Trends in Cognitive Sciences

Learning to be conscious

--Manuscript Draft--

Manuscript Number:	TICS-D-19-00166R2
Article Type:	Opinion
Keywords:	consciousness; Learning; global workspace theory; higher-order theories; predictive processing; metacognition
Corresponding Author:	Axel Cleeremans Université Libre de Bruxelles CP 122 Brussels, BELGIUM
First Author:	Axel Cleeremans
Order of Authors:	Axel Cleeremans
	Dalila Achoui
	Arnaud Beauny
	Lars Keuninckx
	Jean-Remy Martin
	Santiago Muñoz-Moldes
	Laurène Vuillaume
	Adélaïde de Heering
Abstract:	Consciousness remains a formidable challenge. Different theories of consciousness have proposed different mechanisms to account for phenomenal experience. Here, appealing to Global Workspace Theory, Higher-Order Theories, Social Theories, and Predictive Processing, we introduce a novel framework — the Self-Organizing Metarerpresentational Account (SOMA), in which consciousness is viewed as something that the brain learns to do. The brain continuously and unconsciously learns to redescribe its own activity to itself, so developing systems of metarepresentations that qualify target first-order representations. Experiences only occur in experiencers that have learned to know they possess certain first-order states and that have learned to care more about certain states than about others. Thus, consciousness is the brain's (unconscious, embodied, enactive, non-conceptual) theory about itself.

Learning to be conscious

Axel Cleeremans, Dalila Achoui, Arnaud Beauny, Lars Keuninckx, Jean-Remy Martin, Santiago Muñoz-Moldes, Laurène Vuillaume, & Adélaïde de Heering

Affiliation:
Consciousness, Cognition & Computation Group (CO3)
Center for Research in Cognition & Neuroscience (CRCN)
ULB Neuroscience Institute (UNI)
Université libre de Bruxelles
50 ave. F.-D. Roosevelt CP191
B1050 Bruxelles
BELGIUM

Corresponding author: axcleer@ulb.ac.be (Axel Cleeremans)

Keywords: consciousness, learning, global workspace theory, higher-order theories, predictive processing, metacognition

Abstract:

Consciousness remains a formidable challenge. Different theories of consciousness have proposed vastly different mechanisms to account for phenomenal experience. Here, appealing to aspects of Global Workspace Theory, Higher-Order Theories, Social Theories, and Predictive Processing, we introduce a novel framework — the Self-Organizing Metarerpresentational Account (SOMA), in which consciousness is viewed as something that the brain learns to do. By this account, the brain continuously and unconsciously learns to redescribe its own activity to itself, so developing systems of metarepresentations that qualify target first-order representations. Thus, experiences only occur in experiencers that have learned to know they possess certain first-order states and that have learned to care more about certain states than about others. In this sense, consciousness is the brain's (unconscious, embodied, enactive, non-conceptual) theory about itself.

The mystery of consciousness

Consciousness (see Glossary), by which we mean phenomenal experience, remains a genuine mystery — a problem, as Dennett [1] put it, "about which one does not know how to think about yet". Today, after thirty years of concerted scientific research [2-4] dedicated to understanding the biological bases of consciousness, we seem no closer to understanding why it feels like anything at all to be oneself. Different theories offer contrasted accounts of the cognitive functions that consciousness affords (**access consciousness**) [5, 6]; others have attempted to directly address the felt qualities of conscious states (**phenomenal consciousness**) [7-12], but none have achieved sufficient consensus to elicit widespread endorsement [13].

In this *Opinion* piece, we develop a novel perspective on consciousness that we hope will stimulate debate and help integrate different aspects of extant proposals, in particular Global Workspace Theory (GWT) [5, 6, 14], Higher-Order Theories (HOT) [15-17], Social Theories [18-21], and Predictive Processing [22-27]. At its core, our proposal is that consciousness should be viewed as a process that results from continuously operating unconscious learning and plasticity mechanisms. In other words, *consciousness is something that the brain learns to do*, by which we mean to suggest that phenomenal experience, rather than being an intrinsic property of some patterns of neural activation, should instead be viewed as the product of active, plasticity-driven mechanisms through which the brain learns to redescribe its own activity to itself.

Awareness is not sensitivity

To develop this argument, we begin by noting that all sorts of systems are *sensitive* to their environments: Plants, thermostats, computers — all are capable of detecting the states of affairs that they evolved or were designed to be sensitive to, and to react to them in appropriate ways. Yet, few would be willing to attribute any form of awareness to such systems: *Awareness is not sensitivity.* What is the difference, then, between such systems and conscious systems?

Different extant theories address this core challenge in different ways. Amongst the many views that are currently competing, two stand out: Global Workspace Theory (GWT) [5, 28], and Higher-Order theories (HOT) [15, 16, 29] (**BOX 1**). While GWT is not typically taken to be a theory of phenomenal experience, it is fair to say that it links phenomenal experience with global availability: at any point in time, conscious mental states are those that are globally available. HOT, on the other hand, links phenomenal experience with **metarepresentation**: conscious mental states are those that *we* are conscious *of*.

While both perspectives have been criticized [16, 29, 30], and while comparing them offers interesting empirical challenges that are now the object of concerted efforts (i.e., an ongoing Templeton World Charity Foundation initiative aimed at fostering adversarial collaboration), we note that higher-order views are attracting increasing interest [16, 17, 29, 31, 32]. GWT and HOT are often taken to be at odds with each other insofar as core theoretical tenets and empirical evidence are concerned [32]. However, we see reasons to think that they may be usefully reconciled with each other. Different proposals have defended germane (but not identical) ideas. Van Gulick's Higher-Order Global State theory (HOGS, see [33]) is such an

attempt. Likewise, Shea and Frith [34] have recently argued that "the Global Workspace needs **metacognition**".

Here, we take it as a starting point that "phenomenal awareness always involves a form of (subpersonal) metacognition", as Lau (personal communication) recently put it. Thus, we assume that consciousness minimally entails that one is sensitive to one's sensitivity. This segues well with our intuitive understanding of the difference between conscious and unconscious representations: We say that someone is aware of some state of affairs not merely when she is sensitive to that state of affairs, but rather when she knows *that* she is sensitive to that state of affairs. Because the brain only has access to external states of affairs through its sensorium, this suggests that awareness involves (1) a **first-order** representation of the external state of affairs, and (2) a further, **higher-order** representation of the fact that a representation of the target external state of affairs is now active. As we develop later, we surmise that global availability is a consequence of Representational Redescription (RR, **BOX 2**) processes through which unconscious first-order representations become *objects of representation* for the system by means of being indexed, targeted, or otherwise characterized by metarepresentations.

How do we get there?

Regardless of whether one takes GWT or HOT to best characterize the differences between conscious and unconscious cognition, a singularly essential question remains pending: *How do we get there*? How do we *build* the global workspace? How do metarepresentations come to play their role? As Fleming [35] recently asked, "How are awareness states learned?".

This often-ignored question in the consciousness literature is in our view central, for two reasons. The first reason is that learning profoundly shapes consciousness. Expertise creates as well as eliminates contents from phenomenal experience. Tasting wine for the first time is a wholly different experience than that of an oenologist [36] whose phenomenology has been enriched through expertise. But expertise can also eliminate phenomenal contents from awareness, as in the 'find the F's" illusion, whereby observers asked to count the number of instances of the letter "F" in a text passage often fail to produce the correct answer because skilled reading has, through automaticity, eliminated function words (e.g., "*of*") from awareness. Another example of how the contents of consciousness are shaped by expertise is "predictive attenuation". Tickling one's self is far less effective than being tickled [37], for when we tickle ourselves (but not when we are tickled) our brain can leverage previous experience so as to predict the consequences of our actions. Cognitive development also highlights how some changes go unheeded (i.e., the fact that our action and perceptual systems remain adapted despite our limbs growing spectacularly during the first few years), whereas other changes have profound phenomenal consequences (i.e., learning to read). Recent empirical work is strongly suggestive that perception is continuously shaped by learned priors (e.g., [38, 39]). Thus, we argue [40] that *learning shapes conscious experience and that conscious experience shapes learning*: the contents of consciousness are continuously shaped, over different time scales (i.e., development, skill acquisition, time available within a single trial) and over different spaces (interactions within the brain itself, with the world, with other people), by mandatory prediction-driven learning mechanisms, the computational goal of which is to improve control over action and hence to minimize "surprise", as in Predictive Processing [22, 24, 25].

Learning to be conscious

There is a second, more radical claim that we should like to entertain, however. Indeed, acknowledging the fundamental role that learning plays in shaping conscious experience leads to the mesmerizing possibility that learning is in fact instrumental to bootstrapping consciousness, or, to express this hypothesis in other words, that conscious experience is not only shaped by learning, but that its very occurrence depends on it.

From this perspective, experiences only occur in experiencers that have *learned to know* they possess certain first-order states and that have learned to *care* more about certain states than about others. Indeed, what would be the point of doing anything at all if the doing was not doing something to you? It is a distinctive and salient feature of conscious agents that they care about the phenomenal states they find themselves in. The obvious fact that phenomenal states have *value* for the agents who entertain them has equally obvious consequences in accounting for individual differences in phenomenology as they express themselves through a wide range of personality traits such as preference, ability, motivation, and attention [17, 29, 41]. Thus, our claim here is that phenomenal experience, rather than being a mere epiphenomenon associated with rewarding action, as in, say, reinforcement learning, instead has intrinsic value. But this claim only makes sense if agents are able to learn about which phenomenal states they want to find themselves in. As Dennett put it (personal communication), "How do we go from doing things for reasons to having reasons for doing things?". Having reasons for doing things is precisely what differentiates *conscious* agents from agents such as Alpha Go [42], which, despite exhibiting superhuman skill when doing things, remains unable to do so for reasons of its own.

This crucially links conscious experience with *agenthood* [43, 44]. There is no sense in which we can talk about conscious experiences without first assuming there is an experiencer who experiences those experiences. The very notion of conscious experience presupposes the existence of a subject it is the experience of. As Frege [45] pointed out, "It seems absurd to us that a pain, a mood, a wish, should rove about the world without a bearer, independently. An experience is impossible without an experiencer. The inner world presupposes the person whose inner world it is." (p. 299).

In the following, we flesh out these ideas in the form of a novel, integrative proposal based on the ideas expressed in Cleeremans' Radical Plasticity framework [46-49]. We dub this proposal "The Self-Organizing Metarepresentational Account" (SOMA).

The Self-Organizing Metarepresentational Account

The theory is based on three assumptions. The first is that information processing as carried out by neurons is intrinsically unconscious. An implication of this assumption is that consciousness depends on specific mechanisms rather than on intrinsic properties of local neural activity. The second is that information processing as carried out by the brain is graded and cascades [50] in a continuous flow [51] over the multiple levels of a **heterarchy** [52, 53] extending from the posterior to the anterior cortex as evidence accumulates during information processing episodes. An implication of this assumption is that consciousness takes time. The third assumption is that plasticity is mandatory: The brain learns all the time, whether we intend to or not [54] ; each experience leaves a trace in the brain [55].

First-order processing as a necessary condition for consciousness

With these assumptions in place, we surmise that the extent to which a representation is available to different aspects of consciousness (i.e., action, control, and experience) depends on quality of representation [47, 56, 57], a first-order property. **Quality of representation** (QoR) designates graded properties of neural representations, specifically (1) their strength, (2) their stability in time, and (3) their distinctiveness, by which we mean the extent to which they are different from other, competing representations. QoR depends both on bottom-up factors such as stimulus properties (i.e., energy, duration) and on top-down factors such as attention [58]. Crucially, QoR *changes* as a function of learning and plasticity, over different time-scales, so that the weak representations associated with subliminal processing or with the early stages of acquiring a new skill get progressively stronger and more likely to influence behaviour as a function of both time available for processing and plasticity-driven mechanisms that increase their overall quality through learning. Neither the weak representations associated with subliminal processing nor the very strong representations associated with automaticity are available to cognitive control, but for very different reasons that can be understood from an adaptive point of view: Weak representations do not need to be controlled because they only exert weak effects on behaviour. Strong representations, on the other hand, do not need to be controlled either — but only as long as the effects they exert on behaviour can be trusted to be adaptive, as is the case in automaticity. This leaves intermediate representations as the main target of cognitive control, that is, representations that are strong enough that they begin exerting significant effects on action, yet not strong enough that their influence can be left to unfold unfettered. From this, the extent to which given representations are available to form the contents of phenomenal experience is assumed to depend on both their availability to action and their availability to cognitive control [59]. This predicts (1) that weak representations are

simply not available to form such contents, (2) that the intermediate, flexible representations associated with intentional, controlled processing are the most likely to form the contents of phenomenal experience, and (3) that the very strong representations associated with automaticity, while available to form the contents of a processing episode, are typically dimmed out unless amplified through attention. This accounts for the loss of phenomenology associated with automaticity, and also for the fact that metacognitive accuracy often lags first-order performance initially, but *precedes* first-order performance with expertise (i.e., I know that I know the answer to a query before I can actually answer the query). One would thus expect non-monotonic effects as expertise develops, in different paradigms ranging from perception to motor learning. In this continuum, the intermediate representations that are of sufficient QoR that they begin exerting significant effects on behaviour yet not sufficiently automatized that they can exert their influence outside of conscious control are the best candidates for Representational Redescription (RR, see **BOX 2**), and can thus be recoded in different ways, e.g., as linguistic propositions supporting verbal report.

The distinctions introduced here overlap partially with those introduced by other theories — Dehaene's conscious–preconscious–unconscious taxonomy [60], Lamme's Stages 1/2/3/4 framework [61], and Kouider's partial awareness hypothesis [62], but uniquely frame the transitions dynamically as resulting from the consequences of learning and plasticity mechanisms through which the system learns about the geography and dynamics of its own internal representations.

Metarepresentation as a sufficient condition for consciousness?

How do we go from the mere *sensitivity* exhibited by first-order systems to consciousness? As many studies have now demonstrated, even strong, high-quality stimuli can fail to be conscious – this is what happens in change blindness [63], in the attentional blink [64] or in inattentional blindness [65]. Further, states of altered consciousness like hypnosis, and pathological states such as blindsight [66-68] or hemineglect all suggest that high-quality percepts can fail to be consciously represented while (putatively) remaining causally efficacious.

These observations are indicative that merely achieving sufficient quality (i.e., sufficient strength, stability, and distinctiveness), while necessary for a representation to be a conscious representation, is not sufficient. HOT precisely proposes that the contents of first-order representations are only conscious when they are the target of relevant metarepresentations. The densely connected prefrontal cortex (PFC), which we know is involved in conscious report [69, 70] and in metacognition [71] is a good candidate to support such metarepresentations. It is important to note, however, that our perspective does not mandate PFC involvement, and that there remains substantial debate about the role of PFC in subtending conscious experience [69, 72, 73].

Our core suggestion is that a relevant minimal mechanism to support metarepresentation involves Representational Redescription, that is, the ability for a system to redescribe its own representations to itself in ways that make it possible for the relevant action-oriented first-order knowledge it implicitly acquired to be available as data structures to the system as a whole. As Clark and Karmiloff-Smith [74] put it, implicit knowledge "… is knowledge *in* the system, but it is not yet knowledge *to* the system. A procedure for producing a particular output is available

as a whole to other processes, but its component parts (i.e., the knowledge embedded in the procedure) are not." (p. 495).

Figure 1: Tangled loops

One way of enabling a system to be sensitive to its own sensitivity is to have a second, higher-order system act as an observer of a first-order network's internal states (**Figure 1a**). In such a system, one network learns about the world, carrying out first-order decisions. This entire first-network, or layers thereof, is also input to a second-order network, the task of which is to learn something about the representations and the dynamics of the first-order network, endowing it with the ability to express judgments about and to characterize (mental attitudes) what the first-order network knows, so as to develop metarepresentations about the relevant first-order knowledge.

In prior work, we have provided different instantiated computational examples of how such higher-order networks, however elementary, can nevertheless account for many existing patterns of association and dissociation between conscious and unconscious knowledge, or between metacognitive judgements and first-order performance [75-77].

Such metarepresentations subtend not only effective metacognition [71], executive control and verbal report [32], but also, we contend, phenomenal experience itself. Crucially, such redescription processes need neither be conscious, nor conceptual, nor global. The RR mechanism echoes central aspects of both GWT and HOT. Indeed, along with the idea that first-order mental states across sensory modalities and action systems can themselves become *objects of representation* through unconscious RR processes operating through a predictive

inner loop, our proposal leads naturally to the kind of hierarchical structure that enables widespread availability to many transmitting and consuming systems in the brain — the core idea of GWT, but with a higher-order twist [33, 34]. Thus, the very architecture of the global workspace (**Figure 1b)** may simply be the result of repeated representational redescription aimed at improving control over action.

Importantly however, here, and in contrast to Rosenthal's Higher-Order Thought Theory [15], such metarepresentational models (1) may be local and hence exist anywhere in the brain, (2) may be subpersonal, and (3) are subject, just like first-order representations, to plasticity, and can thus themselves become automatic. We note that three recent proposals have expressed germane ideas: Fleming's concept of "verbal reports as inference in a higher-order state space" [35] precisely captures the core idea that reports about our own mental states involve generative models actively monitoring perceptual content. Second, Lau's characterization of consciousness as involving "perceptual reality monitoring" [31, 35, 78] is similarly buttressed on the idea that "consciousness involves subpersonal metacognition". As we develop below, such mechanisms appear necessary to enable a system to distinguish between, say, genuine perceptual input and mental imagery or hallucinations. This, we claim, can only be achieved as long as the observing system has *learned* about the states in which the observed system typically finds itself in. Third, Gershman [79] has recently proposed that phenomenal experience (and abnormalities thereof) results from the interactions between generators (of first-order content) and (higher-order) discriminators in a Generative Adversarial Network (GAN) framework. These recent proposals all share the core intuition that phenomenal experience emerges out of the (learning-driven) interactions between first-order perception-to-action systems and higher-order monitoring and control systems — the central mechanism of metacognition [80].

I am a strange loop

In what way do the learning and plasticity mechanisms that shape interactions between first-order and higher-order systems operate? We assume that they involve similar prediction-driven RR mechanisms that extend over three entangled loops: An inner loop, through which the brain learns about itself, a perception-action loop, through which agents learn about the consequences of action on the world, and a self-other loop, through which they learn about the consequences of action on other agents.

A first, internal or "inner loop", involves the brain redescribing its own representations to itself as a result of its continuous unconscious attempts at predicting how activity in one region influences activity in other regions. The provocative idea here is that the brain *does not know*, e.g., that SMA activity consistently precedes M1 activity. To represent this causal link to itself, it therefore has to learn to redescribe its own activity so that the causal link is now represented explicitly, that is, as an active pattern of neural activity that is available to other systems as a data structure (**Figure 2**). While any layer in a neural network can appropriately be characterized as a redescription of lower-level layers, metarepresentations additionally involve representing the representational relationship itself. As Perner [81] put it: metarepresentations "represent representations *as* representations". Thus, in **Figure 2a**, while neuron B can appropriately be described as representing (as indicating) the activity of neuron A, it takes neuron C (**Figure 2b**) to represent the representational relationship between neurons A and B, so making the implicit information contained in the connection between A and B explicit and available as data to other systems.

Figure 2: Representational Redescription

A substantial pending question here is the extent to which the observing (predictive) systems need to be causally independent from the target first-order systems for them to play out their metarepresentational functions. We note that the same discussion concerning the thorny problem of causality, and in particular circular causality [82] in recurrent systems takes place at other levels of description. For instance, Fleming and Daw [83] distinguish between three classes of metacognitive systems: First-order models, in which actions and confidence are computed based on the same first-order signals, second-order models, in which actions and confidence are computed fully independently, and post-decisional models, in which action information is allowed to influence confidence. The extent to which causal independence is necessary for a representation to count as metarepresentational is a matter of further analysis and empirical research.

It is important to keep in mind that this inner loop involves multiple layers of recurrent connectivity, at different scales throughout the brain. Empirical evidence that the brain "learns about itself" is scant (but see, e.g.,[84], for evidence that the brain anticipates the metabolic needs of specific regions), but we note that plasticity is an integral aspect of all contemporary theories of neural function. This is further broadly consistent with the core assumptions of generative models in general and with the perspective of "radical predictive processing", according to which "cognition is accomplished by a canonical, ubiquitous microcircuit motif replicated across all sensory and cognitive domains in which specific classes of neurons reciprocally pass predictions and prediction errors across all the global neuronal hierarchy" [85, p. 2463].

A second "perception-action loop" results from the agent as a whole predicting the consequences of its actions on the world [13-14]. Not only does perception lead to action, but

acting can itself influence both perception [86] and metacognition [87-89]. Here, our proposal echoes the enactive perspective put forward by O'Regan and Noë [90]. Successful interaction with the world, and, tentatively, our experience of such interactions, depends on learning-driven "mastery of sensorimotor contingencies" and is broadly consistent with the assumptions of active inference — the processes through which internal generative models minimize prediction error through action [22, 24, 25, 27].

We then note that when such prediction-driven learning mechanisms are directed towards improving an agent's ability to act adaptively towards other agents, their operation results in the emergence of systems of internal representations (internal models) that capture the structure and variability of other people's unobservable internal states [19, 91, 92].

This third, "self-other loop", we argue, is the scaffolding that makes it possible for an agent to redescribe its own activity to itself [93] — for now it is endowed with an (implicit, unconscious, enactive, embodied) internal model of what it takes to be an agent [94] — precisely what social theories of consciousness have proposed [19, 21] [20]. This proposal is supported by the hypothesis that **theory of mind** [95, 96] can be understood as rooted in the very same mechanisms of predictive redescriptions as involved when interacting with the world or with oneself [18]. Rather than seeing such redescriptions as internally generated, qualitatively different representations of discrete knowledge about the world, the "social" redescription is an ongoing learning process driven by increasingly complex interactive contexts [97], such as when moving from dyadic to triadic interaction, for instance [98]. Social context as a driving force for learning has, indeed, been recognized in language learning [99], child development [100] and social cognition [101].

Thus, something unique happens when a developing agent has *models of itself* available to it [18] in the form of other agents that it can infer the unobservable internal states of merely by interacting with them [102, 103]. Selves are thus embodied, virtual and transparent renditions of the underlying biological machinery [104] that produces them, and emerge progressively over development as a mandatory consequence of dynamic interactions with other agents [19, 93].

The relationships between theory of mind, **self-awareness** and perceptual awareness are complex, interwoven, and loopy. Here, we argue that they are strongly interdependent on each other: The processing carried out by the inner loop is causally dependent on the existence of both the perception-action loop and the self-other loop, with the entire system thus forming a "tangled hierarchy" (e.g., Hofstadter's concept of "a strange loop" [105, 106]) of predictive internal models [44, 91]. In this light, the social world is thus instrumental in generating conscious experience, for something special happens when we try to build a model of the internal, unobservable states of agents that are just like ourselves [16-17]. As Frith (personal communication) put it, in this sense, "consciousness is for other people". Language, as the metarepresentational tool per excellence, undoubtedly plays a role in explaining the seemingly singular nature of human consciousness [107].

Who is conscious, then? Our perspective predicts that phenomenal awareness depends on (1) The existence of massive information-processing resources that are sufficiently powerful to simulate certain aspects of one's own physical basis and inner workings; (2) the operation of continuously learning systems that attempt to predict future states and (3) immersion in a sufficiently rich social environment, specifically, environments *from which models of yourself*

can be built. Which organisms meet these criteria is, obviously, an open and challenging empirical question.

Concluding remarks

This piece had the main goal of fleshing out the original proposal that conscious experience — what it feels like to have mental states [108] — is the result of continuously operating (unconscious) prediction-driven representational redescription processes, the computational goal of which is to enable better control of action through the anticipation of the consequences of action or activity on the brain itself, on the world, and on other people. Consciousness, from this perspective, is the brain's implicit, embodied, enactive, and non-conceptual theory about itself In other words, we "learn to be conscious". Thus, we broadly espouse the enactive approach [90] [109] — that neural activity is, at its core, driven by action, and that phenomenal experience amounts to learned knowledge of the sensorimotor contingencies — but extend it both inwards (the brain learning about itself) and further outwards (the brain learning about other minds).

Beyond instantiating a search for the "computational correlates of consciousness" [57], our approach also suggests new avenues for empirical research. Our understanding of the differences between conscious and unconscious cognition would clearly benefit from increased focus on documenting the dynamics of consciousness at different scales, from cognitive development [110] to learning situations [38] and individual perceptual episodes [111].

To conclude, a good metaphor for all of this is the following. The brain is as an unconscious biological machine which, in the process of trying to figure out what the consequences of the

actions it carries out through its body, ends up developing a model of itself which is largely shaped based on interactions with other agents. This model is a (sketchy, high-level, unconscious, non-conceptual, prediction-relevant) representation of the inner workings of the machine that produced it. It is self-organizing in the sense that it is through broadly unsupervised learning mechanisms that the brain creates both a first-order sensorium and the higher-level redescriptions that ultimately makes it possible for agents to represent themselves as entertaining mental states. This is where the miracle happens — the rest is a long story about the complex interactions between the machine (the brain) and the representation of itself that it has developed over its existence (see **Outstanding Questions**). Where does consciousness come from in such a system? If one accepts the idea that consciousness amounts to being (unconsciously) sensitive to the fact that one knows, then this is exactly the sort of mechanism we need. Of course, consciousness being such a thorny problem, some will always claim: "But this is just a mechanism!". But consciousness, if it affords a scientific explanation at all, cannot be anything else than a mechanism, as both Seth [112] and Dennett [113] have forcefully argued.

Acknowledgments

This work was supported by European Research Council Advanced Grant #340718 "RADICAL" to Axel Cleeremans. Dalila Achoui, Arnaud Beauny, Lars Keuninckx, Jean-Remy Martin, Santiago Muñoz Moldes, Laurène Vuillaume, & Adélaïde de Heering were supported by the grant. Axel Cleeremans is a Research Director with the Fonds de la Recherche Scientifique (F.R.S.-FNRS, Belgium) and a Senior Fellow of the Canadian Institute for Advanced Research (CIFAR). We thank Matthias Michel for useful comments on a previous version of this manuscript as well as the referees and the editor for their constructive appraisal of the submission.

Bibliography

1. Dennett, D.C. (1991) Consciousness Explained, Little, Brown & Co.
2. Crick, F.H.C. and Koch, C. (1990) Towards a neurobiological theory of consciousness. Seminars in the Neurosciences 2, 263-275.
3. Michel, M. et al. (2019) Opportunities and challenges for a maturing science of consciousness. . Nature Human Behaviour 3, 104-107.
4. Sohn, E. (2019) Decoding the neuroscience of consciousness. Nature 571, S2-S5.
5. Baars, B.J. (1988) A Cognitive Theory of Consciousness, Cambridge University Press.
6. Dehaene, S. et al. (1998) A neuronal model of a global workspace in effortful cognitive tasks. Proceedings of the National Academy of Sciences of the U.S.A. 95 (24), 14529-14534.
7. Dretske, F. (1995) Naturalizing the Mind, MIT Press.
8. Humphrey, N. (2006) Seeing Red, Harvard University Press.
9. O'Regan, J.K. (2011) Why red doesn't sound like a bell: Understanding the feel of consciousness, Oxford University Press.
10. Tye, M. (1995) Ten problems of consciousness, MIT Press.
11. Tononi, G. and Edelman, G.M. (1998) Consciousness and complexity. Science 282 (5395), 1846-1851.
12. Damasio, A. (1999) The feeling of what happens: Body and Emotion in the Making of Consciousness, Harcourt Brace & Company.
13. Michel, M. et al. (2018) An informal internet survey on the current state of consciousness science. Frontiers in Psychology 9.
14. Dehaene, S. and Naccache, L. (2001) Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. Cognition 79, 1-37.
15. Rosenthal, D. (1997) A theory of consciousness. In The Nature of Consciousness: Philosophical Debates (Block, N. et al. eds), MIT Press.
16. Lau, H. and Rosenthal, D. (2011) Empirical support for higher-order theories of consciousness. Trends in Cognitive Sciences 15 (8), 365-373.
17. LeDoux, J.E. and Brown, R. (2017) A higher-order theory of emotional consciousness. Proc Natl Acad Sci U S A 114 (10), E2016-25.

18. Carruthers, P. (2009) How we know our own minds: the relationship between mindreading and metacognition. Behavioral and Brain Sciences 32 (2), 121-138.

19. Frith, C.D. (2007) Making up the mind, Blackwell Publishing.

20. Graziano, M. (2015) Consciousness and the social brain, Oxford University Press.

21. Graziano, M. and Karstner, S. (2011) Human consciousness and its relationship to social neuroscience: A novel hypothesis. Cognitive Neuroscience 2 (2), 98-113.

22. Friston, K. (2006) A free energy principle for the brain. Journal of Physiology (Paris) 100, 70-87.

23. Clark, A. (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences 36 (3), 181-204.

24. Clark, A. (2016) Surfing uncertainty: Prediction, Action, and the Embodied Mind, Oxford University Press.

25. Hohwy, J. (2013) The predictive mind, Oxford University Press.

26. Bar, M. (2009) Predictions: a universal principle in the operation of the human brain. Philosophical Transactions of the Royal Society B 364, 1181-1182.

27. Seth, A. (2014) A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synsethesia. Cognitive Neuroscience 5 (2), 97-118.

28. Dehaene, S. et al. (2003) A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proceedings of the National Academy of Sciences of the U.S.A. 100 (14), 8520-8525.

29. Brown, R. et al. (2019) Understanding the higher-order approach to consciousness. Trends in Cognitive Science 23 (9), 754-768.

30. Block, N. (2011) The higher-order approach to consciousness is defunct. Analysis 71 (3).

31. Lau, H. (2008) A higher-order Bayesian Decision Theory of consciousness. In Models of brain and mind. Physical, computational and psychological approaches. Progress in Brain Research. Progress in Brain Research (Banerjee, R. and Chakrabarti, B.K. eds), pp. 35-48, Elsevier.

32. Dehaene, S. et al. (2017) What is consciousness and could machines have it? Science 358 (1-7).

33. Van Gulick, R. (2004) Higher-Order Global States (HOGS): An alternative Higher-Order model of consciousness. In Higher-Order Theories of Consciousness: An anthology (Gennaro, R.J. ed), pp. 67-90, John Benjamins.

34. Shea, N. and Frith, C.D. (2019) The Global Workspace Needs Metacognition. Trends in Cognitive Sciences 23 (7), 560-571.

35. Fleming, S.M. (2019) Awareness reports as inference in a higher-order state space. arXiv:1906.00728 [q-bio].

36. Smith, B.C. (2006) Questions of taste: The philosophy of wine, Oxford University Press.

37. Blakemore, S.J. et al. (1998) Central cancellation of self-produced tickle sensation. Nature Neuroscience 1 (7), 635-640.

38. Schwiedrzik, C.M. et al. (2009) Sensitivity and perceptual awareness increase with practice in metacontrast masking. Journal of Vision 9, 1-18.

39. de Lange, F.P. et al. (2018) How do expectations shape perception? Trends in Cognitive Science 22 (9), 764-779.

40. Perruchet, P. and Vinter, A. (2002) The self-organizing consciousness. Behavioral and Brain Sciences 25 (3), 297-330.

41. Hornsby, A.N. and Love, B.C. (2019) How decisions and the desire for coherency shape subjective preferences over time. PsyarXiv.

42. Silver, D. et al. (2017) Mastering the game of Go without human knowledge. Nature 550, 354.

43. Bayne, T. (2013) Agency as a marker of consciousness. In Decomposing the will (Clark, A. et al. eds), pp. 160-177, Oxford University Press.

44. Pacherie, E. (2008) The phenomenology of action: A conceptual framework. Cognition 107, 179-217.

45. Frege, G. (1918/1956) The thought: A logical enquiry. Mind 65 (259), 289-311.

46. Cleeremans, A. (2008) Consciousness: the radical plasticity thesis. In Models of Brain and Mind: Physical, Computational and Psychological Approaches. Progress in Brain Research (Banerjee, R. and Chakrabarti, B.K. eds), pp. 19-33, Elsevier.

47. Cleeremans, A. (2011) The radical plasticity thesis: How the brain learns to be conscious. Frontiers in Psychology 2, 1-12.

48. Cleeremans, A. (2014) Connecting conscious and unconscious cognition. Cognitive Science 38 (6), 1286-1315.

49. Cleeremans, A. (2019) Consciousness (unconsciously) designs itself. Journal of Consciousness Studies 26 (3-4), 88-111.

50. McClelland, J.L. (1979) On the time-relations of mental processes: An examination of systems in cascade. Psychological Review 86, 287-330.

51. Eriksen, C.W. and Schultz, D.W. (1979) Information processing in visual search: A continuous flow conception and experimental results. Attention, Perception & Psychophysics 25 (4), 249-263.

52. McCulloch, W.S. (1945) A heterarchy of values determined by the topology of nervous nets. Bull. Math. Biophys. 7, 89-93.

53. Fuster, J.M. (2008) The prefrontal cortex, 4th edn., Academic Press.

54. Cleeremans, A. et al. (1998) Implicit learning: News from the front. Trends in Cognitive Sciences 2, 406-416.

55. Kreiman, G. et al. (2002) Single-neuron correlates of subjective vision in the human medial temporal lobe. Proceedings of the National Academy of Sciences of the U.S.A. 99 (8378-8383).

56. Farah, M.J. (1994) Neuropsychological inference with an interactive brain: A critique of the "locality" assumption. Behavioral and Brain Sciences 17, 43-104.

57. Cleeremans, A. (2005) Computational correlates of consciousness. Boundaries of Consciousness: Neurobiology and Neuropathology 150, 81-98.

58. Dehaene, S. et al. (2006) Conscious, preconscious, and subliminal processing: A testable taxonomy. Trends in Cognitive Sciences 10 (5), 204-211.

59. Shallice, T. (1978) The dominant action system: An information-processing approach to consciousness. In The steam of consciousness (Pope, K. and Singer, J.L. eds), pp. 117-157, Springer.

60. Dehaene, S. et al. (2006) Conscious, preconscious, and subliminal processing: A testable taxonomy. Trends in Cognitive Sciences 10 (5), 204-211.

61. Lamme, V.A.F. (2006) Toward a true neural stance on consciousness. Trends in Cognitive Sciences 10 (11), 494-501.

62. Kouider, S. et al. (2010) How rich is consciousness: The partial awareness hypothesis. Trends in Cognitive Sciences 14 (7), 301-307.

63. Simons, D.J. and Levin, D.T. (1997) Change Blindness. Trends in Cognitive Sciences 1, 261-267.

64. Shapiro, K.L. et al. (1997) The Attentional Blink. Trends in Cognitive Sciences 1, 291-295.

65. Mack, A. and Rock, I. (1998) Inattentional Blindness, MIT Press.

66. Muckli, L. et al. (2009) Bilateral visual field maps in a patient with only one hemisphere. Proceedings of the National Academy of Sciences 106 (31), 13034-13039.

67. Silvanto, J. and Rees, G. (2011) What does Neural Plasticity Tell us about Role of Primary Visual Cortex (V1) in Visual Awareness? Frontiers in Psychology 2.

68. Weiskrantz, L. (1986) Blindsight: A case study and implications, Oxford University Press.
69. Tsuchiya, N. et al. (2015) No-Report paradigms: Extracting the true neural correlates of consciousness. Trends in Cognitive Science 19 (12), 757-770.
70. Block, N. (2019) What is wrong with the no-report paradigm and how to fix it. Trends in Cognitive Science.
71. Fleming, S.M. et al. (2010) Relating instrospective accuracy to individual differences in brain structure. Science 329 (5998), 1541-1543.
72. Odergaard, B. et al. (2017) Should a few null findings falsify prefrontal theories of conscious perception? Journal of Neuroscience 37 (40), 9593-9602.
73. Boly, M. et al. (2017) Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and neuroimaging evidence. Journal of Neuroscience 2017 (37), 9603-9613.
74. Clark, A. and Karmiloff-Smith, A. (1993) The cognizer's innards: A psychological and philosophical perspective on the development of thought. Mind and Language 8, 487-519.
75. Cleeremans, A. et al. (2007) Consciousness and metarepresentation: A computational sketch. Neural Networks 20 (9), 1032-1039.
76. Pasquali, A. et al. (2010) Know thyself: Metacognitive networks and measures of consciousness. Cognition 117, 182-190.
77. Timmermans, B. et al. (2012) Higher order thoughts in action: consciousness as a unconscious re-description process. Philosophical Transcations of the Royal Society B 367, 1412-1423.
78. Lau, H. (2019) Consciousness, Metacognition, & Perceptual Reality Monitoring.
79. Gershman, S.J. (submitted) The generative adversarial brain.
80. Nelson, T.O. and Narens, L. (1990) Metamemory: A theoretical framework and new findings. The Psychology of Learning and Motivation 26, 125-173.
81. Perner, J. (1991) Understanding the representational mind, MIT Press.
82. Haken, H. (1977) Synergestics - An introduction: Nonequilibrium phase transitions and self-organization in physics, chemistry, and biology., Springer Verlag.
83. Fleming, S.M. and Daw, N.D. (2017) Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. Psychological Review 124, 91-114.
84. Sirotin, Y.B.D.A. (2009) Anticipatory haemodynamic signals in sensory cortex not predicted by local neuronal activity. Nature 457, 475-480.

85. Allen, M. and Friston, K. (2018) From cognitivism to autopoesis: towards a computational framework for the embodied mind. Synthèse 195, 2459-2482.

86. Strack, F. et al. (1998) Inhibiting and facilitating conditions of the human smile: A nonobstrusive test of the facial feedback hypothesis. . Journal of Personality and Social Psychology 54 (5), 768-777.

87. Fleming, S.M. et al. (2015) Action-specific disruption of perceptual confidence. Psychologial Science 26 (1), 89-98.

88. Wokke, M. et al. (2019) Action information contributes to metacognitive decision-making. bioRxiv.

89. Siedlecka, M. et al. (2016) But I was so sure! Metacognitive judgments are less accurate given prospectively than retrospectively. Frontiers in Psychology 7.

90. O'Regan, J.K. and Noë, A. (2001) A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences 24 (5), 883-917.

91. Wolpert, D.M. et al. (2004) A unifying computational framework for motor control and social interaction. In The neuroscience of social interaction (Frith, C.D. and Wolpert, D.M. eds), pp. 305-322, Oxford University Press.

92. Prinz, W. (2012) Open minds: The social making of agency and intentionality, MIT Press.

93. Shea, N. et al. (2014) Supra-personal cognitive control and metacognition. Trends in cognitive sciences 18 (186-193).

94. Seth, A. and Tsakiris, M. (2018) Being a beast machine: The somatic basis of selfhood. Trends in Cognitive Sciences 22 (11), 969-981.

95. Leslie, A.M. et al. (2004) Core mechanisms in "theory of mind". Trends in Cognitive Sciences 8 (12), 528-533.

96. Carruthers, P. and Smith, P.K. (1996) Theories of theories of mind, Cambridge University Press.

97. Schilbach, L. et al. (2013) Toward a second-person neuroscience. Behavioral and Brain Sciences 36 (4), 393-414.

98. Carpendale, J.I. and Lewis, C. (2004) Constructing an understanding of mind: The development of children's social understanding within social interaction. Behavioral and brain sciences 27 (1), 79–96.

99. Kuhl, P.K. (2007) Is speech learning "gated" by the social brain? Developmental Science 10, 110-120.

100. Reddy, V. (2008) How infants know minds, Harvard University Press.

101. Becchio, C. et al. (2010) Toward you: The social side of actions. Current Directions in Psychological Science 19 (3), 183-188.
102. Tamir, D. and Thornton, M.A. (2018) Modeling the predictive social mind. Trends in Cognitive Science 22 (3), 201-212.
103. Thornton, M.A. et al. (2019) The social brain automatically predicts other's future mental states. Journal of Neuroscience 39 (1), 140-148.
104. Metzinger, T. (2003) Being No One: The self-model theory of subjectivity, Bradford Books, MIT Press.
105. Hofstadter, D.R. and Dennett, D.C. (1981) The mind's I: Fantasies and reflections on self and soul, Penguin.
106. Hofstadter, D.R. (2007) I am a strange loop, Basic Books.
107. Frith, C. (2012) The role of metacognition in human social interactions. Philisophical Transactions of the Royal Society of London, B. 367, 2213-2223.
108. Nagel, T. (1974) What is like to be a bat? Philosophical Review 83, 434-450.
109. Varela, F.J. et al. (1991) The Embodied Mind: Cognitive Science and Human Experience, MIT Press.
110. Kouider, S. et al. (2013) A neural marker of perceptual consciousness in infants. Science 50 (14), 3736-3744.
111. Del Cul, A. et al. (2007) Brain dynamics underlying the nonlinear threshold for access to consciousness. PloS Biology 5 (10), e260.
112. Seth, A. (2016) The real problem. Aeon.
113. Dennett, D.C. (2001) Are we explaining consciousness yet? Cognition 79, 221-237.
114. Dienes, Z. and Perner, J. (1999) A theory of implicit and explicit knowledge. Behavioral and Brain Sciences 22, 735-808.
115. Karmiloff-Smith, A. (1992) Beyond modularity : A developmental perspective on cognitive science, MIT Press.
116. Piaget, J. (1970) Genetic Epistemology, Columbia University Press.

Glossary

- **Consciousness:** Consciousness is a mongrel concept that involves at least three distinctions: The distinction between **phenomenal consciousness** and **access consciousness**; the distinction between awareness of the world (perceptual awareness), **self-awareness**, and awareness of other people's mental states (**theory of mind**); and the distinction between *states* (e.g., sleep *versus* wakefulness) and *contents* of consciousness. Here, we use the term "consciousness" to refer to information processing that is associated with phenomenal experience.

- **Access consciousness:** Access consciousness refers to the fact that, unlike unconscious mental states, conscious mental states are available to cognitive functions such as reasoning, verbal report, memory, planning, or goal-directed behaviour.

- **Phenomenal consciousness:** Phenomenal consciousness refers to the felt subjective qualities associated with conscious mental states; "what it is like", as Thomas Nagel famously put it, to be a bat, to smell cheese, to listen to Bach, to remember a vacation, or to imagine having one next year.

- **First-order (representation):** A first-order representation is a neuronal state representing a state of affairs from the world (perception) or from one's body (interoception). First-order representations are the result of the neural computations of the constitutive sensory properties of objects such as their shape, colour, size, pitch, and so on, that are necessary to successfully drive action and decision-making.

- **Heterarchy:** Unlike hierarchies, heterarchies are connected networks in which all nodes are equipotent and may thus play different roles, including hierarchical roles, as a function of context.

- **Higher-order (representation):** Higher-order representations are, in our perspective, identical to **metarepresentations**.

- **Quality of representation:** A core concept of the proposed framework, quality of representation is a construct aimed at characterizing core properties of representations in a graded manner: Their strength, their stability in time, and their distinctiveness.

- **Metarepresentation:** A metarepresentation – or second-order representation – is a representation that conveys information about other representations in the brain, for instance, the fact that the target (first-order) representation exists, the probability that it correctly represents a true state of affairs (confidence), it emotional value, its kind (a belief, a hope, a regret, and so on).

- **Metacognition:** By metacognition (cognition about cognition), we mean the operations by which one consciously evaluates and controls one's own cognitive processes. Metacognition depends on the existence of metarepresentations.

- **Self awareness:** The sense that we have (or not) of being a conscious agent distinct from the world and from other agents. Self-awareness depends on introspection and on interoception. Here, following Carruthers, we argue that self-awareness engages the same mechanisms as theory of mind.

- **Theory of mind:** Here, by theory of mind, we mean the processes that make it possible for an agent to ascribe mental states (beliefs, desires, intentions) to other agents (including oneself).

BOX 1: *Global Workspace Theory and Higher-Order Theories*

According to Global Workspace Theory (GWT), conscious representations are made globally available to cognitive functions in a manner that unconscious representations are not. Global availability, that is, the capacity for a given representation to influence processing on a global scale (supporting, in particular, verbal report, but also goal-directed decision-making), is achieved by means of "the global neuronal workspace", a large network of high-level neural "processors" linked to each other by long-distance cortico-cortical connections. Thus, while information processing can take place without consciousness in any given specialized processor, once the contents processed by that processor enter in contact with the neural workspace, they trigger a non-linear transition dubbed "ignition" and are "broadcasted" to the entire brain, so achieving what Dennett [113] has called "fame in the brain". GWT thus solves the quandary of explaining the differences between conscious and unconscious cognition by distinguishing between causal efficacy and conscious access through architecture: Information that is *in* the neural workspace is globally available and hence conscious; information that is *outside of it* and embedded in peripheral modules is not (despite potentially retaining causal efficacy). While GWT makes no attempt to explain phenomenal awareness in and of itself, it is fair to say that it implicitly assumes that global availability is a correlate of phenomenal experience.

Higher-Order theories of consciousness, of which there are different instantiations [15, 16, 29, 31, 78], have a very different flavour. According to HOT, a mental state is conscious when the agent entertains, in a non-inferential manner, thoughts to the effect that it currently is in that mental state. Importantly, for Rosenthal, it is in virtue of occurrent higher-order thoughts that the target first-order representations become conscious. In other words, a particular

representation, say, a representation of the printed letter "J", will only be a conscious representation to the extent that there exists another (unconscious) representation (in the same brain) that indicates the fact that a (first-order) representation of the letter "J" exists at time *t*. Dienes and Perner [114] have elaborated this idea by analysing the implicit-explicit distinction as reflecting a hierarchy of the different manners in which a given representation can be explicit. Thus, a representation can explicitly indicate a property (e.g., "yellow"), predication to an individual ("the flower is yellow"), factivity ("it is a fact and not a belief that the flower is yellow") and attitude ("I know that the flower is yellow"). Fully conscious knowledge is thus knowledge that is "attitude-explicit", and conscious states are necessarily states that the subject is aware *of*. While this sounds highly counterintuitive to some authors (most notably Ned Block, see e.g., [30]), it captures the central intuition that it is precisely the fact that I am aware (that I experience the fact, that I feel) that I possess some knowledge that makes this knowledge conscious. HOT thus solves the problem of distinguishing between conscious and unconscious cognition in a completely different manner than GWT, specifically by assuming the involvement of specific kinds of representations, the function of which it is to denote the existence of and to qualify target first-order representations. Such higher-order states, or meta-representations, do not need to be localized in any particular brain region, but of course the densely interconnected prefrontal cortex is a good candidate for such metarepresentations to play out their functions [29].

BOX 2. *Representational Redescription*

Representational Redescription (RR) is a theory of cognitive development introduced by Karmiloff-Smith [115] and further developed by Clark & Karmiloff-Smith [74] about human knowledge, its processes and its by-products. The starting point of the theory is the observation that "human learning goes beyond success", that is, that children's learning often exhibits u-curved-shaped developmental trajectories whereby early behavioural mastery of a particular task is paradoxically followed by an increase of errors before a final recovery. Karmiloff-Smith interprets this pattern as reflecting the increased cognitive load induced by the reorganization of internal knowledge over the course of learning. For instance, in French, the same form ("*un"*) is used as an indefinite pronoun, i.e., "a *truck*" (vs. a car) or to denote number, i.e., "*one* truck" (vs. two). Over development, children learning French start explicitly marking the different usages of "un" by committing errors such as producing "*un de* camion" in contexts were the intent is to denote kind rather than number. Karmiloff-Smith takes such cases as indications that the underlying representations are in the process of being reorganized so as to capture formerly implicit distinctions. To account for such patterns, the RR theory distinguishes between four knowledge "levels". Implicit (level I) representations are individuated and procedural — they are effective procedures to drive behaviour but fail to be available as *objects of representation* to the system. Three further levels characterize explicit knowledge: E1 knowledge is knowledge that has been successfully redescribed into an explicit format that enables generalization. E2 knowledge is conscious knowledge. E3 knowledge is available for verbal report, that is, it can be used to justify one's decisions. Overall, the theory aimed to move away from traditional perspectives on cognitive development, in particular the idea that it proceeds by broad cross-domain stages [116]. Clark and Karmiloff-Smith [74] later elaborated on these ideas by framing them in the larger context of understanding the differences

between classical and connectionist approaches to cognition, and by asking what kinds of mechanisms might support representational redescription. Clark and Karmiloff-Smith argued that knowledge in connectionist networks is always implicit: A first-order network never knows *that* it knows. Explicit knowledge, in contrast, in the form of rules for instance, always entails awareness. The difference, according to the authors, stems precisely from the system's ability to be sensitive to its own internal representations by means of representational redescription: "For the genuine thinkers, we submit, are endowed with an internal organization which is geared to the repeated redescription of its own stored knowledge" (p. 488). Clark and Karmiloff-Smith speculated about possible mechanisms that would enable connectionist networks to learn to become sensitive to their own internal states in the way suggested by RR. We subsequently proposed possible implementations [75-77].

Figure Legends

Figure 1: Tangled loops (a): Three interacting prediction-driven loops define the dynamics of a core representational redescription (RR) system in which a first-order network mapping perception to action constitutes input to a higher-order network, the task of which is to re-represent first-order states in order to serve other computational goals, such as computing confidence and value, monitoring first-order states and dynamics, and predicting its future states (inner loop). Two further prediction-driven loops augment this core system: A perception-action loop that extends over interactions with the world, and a self-other loop that extends over interactions with other agents. The three loops form a tangled hierarchy in the sense that the operation of the inner loop, and the resulting metarepresentations, are causally dependent on the operation of the other loops. (b) Many RR systems linked to each other lead naturally to the architecture of the global workspace, the higher-level states of which should now be viewed as fundamentally metarepresentational in the sense that their core function is to redescribe first-order knowledge in such a way that they can be shared across many systems.

Figure 2: Representational Redescription (a): Neuron A is connected to Neuron B and can drive its activity, but the causal link between A and B is only implicitly represented in the connection itself. Neither A nor B explicitly represent the fact that A is causally linked to B. (b) Making the causal link between A and B explicit minimally requires a third neuron, C, the state of which can then explicitly represent the fact that neurons A and B's states are causally linked to each other. This information is then available for further representation by other systems.

Outstanding questions

- How causally independent does a monitoring system need to be from the monitored system for it to count as genuinely metarepresentational?

- Which empirical evidence would support the hypothesised mechanisms that subtend the inner loop (the brain learning about itself)?

- How does consciousness develop in exceptional social contexts, such as complete isolation or prolonged exposure to non-human agents, as is the case for feral children?

- Which neural network architecture is best suited to capture the intuitions behind Clark and Karmiloff-Smith's Representational Redescription hypothesis?

- What should be the computational goals of the Representational Redescription systems, beyond computing confidence?

- What are the links between theory of mind, self-awareness, and perceptual awareness? Is theory of mind a precursor to self-awareness? Is minimal self-awareness required for or phenomenal experience?

- If phenomenal awareness depends on specific cognitive mechanisms, as suggested here, then artificial consciousness is possible. How do we get there? Do we want to get there?

Highlights

- We introduce a novel framework dubbed the Self-organizing metarepresentational Account (SOMA), according to which consciousness is something that the brains learns to do.

- We propose Representational Redescription as the core mechanism through which higher-order monitoring systems learn about and characterize target first-order states

- We suggest that theory of mind, self-awareness, and perceptual awareness share common prediction-driven learning mechanisms that operate over three loops: An inner loop, a perception-action loop, and a self-other loop

Figure 1