our view on consciousness. Cogn. Neurosci. 1, 204–220.

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

Lau, H., 2023, Where is the "posterior hot zone"? Open Review of Ferrante et al (2023): "An Adversarial Collaboration to Critically Evaluate Theories of Consciousness" (by the ARC-Cogitate Consortium). psyArXiv. https://doi.org/10.31234/osf.io/93ufe.

Lau, H., Rosenthal, D., 2011. Empirical support for higher-order theories of conscious awareness. Trends Cogn. Sci. 15, 365373.

Lau, H., Michel, M., LeDoux, J.E., Fleming, S.M., 2022. The mnemonic basis of subjective experience. Nat. Rev. Psychol. 1, 10.

Lindsay, G.W., 2020. Attention in Psychology, Neuroscience, and Machine Learning. Front. Comput. Neurosci. 14, 29.

Mante, V., Sussillo, D., Shenoy, K.V., Newsome, W.T., 2013. Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature 503, 78–84.

Marr, D., 1982, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. New York Freeman.

Mashour, G.A., Roelfsema, P., Changeux, J.-P., Dehaene, S., 2020. Conscious Processing and the Global Neuronal Workspace Hypothesis. Neuron 105, 776–798.

McCurdy, L.Y., Maniscalco, B., Metcalfe, J., Liu, K.Y., de Lange, F.P., Lau, H., 2013. Anatomical Coupling between Distinct Metacognitive Systems for Memory and Visual Perception. J. Neurosci. 33, 1897–1906.

Megumi, F., Yamashita, A., Kawato, M., Imamizu, H., 2015. Functional MRI neurofeedback training on connectivity between two regions induces long-lasting changes in intrinsic functional network. Front. Hum. Neurosci. 9, 160.

Michel, M., 2022, Confidence in consciousness research. WIREs Cognitive Science.

Michel, M., Fleming, S.M., Lau, H., Lee, A.L.F., Martinez-Conde, S., Passingham, R.E., Peters, M.A.K., Rahnev, D., Sergent, C., Liu, K., 2018. An Informal Internet Survey on the Current State of Consciousness Science. Front. Psychol. 9, 2134.

Milner, A.D., 2012. Is visual processing in the dorsal stream accessible to consciousness? Proc. Biol. Sci. 279, 2289–2298.

Miyoshi, K., Lau, H., 2020. A Decision-Congruent Heuristic Gives Superior Metacognitive Sensitivity under Realistic Variance Assumptions. Psychological Review 127 (5): 655–71.

Morales, J., Lau, H., Fleming, S., 2018. Domain-General and Domain-Specific Patterns of Activity Supporting Metacognition in Human Prefrontal Cortex. J. Neurosci. 2360–2317.

Morch, H.H., 2019. Is Consciousness Intrinsic?: A Problem for the Integrated Information Theory. J. Conscious. Stud. 26, 133–162.

Nagel, T., 1974. What is it like to be a bat? Philos. Rev. 83, 435.

Norman, K.J., Riceberg, J.S., Koike, H., Bateh, J., McCraney, S.E., Caro, K., Kato, D., Liang, A., Yamamuro, K., Flanigan, M.E., Kam, K., Falk, E.N., Brady, D.M., Cho, C., Sadahiro, M., Yoshitake, K., Maccario, P., Demars, M.P., Waltrip, L., Varga, A.W., Russo, S.J., Baxter, M.G., Shapiro, M.L., Rudebeck, P.H., Morishita, H., 2021. Post-error recruitment of frontal sensory cortical projections promotes attention in mice. Neuron 109, 1202–1213. .e5.

Noudoost, B., Moore, T., 2011. Control of visual cortical signals by prefrontal dopamine. Nature 474, 372–375.

Oizumi, M., Albantakis, L., Tononi, G., 2014. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. PLoS Comput. Biol. 10 https://doi.org/10.1371/journal.pcbi.1003588.

Overgaard, M., Fazekas, P., 2016. Can No-Report Paradigms Extract True Correlates of Consciousness? Trends Cogn. Sci.

Panagiotaropoulos, T.I., Dwarakanath, A., Kapoor, V., 2020. Prefrontal Cortex and Consciousness: Beware of the Signals. Trends Cogn. Sci. 0 https://doi.org/10.1016/j.tics.2020.02.005.

Petersen, S.E., Posner, M.I., 2012. The attention system of the human brain: 20 years after. Annu. Rev. Neurosci. 35, 73–89.

Pophale, A., Shimizu, K., Mano, T., Iglesias, T.L., Martin, K., Hiroi, M., Asada, K., Andaluz, P.G., Van Dinh, T.T., Meshulam, L., Reiter, S., 2023. Wake-like skin patterning and neural activity during octopus sleep. Nature. https://doi.org/10.1038/s41586-023-06203-4.

Pu, M., Heleven, E., Delplanque, J., Gibert, N., Ma, Q., Funghi, G., Van Overwalle, F., 2020. The posterior cerebellum supports the explicit sequence learning linked to trait attribution. Cogn. Affect. Behav. Neurosci. 20, 798–815.

Rao, R.P., Ballard, D.H., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2, 79–87.

Rotondo, A.P., Raman, D.V., O'Leary, T., 2023, How Cerebellar Architecture and Dense Activation Patterns Facilitate Online Learning in Dynamic Tasks. bioRxiv. https://doi.org/10.1101/2022.10.20.512268.

Schapiro, A.C., McDevitt, E.A., Rogers, T.T., Mednick, S.C., Norman, K.A., 2018. Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. Nat. Commun. 9, 3920.

Scott, R.B., Samaha, J., Chrisley, R., Dienes, Z., 2018. Prevailing theories of consciousness are challenged by novel cross-modal associations acquired between subliminal stimuli. Cognition 175, 169–185.

Sendhilnathan, N., Semework, M., Goldberg, M.E., Ipata, A.E., 2020. Neural Correlates of Reinforcement Learning in Mid-lateral Cerebellum. Neuron 106, 188–198.e5.

Seth, A.K., Bayne, T., 2022. Theories of consciousness. Nat. Rev. Neurosci. 1–14.

Shanahan, M., Baars, B., 2005. Applying global workspace theory to the frame problem. Cognition 98, 157–176.

Shibata, K., Watanabe, T., Sasaki, Y., Kawato, M., 2011. Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. Science 334, 1413–1415.

Simons, J.S., Garrison, J.R., Johnson, M.K., 2017. Brain Mechanisms of Reality Monitoring. Trends Cogn. Sci. 21, 462–473.

Sugimoto, N., Haruno, M., Doya, K., Kawato, M., 2012. MOSAIC for multiple-reward environments. Neural Comput. 24, 577–606.

Tanaka, S., Doya, K., Okada, G., Ueda, K., Okamoto, Y., Yamawaki, S., 2004. Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. Nat. Neurosci. 7, 887–893.

Tani, J., 2016, Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena. Oxford University Press.

Tedesco, A.M., Chiricozzi, F.R., Clausi, S., Lupo, M., Molinari, M., Leggio, M.G., 2011. The cerebellar cognitive profile. Brain 134, 3672–3686.

Tong, F., Meng, M., Blake, R., 2006. Neural bases of binocular rivalry. Trends Cogn. Sci. 10, 502511.

Tononi, G., 2004. An information integration theory of consciousness. BMC Neurosci. 5, 42.

Tsuchiya, N., Wilke, M., Frässle, S., Lamme, V.A.F., 2015. No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. Trends Cogn. Sci. 19, 757–770.

Vidaurre, D., Smith, S., Woolrich, M., 2017. Brain network dynamics are hierarchically organized in time. Proc. Natl. Acad. Sci. 114, 12827–12832.

Wang, M., Foster, D.J., Pfeiffer, B.E., 2020. Alternating sequences of future and past behavior encoded within hippocampal theta oscillations. Science 370, 247–250.

Watanabe, T., Sasaki, Y., Shibata, K., Kawato, M., 2017. Advances in fMRI real-time neurofeedback. Trends Cogn. Sci. 21, 997–1010.

Weilnhammer, V., Fritsch, M., Chikermane, M., Eckert, A.-L., Kanthak, K., Stuke, H., Kaminski, J., Sterzer, P., 2021. An active role of inferior frontal cortex in conscious experience. Curr. Biol. https://doi.org/10.1016/j.cub.2021.04.043.

Wolpert, D., Kawato, M., 1998. Multiple paired forward and inverse models for motor control. Neural Netw. 11, 1317–1329.

Young, J.Z., 1971, The anatomy of the nervous system of Octopus vulgaris. Oxford University Press, London, England.

Zikopoulos, 4. Basilis, Barbas, H., 2007, Circuits for multisensory integration and attentional modulation through the prefrontal cortex and the thalamic reticular nucleus in primates. Rev. Neurosci.