frontiers in PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 09 May 2011
doi: 10.3389/fpsyg.2011.00086

The radical plasticity thesis: how the brain learns to be conscious

Axel Cleeremans*

Consciousness, Cognition and Computation Group, Université Libre de Bruxelles, Bruxelles, Belgium

Edited by:
Morten Overgaard, Aarhus University, Aarhus University Hospital, Denmark

Reviewed by:
Hakwan C. Lau, Columbia University in the City of New York, USA
Zoltan Dienes, University of Sussex, UK

****Correspondence:***
Axel Cleeremans, Consciousness, Cognition and Computation Group, Université Libre de Bruxelles CP 191, 50 Avenue F.-D. Roosevelt, B1050 Bruxelles, Belgium.
e-mail: axcleer@ulb.ac.be

In this paper, I explore the idea that consciousness is something that the brain learns to do rather than an intrinsic property of certain neural states and not others. Starting from the idea that neural activity is inherently unconscious, the question thus becomes: How does the brain learn to be conscious? I suggest that consciousness arises as a result of the brain's continuous attempts at predicting not only the consequences of its actions on the world and on other agents, but also the consequences of activity in one cerebral region on activity in other regions. By this account, the brain continuously and unconsciously learns to redescribe its own activity to itself, so developing systems of meta-representations that characterize and qualify the target first-order representations. Such learned redescriptions, enriched by the emotional value associated with them, form the basis of conscious experience. Learning and plasticity are thus central to consciousness, to the extent that 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. This is what I call the "Radical Plasticity Thesis." In a sense thus, this is the enactive perspective, but turned both inwards and (further) outwards. Consciousness involves "signal detection on the mind"; the conscious mind is the brain's (non-conceptual, implicit) theory about itself. I illustrate these ideas through neural network models that simulate the relationships between performance and awareness in different tasks.

Keywords: consciousness, learning, subjective experience, neural networks, emotion

Consider the humble but proverbial thermostat. A thermostat is a simple device that can turn a furnace on or off depending on whether the current temperature exceeds a set threshold. Thus, the thermostat can appropriately be said to be *sensitive* to temperature. But is there some sense in which the thermostat can be characterized as being *aware* of temperature? Contra Chalmers (1996), I will argue that there is no sense in which the thermostat can be characterized as being aware of temperature. There are two important points that I would like to emphasize in developing this argument. The first is that there is no sense in which the thermostat can be characterized as being aware of temperature because it does not *know that* it is sensitive to temperature. The second point is that there is no sense in which the thermostat can be characterized as being aware of temperature because it does not *care* about whether its environment is hot or cold. I will further argue that these two features – knowledge of one's own internal states and the emotional value associated with such knowledge – are constitutive of conscious experience. Finally, I will argue that learning (or, more generally, plasticity) is necessary for both features to emerge in cognitive systems. From this, it follows that consciousness is something that the brain learns to do through continuously operating mechanisms of neural plasticity. This I call the "Radical Plasticity Thesis."

Information processing can undoubtedly take place without consciousness, as abundantly demonstrated not only by empirical evidence (the best example of which is probably blindsight), but also by the very fact that extremely powerful information-processing machines, namely computers, have now become ubiquitous. Only but a few would be willing to grant any quantum of conscious experience to contemporary computers, yet they are undeniably capable of sophisticated information processing – from recognizing faces to analyzing speech, from winning chess tournaments to helping prove theorems. Thus, consciousness is not information processing; experience is an "extra ingredient" (Chalmers, 2007a) that comes over and beyond mere computation.

With this premise in mind – a premise that just restates Chalmers' (1996) *hard problem*, that is, the question of *why* it is the case that information processing is accompanied by experience in humans and other higher animals, there are several ways in which one can think about the problem of consciousness.

One is to simply state, as per Dennett (e.g., Dennett, 1991, 2001) that there is nothing more to explain. Experience is *just* (a specific kind of) information processing in the brain; the contents of experience are *just* whatever representations have come to dominate processing at some point in time ("fame in the brain"); consciousness is *just* a harmless illusion. From this perspective, it is easy to imagine that machines will be conscious when they have accrued sufficient complexity; the reason they are not conscious now is simply because they are not sophisticated enough: They lack the appropriate architecture perhaps, they lack sufficiently broad and diverse information-processing abilities, and so on. Regardless of what is missing, the basic point here is that there is no reason to assume that conscious experience is anything special. Instead, all that is required is one or several yet-to-be-identified functional mechanisms: Recurrence, perhaps (Lamme, 2003), stability of representation (O'Brien and Opie, 1999), global availability (Baars,

1988; Dehaene et al., 1998), integration and differentiation of information (Tononi, 2003, 2007), or the involvement of higher-order representations (Rosenthal, 1997, 2006), to name just a few.

Another perspective is to consider that *experience* will never be amenable to a satisfactory functional explanation. Experience, according to some (e.g., Chalmers, 1996), is precisely what is left over once all functional aspects of consciousness have been explained. Notwithstanding the fact that so defined, experience is simply not something one can approach from a scientific point of view, this position recognizes that consciousness is a unique (a *hard*) problem in the Cognitive Neurosciences. But that is a different thing from saying that a reductive account is not possible. A non-reductive account, however, is exactly what Chalmers' Naturalistic Dualism attempts to offer, by proposing that information, as a matter of ontology, has a dual aspect, – a physical aspect and a phenomenal aspect. "Experience arises by virtue of its status as one aspect of information, when the other aspect is found embodied in physical processing" (Chalmers, 2007b, p. 366). This position leads him to defend the possibility that experience is a fundamental aspect of reality. Thus, even thermostats, for instance, may be endowed with very simple experiences, in virtue of the fact that they can toggle in two different states.

What, however, do we mean when we speak of "subjective experience" or of "quale"? The simplest definition of these concepts (Nagel, 1974) goes right to the heart of the matter: "Experience" is *what it feels like* for a conscious organism to be that organism. There is something it is like for a bat to be a bat; there is nothing it is like for a stone to be a stone. As Chalmers (2007a) puts it: "When we see, for instance, we *experience* visual sensations: The felt quality of redness, the experience of dark and light, the quality of depth in a visual field" (p. 226).

Let us try to engage in some phenomenological analysis at this point to try to capture what it means for each of us to have an experience. Imagine you see a patch of red (Humphrey, 2006). You now have a *red* experience – something that a camera recording the same patch of red will most definitely *not* have. What is the difference between you and the camera? Tononi (2007), from whom I borrow this simple thought experiment, points out that one key difference is that when you see the patch of red, the state you find yourself in is but one among billions, whereas for a simple light-sensitive device, it is perhaps one of only two possible states – thus the state conveys a lot more *differentiated information* for you than for a light-sensitive diode. A further difference is that you are able to *integrate* the information conveyed by many different inputs, whereas the chip on a camera can be thought of as a mere array of independent sensors among which there is no interaction.

Hoping not to sound presumptuous, it strikes me, however, that both Chalmers' (somewhat paradoxically) and Tononi's analyses miss fundamental facts about experience: Both analyze it as a rather abstract dimension or aspect of information, whereas experience – *what it feels like* – is anything but abstract. On the contrary, what we mean when we say that seeing a patch of red elicits an "experience" is that the seeing *does something to us* – in particular, we might feel one or several emotions, and we may associate the redness with memories of red. Perhaps seeing the patch of red makes you remember the color of the dress that your prom night date wore 20 years ago. Perhaps it evokes a vague anxiety, which we now know is also shared by monkeys (Humphrey, 1971). To a synesthete, perhaps seeing the color red will evoke the number 5. The point is that if conscious experience is what it feels like to be in a certain state, then "What it feels like" can only mean the specific set of associations that have been established by experience between the stimulus or the situation you now find yourself in, on the one hand, and your memories, on the other. This is what one means by saying that there is something it is like to be you in this state rather than nobody or somebody else: The set of memories evoked by the stimulus (or by actions you perform, etc.), and, crucially, the set of emotional states associated with each of these memories. This is essentially the perspective that Damasio (2010) defends.

Thus, a first point about the very notion of subjective experience I would like to make here is that it is difficult to see what experience could mean beyond (1) the emotional value associated with a state of affairs, and (2) the vast, complex, richly structured, experience-dependent network of associations that the system has learned to associate with that state of affairs. "What it feels like" for me to see a patch of red at some point seems to be entirely exhausted by these two points. Granted, one could still imagine an agent that accesses specific memories, possibly associated with emotional value, upon seeing a patch of red and who fails to "experience" anything. But I surmise that this would be mere simulation: One *could* design such a zombie agent, but any real agent that is driven by self-developed motivation, and that cannot help but be influenced by his emotional states will undoubtedly have experiences much like ours.

Hence, there is nothing it is like for the camera to see the patch of red simply because it does not care: The stimulus is meaningless; the camera lacks even the most basic machinery that would make it possible to ascribe any interpretation to the patch of red; it is instead just a mere recording device for which nothing matters. There is nothing it is like to be that camera at that point in time simply because (1) the experience of different colors do not do anything to the camera; that is, colors are not associated with different emotional valences; and (2) the camera has no brain with which to register and process its own states. It is easy to imagine how this could be different. To hint at my forthcoming argument, a camera could, for instance, keep a record of the colors it is exposed to, and come to "like" some colors better than others. Over time, your camera would like different colors than mine, and it would also know that in some non-trivial sense. Appropriating one's mental contents for oneself is the beginning of individuation, and hence the beginning of a *self*.

Thus a second point about experience that I perceive as crucially important is that it does not make any sense to speak of experience without an *experiencer* who experiences the experiences. Experience is, almost by definition ("what it feels like"), something that takes place not in *any* physical entity but rather only in special physical entities, namely cognitive agents. Chalmers' (1996) thermostat fails to be conscious because, despite the fact that it can find itself in different internal states, it lacks the ability to remove itself from the causal chain which it instantiates. In other words, it lacks knowledge *that* it can find itself in different states; it is but a mere mechanism that responds to inputs in certain ways. While there is indeed something to be experienced there (the different states the thermostat can find itself in), there is no one home to be the *subject* of these experiences – the thermostat simply lacks

the appropriate machinery to do so. The required machinery, I surmise, minimally involves the ability to *know that* one finds itself in such or such a state.

This point can be illustrated by means of well-known results in the connectionist, or artificial neural network modeling literature. Consider for instance Hinton's (1986) famous demonstration that neural networks trained through associative learning mechanisms can learn about abstract dimensions of the training set. Hinton's (1986) network was a relatively simple back-propagation network trained to process linguistic expressions consisting of an agent, a relationship, and a patient, such as for instance "Maria is the wife of Roberto." The stimulus material consisted of a series of such expressions, which together described some of the relationships that exist in the family trees of an Italian family and of an English family The network was required to produce the patient of each agent–relationship pair it was given as input. For instance, the network should produce "Roberto" when presented with "Maria" and "wife." Crucially, each person and each relationship were presented to the network by activating a single input unit. Hence there was no overlap whatsoever between the input representations of, say, Maria and Victoria. Yet, despite this complete absence of surface similarity between training exemplars, Hinton (1986) showed that after training, the network could, under certain conditions, develop internal representations that capture relevant abstract dimensions of the domain, such as nationality, sex, or age!

Hinton's (1986) point was to demonstrate that such networks were capable of learning richly structured internal representations as a result of merely being required to process exemplars of the domain. Crucially, the structure of the internal representations learned by the network is determined by the manner in which different exemplars interact with each other, that is, by their *functional similarity,* rather than by their mere *physical similarity* expressed, for instance, in terms of how many features (input units) they share. Hinton (1986) thus provided a striking demonstration of this important and often misunderstood aspect of associative learning procedures by showing that under some circumstances, specific hidden units of the network had come to act as detectors for dimensions of the material that had never been presented explicitly to the network. These results truly flesh out the notion that rich, abstract knowledge can simply emerge as a by-product of processing structured domains. It is interesting to note that the existence of such single-unit "detectors" has recently been shown to exist in human neocortex (Kreiman et al., 2002): Single-neuron recording of activity in hippocampus, for instance, has shown that some individual neurons exclusively respond to highly abstract entities, such as the words "Bill Clinton" and images of the American president.

Now, the point I want to make with this example is as follows: One could certainly describe the network as being *sensitive* to nationality, in the sense that it exhibits differential responding (hence, behavioral sensitivity) to inputs that involve Italian agents vs. English agents. But, obviously, the network does not *know* anything about nationality. It does not even know that it has such and such representations of the inputs, nor does it know anything about its own, self-acquired sensitivity to the relevant dimensions. Instead, the rich, abstract, structured representations that the network has acquired over training forever remain embedded in a causal chain that begins with the input and ends with the network's responses. As Clark and Karmiloff-Smith (1993) insightfully pointed out, such representations are "first-order" representations to the extent that they are representations *in the system* rather than representations *for the system* that is, such representations are not accessible to the network *as representations.*

In other words, such a (first-order) network can never know *that* it knows: It simply lacks the appropriate machinery. This points to a fundamental difference between sensitivity and awareness. Sensitivity merely entails the ability to respond in specific ways to certain states of affairs. Sensitivity does not require consciousness in any sense. A thermostat can appropriately be characterized as being sensitive to temperature, just as the carnivorous plant Dionaea Muscipula may appropriately be described as being sensitive to movement on the surface of its leaves. But our intuitions (at least, my intuitions) tell us that such sensitive systems (thermostats, photodiodes, transistors, cameras, carnivorous plants) are not conscious. They do not have "elementary experiences," they simply have no experiences whatsoever. Sensitivity can involve highly sophisticated knowledge, and even learned knowledge, as illustrated by Hinton's (1986) network, but such knowledge is always first-order knowledge, it is always knowledge that is necessarily embedded in the very same causal chain through which first-order processing occurs and that can therefore only be expressed through action as a direct result of perception.

Awareness, on the other hand, always seems to minimally entail the ability of knowing *that* one knows. This ability, after all, forms the basis for the verbal reports we take to be the most direct indication of awareness. And when we observe the absence of such ability to report on the knowledge involved in our decisions, we rightfully conclude that the decision was based on unconscious knowledge. Thus, it is when an agent exhibits *knowledge* of the fact that he is sensitive to some state of affairs that we take this agent to be a conscious agent. This *second-order* knowledge, I argue, critically depends on *learned* systems of meta representations, and forms the basis for conscious experience provided the agent also *cares* about certain states of affairs more than about others.

Consciousness thus not only requires ability to learn about the geography of one's own representations, but it also requires that the resulting knowledge reflects the dispositions and preferences of the agent. This is an important point, for it would be easy to program a thermostat that is capable not only of acting based on the current temperature, but also to report on its own states. Such a talking thermostat would constantly report on the current temperature and on its decisions. Would that make the thermostat conscious? Certainly not, for it is clear that the reporting is but a mere additional process tacked on the thermostat's inherent ability to switch the furnace according to the temperature. What would go some way toward making the thermostat conscious is to set it up so that it *cares* about certain temperatures more than about others, and that these preferences emerge as a result of learning.

What would it take for a network like Hinton's (1986) to be able to access its own representations, and what difference would that make with respect to consciousness? To answer the first question, the required machinery is the machinery of agenthood; in a nutshell, the ability to do something not just with external states of affairs, but rather with one own's representations of such external states. This crucially requires that the agent be able to access,

inspect, and otherwise manipulate its own representations, and this in turn, I surmise, requires mechanisms that make it possible for an agent to redescribe its own representations to itself. The outcome of this continuous "representational redescription" (Karmiloff-Smith, 1992) process is that the agent ends up knowing something about the geography of its own internal states: It has, in effect, *learned* about its own representations. Minimally, this could be achieved rather simply, for instance by having another network take both the input (i.e., the external stimulus as represented proximally) to the first-order network and its internal representations of that stimulus as inputs themselves and do something with them.

One elementary thing the system consisting of the two interconnected networks (the first-order, observed network and the second-order, observing network) would now be able to do is to make decisions, for instance, about the extent to which an external input to the first-order network elicits a familiar pattern of activation over its hidden units or not. This would in turn enable the system to distinguish between hallucination and blindness (see Lau, 2008), or to come up with judgments about the performance of the first-order network (Persaud et al., 2007; Dienes, 2008).

To address the second question (what difference would representational redescription make in terms of consciousness), I appeal to Rosenthal's (1997, 2006) higher-order thought (HOT) theory of consciousness. While I do not feel perfectly happy with all aspects of HOT Theory, I do believe, however, that higher-order representations (I will call them meta-representations in what follows) play a crucial role in consciousness.

An immediate objection to this idea is as follows: If there is nothing intrinsic to the existence of a representation in a cognitive system that makes this representation conscious, why should things be different for meta-representations? After all, meta-representations are representations also. Yes indeed, but with a crucial difference: Meta-representations inform the agent about its own internal states, making it possible for it to develop an understanding of its own workings. And this, I argue, forms the basis for the contents of conscious experience, provided of course – which cannot be the case in an contemporary artificial system – that the system has learned about its representations by itself, over its development, and provided that it cares about what happens to it, that is, provided its behavior is rooted in emotion–laden motivation (to survive, to mate, to find food, etc.).

THE RADICAL PLASTICITY THESIS

I would thus like to defend the following claim: Conscious experience occurs if and only if an information-processing system has *learned* about its own representations of the world in such a way that these representations have acquired value for it. To put this claim even more provocatively: Consciousness is the brain's (emphatically non-conceptual) theory about itself, gained through experience interacting with the world, with other agents, and, crucially, with itself. I call this claim the "*Radical Plasticity Thesis*," for its core is the notion that learning is what makes us conscious.

Before getting to the core of the argument, I should briefly sketch a framework through which to characterize the relationships between learning and consciousness. If the main cognitive function of consciousness is to make adaptive control of behavior possible, as is commonly accepted, then consciousness is necessarily closely related to processes of learning, because one of the central consequences of successful adaptation is that conscious control is no longer required over the corresponding behavior. Indeed, it might seem particularly adaptive for complex organisms to be capable of behavior that does not require conscious control, for instance because behavior that does not require monitoring of any kind can be executed faster or more efficiently than behavior that does require such control. What about conscious experience? Congruently with our intuitions about the role of consciousness in learning, we often say of somebody who failed miserably at some challenging endeavor, such as completing a paper by the deadline, that the failure constitutes "a learning experience." What precisely do we mean by this? We mean that the person can now learn from her mistakes, that the experience of failure was sufficiently imbued with emotional value that it has registered in that person's brain. The experience *hurt*, it made one realize what was at stake, it made us think about it, in other words, it made us consciously aware of what failed and why.

But this minimally requires what Kirsh (1991) has called "explicit representation," namely the presence of representations that directly represent the relevant information. By "direct" here, I mean that the information is represented in such a manner that no further computation is required to gain access to it. For instance, a representation that is explicit in this sense might simply consist of a population of neurons that fire whenever a specific condition holds: A particular stimulus is present on the screen, my body is in a particular state (i.e., pain, or hunger).

By assumption, however, such "explicit" representations are not necessarily conscious. Instead, they are merely good candidates to enter conscious awareness in virtue of features such as their stability, their strength, or their distinctiveness (Cleeremans, 1997; Cleeremans and Jiménez, 2002). What is missing, then? What is missing is that such representations be themselves the target of other representations. And how would this make any difference? It makes a crucial difference, for the relevant first-order *representations* are now part of the agent's known repertoire of mental states; such representations are then, and only then, recognized by the agent as playing the function of representing some other (internal or external) state of affairs.

NECESSARY CONDITIONS FOR AWARENESS

Let us now focus on the set of assumptions that together form the core of a framework that characterizes how learning shapes availability to consciousness (see Cleeremans and Jiménez, 2002; Cleeremans, 2008, for more detailed accounts). It is important to keep it in mind that the framework is based on the connectionist framework (Rumelhart and McClelland, 1986). It is therefore based on many central ideas that characterize the connectionist approach, such as the fact that information processing is graded and continuous, and that it takes place over many interconnected modules consisting of processing units. In such systems, long-term knowledge is embodied in the pattern of connectivity between the processing units of each module and between the modules themselves, while the transient patterns of activation over the units of each module capture the temporary results of information processing.

This being said, a first important assumption is that *representations are graded, dynamic, active, and constantly causally efficacious* (Cleeremans, 1994, 2008). Patterns of activation in neural networks

and in the brain are typically distributed and can therefore vary on a number of dimensions, such as their stability in time, their strength. or their distinctiveness. *Stability* in time refers to how long a representation can be maintained active during processing. There are many indications that different neural systems involve representations that differ along this dimension. For instance, prefrontal cortex, which plays a central role in working memory, is widely assumed to involve circuits specialized in the formation of the enduring representations needed for the active maintenance of task-relevant information. *Strength* of representation simply refers to how many processing units are involved in the representation, and to how strongly activated these units are. As a rule, strong activation patterns will exert more influence on ongoing processing than weak patterns. Finally, *distinctiveness* of representation is inversely related to the extent of overlap that exists between representations of similar instances. Distinctiveness has been hypothesized as the main dimension through which cortical and hippocampal representations differ (McClelland et al., 1995; O'Reilly and Munakata, 2000), with the latter becoming active only when the specific conjunctions of features that they code for are active themselves.

In the following, I will collectively refer to these different dimensions as "quality of representation" (Farah, 1994). The most important notion that underpins these different dimensions is that representations, in contrast to the all-or-none propositional representations typically used in classical theories, instead have a *graded* character that enables any particular representation to convey the extent to which what it refers to is indeed present.

Another important aspect of this characterization of representational systems in the brain is that, far from being static propositions waiting to be accessed by some process, representations instead continuously influence processing regardless of their quality. This assumption takes its roots in McClelland's (1979) analysis of cascaded processing which, by showing how modules interacting with each other need not "wait" for other modules to have completed their processing before starting their own, demonstrated how stage-like performance could emerge out of such continuous, non-linear systems. Thus, even weak, poor-quality traces are capable of influencing processing, for instance through associative priming mechanisms, that is, in *conjunction* with other sources of stimulation. Strong, high-quality traces, in contrast have *generative capacity*, in the sense that they can influence performance independently of the influence of other constraints, that is, whenever their preferred stimulus is present.

A second important assumption is that *learning is a mandatory consequence of information processing*. Indeed, every form of neural information-processing produces adaptive changes in the connectivity of the system, through mechanisms such as long-term potentiation (LTP) or long-term depression (LTD) in neural systems, or Hebbian learning in connectionist systems. An important aspect of these mechanisms is that they are mandatory in the sense that they take place whenever the sending and receiving units or processing modules are co-active. O'Reilly and Munakata (2000) have described Hebbian learning as instantiating what they call *model learning*. The fundamental computational objective of such unsupervised learning mechanisms is to enable the cognitive system to develop useful, informative models of the world by capturing its correlational structure. As such, they stand in contrast with *task learning* mechanisms, which instantiate the different computational objective of mastering specific input–output mappings (i.e., achieving specific goals) in the context of specific tasks through error-correcting learning procedures.

Stability, strength, or distinctiveness can be achieved by different means. Over short time scales, they can result, for instance, from increased stimulus duration, from the simultaneous top-down and bottom-up activation involved in so-called "reentrant processing" (Lamme, 2006), from processes of "adaptive resonance" (Grossberg, 1999), from processes of "integration and differentiation" (Edelman and Tononi, 2000), or from contact with the neural workspace, brought about by "dynamic mobilization" (Dehaene and Naccache, 2001). It is important to realize that the ultimate effect of any of these putative mechanisms is to make the target representations stable, strong, and distinctive. These properties can further be envisioned as involving graded or dichotomous dimensions (see also Maia and Cleeremans, 2005 for an exploration of how connectionist principles are relevant to the study of consciousness).

Over longer time scales, however, high-quality representations arise as a result of learning or cognitive development. Weak, fragile representations become progressively stronger and higher-quality. As a result, they exert more of an influence on behavior. In most cases, this is a good outcome because the stronger a representation is, the less it will require conscious control and monitoring. Thus, in any domain of experience (from being able to stand up to wine-tasting, from recognizing faces to reading) we begin with weak representations, which are characteristic of implicit cognition and do not require control because they only exert weak effects on behavior. Such representations, because of their poor quality, are also only weakly available to form the contents of consciousness. As learning progresses, the relevant representations become stronger, yet not so strong that they can be "trusted" to do their job properly. This is when cognitive control is most necessary. This is also the point where such explicit representations are most likely to form the contents of consciousness. Finally, with further training, the relevant representations become even stronger and eventually fully adapted. As such, these high-quality representations characteristic of automaticity no longer require cognitive control either, but this is so for completely different reasons than the weak representations characteristic of implicit cognition.

Thus, when I respond faster to a target stimulus in virtue of the fact that the target was preceded by a congruent subliminal prime, I can properly say that there exists a state *c* such that its existence made me respond faster, but by assumption I am not sensitive to the fact that this state *c* is different from state *i* where the target stimulus was preceded by an incongruent prime. States *c* and *i* are thus not conscious states – they merely exert their effects on behavior, so reflecting the agent's *sensitivity* to their existence, but crucially not its *awareness* of their existence. The reason such states are not conscious states has to do with the properties of the corresponding first-order states: It is not so much that there is a failure of a higher-order system to target these states, but rather that the first-order states are too weak to be appropriate targets. You cannot know what is not (sufficiently) there.

Likewise, but perhaps more controversially so, habitual, automatic behavior is often described as involving unconscious knowledge: The behavior unfolds whether you intend to or not, it can

unfold with attention engaged elsewhere, and so on. In such cases, behavior is driven by very high-quality representations that have become, through experience, optimally tuned to drive behavior. While such very high-quality representations are appropriate objects for redescription, the redescriptions either no longer play a functional role or are prevented from taking place (for instance because the agent's attention is engaged elsewhere). Automatic behavior is thus not truly unconscious behavior (Tzelgov, 1997). Rather, it is behavior for which awareness has become optional. You can be perfectly aware of behavior that occurs automatically – you just seldom do so for it is neither necessary nor desirable for you to become aware of such behavior. That is precisely *why* the behavior has become automatic: Because it so adapted that it can unfold without the need for conscious monitoring.

Hence a first important computational principle through which to distinguish between conscious and unconscious representations is the following:

> *Availability to consciousness depends on quality of representation, where quality of representation is a graded dimension defined over stability in time, strength, and distinctiveness.*

While being of high-quality thus appears to be a necessary condition for a representation's availability to consciousness, one should ask, however, whether it is a sufficient condition. Cases such as hemineglect or blindsight (Weiskrantz, 1986) clearly suggest that quality of representation alone does not suffice, for even strong stimuli can fail to enter conscious awareness in such conditions. In normal participants, the attentional blink (Shapiro et al., 1997), as well as inattentional (Mack and Rock, 1998) and change blindness (Simons and Levin, 1997), are all suggestive that high-quality stimuli can simply fail to be experienced unless attended to. Likewise, merely achieving stable representations in an artificial neural network, for instance, will not make this network conscious in any sense – this is the problem pointed out by Clark and Karmiloff-Smith (1993) about the limitations of what they called first-order networks: In such networks, even explicit knowledge (e.g., a stable pattern of activation over the hidden units of a standard back-propagation network that has come to function as a "face detector") remains knowledge that is in the network as opposed to knowledge for the network. In other words, such networks might have learned to be informationally sensitive to some relevant information, but they never know that they possess such knowledge. Thus the knowledge can be deployed successfully through action, but only in the context of performing some particular task.

Hence it could be argued that it is a defining feature of consciousness that when one is conscious of something, one is also, at least potentially so, conscious *that* one is conscious of being in that state. This is the gist of so-called HOT theories of consciousness (Rosenthal, 1997), according to which a mental state is conscious when the agent entertains, in an non-inferential manner, thoughts to the effect that it currently is in that mental state. Importantly, for Rosenthal, it is in virtue of occurrent HOTs that the target first-order representations become conscious. Dienes and Perner (1999) have developed this idea by analyzing the implicit-explicit distinction as reflecting a hierarchy of different manners in which the 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 just a possibility that the flower is yellow) and attitude (I know that the flower is yellow). Fully conscious knowledge is thus knowledge that is "attitude-explicit."

This analysis suggests that a further important principle that differentiates between conscious and unconscious cognition is the extent to which a given representation endowed with the proper properties (stability, strength, distinctiveness) is itself the target of meta-representations.

Hence a second important computational principle through which to distinguish between conscious and unconscious representations is the following:

> *Availability to consciousness depends on the extent to which a representation is itself an object of representation for further systems of representation.*

It is interesting to consider under which conditions a representation will remain unconscious based on combining these two principles (Cleeremans, 2008). There are at least four possibilities. First, knowledge that is embedded in the connection weights within and between processing modules can never be directly available to conscious awareness and control. This is simply a consequence of the fact that consciousness, by assumption, necessarily involves representations (patterns of activation over processing units). The knowledge embedded in connection weights will, however, shape the representations that depend on it, and its effects will therefore detectable – but only indirectly, and only to the extent that these effects are sufficiently marked in the corresponding representations. This is equivalent to Dehaene and Changeux's (2004) principle of "active firing."

Second, to enter conscious awareness, a representation needs to be of sufficiently high-quality in terms of strength, stability in time, or distinctiveness. Weak representations are therefore poor candidates to enter conscious awareness. This, however, does not necessarily imply that they remain causally inert, for they can influence further processing in other modules, even if only weakly so. This forms the basis for a host of sub-threshold effects, including, in particular, subliminal priming.

Third, a representation can be strong enough to enter conscious awareness, but failed to be associated with relevant meta-representations. There are thus many opportunities for a particular conscious content to remain, in a way, implicit, not because its representational vehicle does not have the appropriate properties, but because it fails to be integrated with other conscious contents.

Finally, a representation can be so strong that its influence can no longer be controlled – automaticity. In theses cases, it is debatable whether the knowledge should be taken as genuinely unconscious, because it can certainly become fully conscious as long as appropriate attention is directed to them (Tzelgov, 1997), but the point is that such very strong representations can trigger and support behavior without conscious intention and without the need for conscious monitoring of the unfolding behavior.

SUFFICIENT CONDITIONS FOR AWARENESS?

Strong, stable, and distinctive representations are thus *explicit* representations, at least in the sense put forward by Koch (2004): They indicate what they stand for in such a manner that their

reference can be retrieved directly through processes involving low computational complexity (see also Kirsh, 1991, 2003). Conscious representations, in this sense, are explicit representations that have come to play, through processes of learning, adaptation, and evolution, the functional role of denoting a particular content for a cognitive system. Importantly, quality of representation should be viewed as a *graded* dimension. This is essential to capture the fact that phenomenal experience, particularly ordinary phenomenal experience, appears graded itself. Gradedness can be achieved in different ways in a complex system such as the brain. One possibility is that representations are inherently graded because their vehicles are patterns of activation distributed over populations of firing neurons. Another is that representations tend to be all-or-none, but always involve multiple levels of a hierarchy (Kouider et al., 2010).

Once a representation has accrued sufficient strength, stability, and distinctiveness, it may be the target of meta-representations: The system may then "realize," if it is so capable, that is, if it is equipped with the mechanisms that are necessary to support self-inspection, that it has learned a novel partition of the input; that it now possesses a new "detector" that only fires when a particular kind of stimulus, or a particular condition, is present. Humphrey (2006) emphasizes the same point when he states that "This self-monitoring by the subject of his own response is the prototype of the "feeling sensation" as we humans know it" (p. 90). Importantly, my claim here is that such meta-representations are learned in just the same way as first-order representations, that is, by virtue of continuously operating learning mechanisms. Because meta-representations are also representations, the same principles of stability, strength, and distinctiveness therefore apply. An important implication of this observation is that activation of meta-representations can become automatic, just as it is the case for first-order representations.

What might be the function of such meta-representations? One possibility is that their function is to indicate the mental attitude through which a first-order representation is held: Is this something I know, hope, fear, or regret? Possessing such metaknowledge about one's knowledge has obvious adaptive advantages, not only with respect to the agent himself, but also because of the important role that communicating such mental attitudes to others plays in both competitive and cooperative social environments.

What is the mechanism through which such redescription is achieved? Minimally, enabling redescription of one's own internal states requires such internal states to be *available* to redescription, where *availability* is contingent, as described above, on such internal states being *patterns of activation* endowed with certain characteristics such as their strength, their stability in time, and their distinctiveness. Note that these assumptions rule out many potential sources of internal knowledge. For instance, the sort of weak, fleeting representations presumably resulting from the presentation of a brief stimulus would be poor candidates to be available to further processing. Likewise, the associative links that exist between representations, if implemented through patterns of connectivity between groups of units (as they are in connectionist networks) would likewise be inaccessible. Finally, and though this is more speculative (but see Brunel et al., 2010), it may also be the case that the highly distributed representations typical of semantic knowledge (i.e., my knowledge of a typical dog) are less available to form the contents of conscious experience then are the highly distinctive representations characteristic of episodic memory.

Second, those representations that meet these minimal requirements for redescription need to be accessed by another, independent part of the system whose function it is to redescribe them. It is important to note here that mere redescription probably does not cut it, for even in a simple feedforward network, each layer can be thought of as being a redescription of the input. The brain is massively hierarchical and thus contains multiple such redescriptions of any input. Instead of being strictly hierarchically organized, however, the redescriptions that count for the mechanism I have in mind should be removed from the causal chain responsible for the first-order processing. Hence, we need some mechanism that can access and redescribe first-order representations in a manner that is independent from the first-order causal chain.

I suggest that the general form of such mechanisms is something similar to what is depicted in **Figure 1**. Two independent networks (the first-order network and the second-order network) are connected to each other in such a way that the entire first-order network is input to the second-order network. Both networks are simple feedforward back-propagation networks. The first-order network consists of thee pools of units: a pool of input units, a pool of hidden units, and a pool of output units. Let us further imagine that this network is trained to perform a simple discrimination task, that is, to produce what is named Type I response in the language of Signal-Detection Theory. My claim is that there is nothing in the computational principles that characterize how this network performs its task that is intrinsically associated with awareness. The network simply performs the task. While it will develop knowledge of the associations between its inputs and outputs over its hidden units, and while this knowledge may be in some cases very sophisticated, it will forever remain knowledge that is "in" the network as opposed to being knowledge "for" the network. In other words, such a (first-order) network can never know *that* it knows: It simply lacks the appropriate machinery to do so. Likewise, in Signal-Detection Theory, while Type 1 responses always reflect sensitivity to some state of affairs, this sensitivity may or may not be conscious sensitivity. That is, a participant may be successful in discriminating one stimulus from another, yet fail to be aware *that* he is able to do so and thus claim, if asked, that he is merely guessing or responding randomly. In its more general form, as depicted in **Figure 1**, such an architecture would also be sufficient for the second-order network to also perform other judgments, such as distinguishing between an hallucination and a veridical perception, or developing knowledge about the overall geography of the internal representations developed by the first-order network (see also Nelson and Narens, 1990).

Can we use such architectures to account for relevant data? That is the question we set out to answer in recent work (e.g., Cleeremans et al., 2007; Pasquali et al., 2010) aimed at exploring the relationships between performance and awareness. We have found that different approaches to instantiating the general principles we have described so far are required to capture empirical findings. In one, as hinted at above, the first-order and the second-order network are part of the same causal chain, but are trained on different tasks, one corresponding to first-order decisions and the second

corresponding to metacognitive decisions. In a second approach, the two networks are truly independent. Note that in either case, our assumptions are oversimplified, for a complete implementation of the theory would require that the second-order network may influence processing as it takes place in the first-order network by means of recurrence.

In the following, I will illustrate the first approach, through which we have focused on architectures in which the first and second-order networks function as part of the same causal chain. Post-decision wagering was recently introduced by Persaud et al. (2007) as a measure of awareness through which participants are required to place a high or a low wager on their decision, such as relative to stimulus identification for example. The intuition behind this measure is that people will place a high wager when they have conscious knowledge about the reasons for their decisions, and a low wager when they are uncertain of their decisions. In this, wagering is thus similar to other subjective measures of awareness (Seth et al., 2008; Sandberg et al., 2010). According to Persaud et al. (2007) wagering provides an incentive for participants not to withhold any conscious information, as well as not to guess, making it a more objective measure of awareness than confidence judgment. Despite recent criticism of Persaud et al.'s claims (Dienes and Seth, 2010; Sandberg, et al., 2010), wagering certainly reflects the extent to which an agent is sensitive to its own internal states. In Cleeremans et al. (2007), we therefore aimed at creating a wagering network, for wagering affords easy quantification and thus appeared more readily amenable to computational simulation than other metacognitive measures such as confidence. In one of our simulations, which I will describe in more detail here, the first-order feedforward back-propagation network (see **Figure 2**) consisted of 7 input units representing digit shapes (as on a digital watch), 100 hidden units, and 10 output units for the 10 digits. The task of the first-order network is a simple one: It consists of identifying the "visual" representations of the digits 0–9. This is achieved by training the first-order network to respond to each input by activating one of its 10 output units. The 100 first-order hidden units connected to a different pool of 100 hidden units of the second-order feedforward network, with 2 output units representing a high and a low wager, as shown in **Figure 2**. The task of the higher-order network consisted of wagering high if it "thought" that the first-order network was providing a correct answer (correct identification of the digit), and to wager low in case the first network gave a wrong answer (misidentification of the digit). Note that as implemented here, there is no substantial difference between wagering and merely expressing confidence judgments.

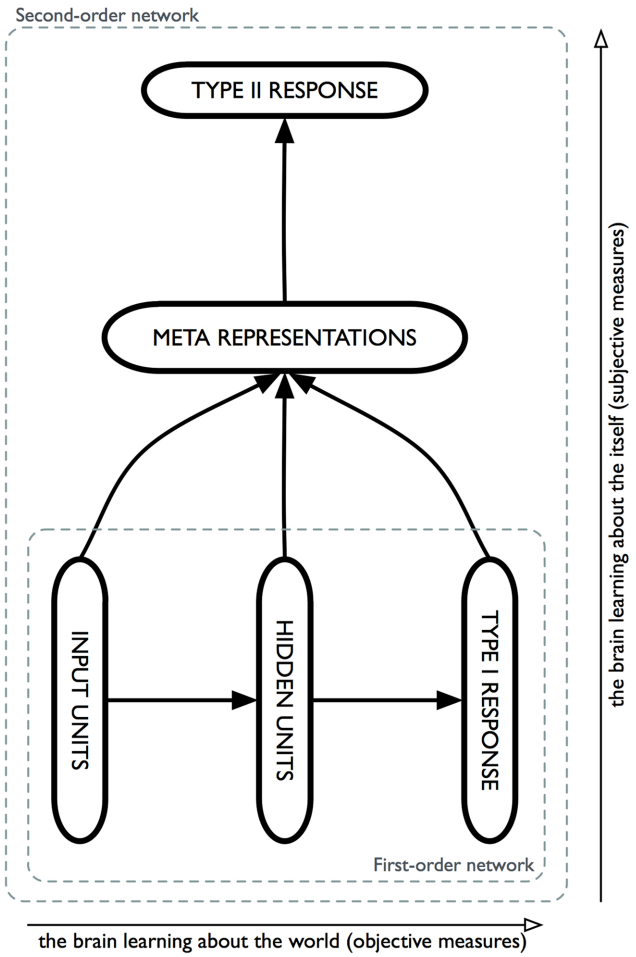

FIGURE 1 | General architecture of a metacognitive network. A first-order network, consisting for instance of a simple three-layers back-propagation network, has been trained to perform a simple classification task and thus contains knowledge that links inputs to outputs in such a way that the network can produce Type I responses. By design, this entire first-order network then constitutes the input to a second-order network, the task of which consists of redescribing the activity of the first-order network in some way. Here, the task that this second-order network is trained to perform is to issue Type II responses, that is, judgments about the extent to which the first-order network has performed *its* task correctly. One can think of the first-order network as instantiating cases where the brain learns about the world, and of the second-order network as instantiating cases where the brain learns about itself.

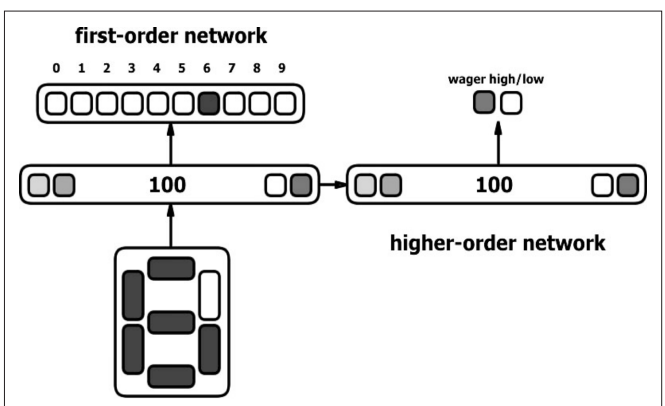

FIGURE 2 | Architecture of a wagering network. A first-order network instantiates a simple pattern classifier trained to classify "visual" input patterns representing the shapes of digits 0–9 in 10 categories. A second-order network is assigned the task of wagering on the first-order network's performance based on the latter's internal representations of the stimulus. The second-order network thus performs judgments about the extent to which the first-order network is correct in its own decisions.

A learning rate of 0.15 and a momentum of 0.5 were used during training of the first-order network. In a first condition of "high awareness," the second network was trained with a learning rate of 0.1, and in a second condition of "low awareness," a learning rate of 10^{-7} was applied. Ten networks were trained to perform their tasks concurrently throughout 200 epochs of training and their performance averaged. The performance of all three networks is depicted in **Figure 3**. Chance level for the first-order network is 10% (there is one chance of out 10 of correctly identifying one digit amongst 10); it is 50% for the second-order network (one chance out of two of placing a correct bet). The figure shows that the first-order network simply gradually learns to improve its classification performance continuously until it achieves 100% correct responses at the end of training. The performance of the "high awareness" second-order network, however, exhibits a completely different pattern. Indeed, one can see that the second-order network initially performs quite well, only to show decreasing performance up until about epoch 40, at which point its performance has sagged to chance level. From epoch 40 onwards, the second-order network's performance increases in parallel with that of the first-order network. This u-shaped performance pattern is replicated, to a lesser degree and with slightly different dynamics, in the "low awareness" second-order network.

One can understand this performance pattern as follows. Initially, the second-order network quickly learns that the first-order network is systematically incorrect in classifying the digits. (which is expected since it has not begun to learn how to perform the task). The safest response (i.e., the response that minimizes error) is thus to always bet low. This, incidentally, is what any rational agent would do. However, as the first-order network quickly begins to exceed chance level performance on its digit classification task, the performance of the second-order network begins to decrease. This corresponds to a stage where the second-order network is beginning to bet "high" on some occasions as it learns to categorize states of the first-order network that are predictive of a correct classification. An interesting pattern of dissociation then occurs, for the second-order network is performing rather poorly just when the first-order network is beginning to truly master its own digit classification task. One can think of that stage as corresponding to a point in training where the system as a whole is essentially acting based on unconscious knowledge: First-order performance on the digit classification task is well above chance level, yet, wagering by the second-order network is close to chance, and is at chance on epoch 40. Later on, after epoch 40, the second-order network has learned enough about when the first-order network will be correct vs. incorrect to begin attempting to maximize its own wagering performance. Thus, epoch 40 corresponds to the second-order network's "most doubtful moment." One could view this as the moment at which the higher-order network abandons a simple "safe" strategy of low wagers and explores the space of first-order hidden unit representations, looking for a criterion that will allow it to separate good from bad identifications.

Thus, as the two networks learn simultaneously to perform their respective tasks, one sees the entire system shifting from a situation where there is no relationship between first- and second-order performance to a situation where the two are correlated. This transition reflects, under our assumptions, a shift between unconscious vs. conscious processing.

In later work (Pasquali, et al., 2010), we have explored similar models based on germane or identical architectures and shown that they are capable of accounting for the data reported by Persaud et al.

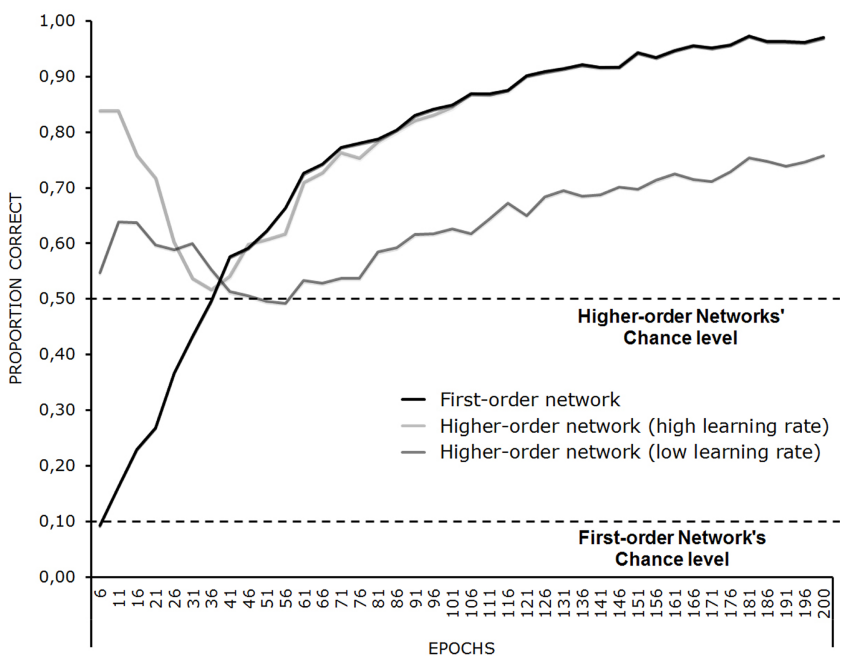

FIGURE 3 | Performance of the first-order and second-order networks, as a function of training expressed as number of epochs.

(2007) in three different domains: Artificial Grammar Learning, Blindsight, and the Iowa Gambling Task. In all three cases, our simulations replicate the patterns of performance observed in human participants with respect to the relationship between task performance and wagering. The blindsight and Artificial Grammar learning simulations instantiate the second approach briefly described above in that they use an architecture in which the processing carried out in second-order network is completely independent from that carried out in the first-order network. In such architectures, the two networks are connected by means of fixed connections that instantiate "comparator units." The Iowa Gambling Task simulation, on the other hand, relies on the same mechanisms as described for the digits task. Interestingly, in this latter case, we were able to additionally capture the fact that asking participants to reflect upon their own performance helps them improve metacognitive awareness (Maia and McClelland, 2004) and hence, the relationship between first-order performance and wagering. The fact that the relationships between first-order and metacognitive performance can vary as a function of task instructions is borne out by a recent study of Fleming et al. (2010) which indicates large individual differences in people's ability to judge their own performance. Strikingly, the authors found that differences in metacognitive ability were subtended not only by differences in the activity of anterior prefrontal cortex, but also by structural differences in the white matter of these regions.

It may seem that the proposed mechanism is identical with signal-detection accounts of metacognition (e.g., Scott and Dienes, 2008). However, there is a crucial difference. Signal-detection accounts typically make the second-order distinction between confidence and guessing (high vs. low wagers) on the very signal that is used for first-order classifications by setting two boundaries on the signal: One boundary that accounts for the first-order classification, and a second boundary (on either side of the first-order boundary) that distinguishes between guessing (cases that fall within the area defined by the second boundaries) and cases that fall outside of these boundaries (on the extremes of the distribution). In such an account, confidence thus depends directly on first-order signal strength (but see Maniscalco and Lau, 2010; Pleskac and Busemeyer, 2010 for further discussion). However, in some of the models we have proposed, the second-order classification does not depend on the same signal as the first-order task. Indeed, instead of wagering high or low based on signal strength, the second-order network re-represents the first-order error as a new pattern of activation. Thus, before it can wager correctly, the second-order network, like the first-order network, has to learn to make a new, single-boundary classification based on this second-order representation (the error representation). Thus, the second-order network actually learns to judge the first-order network's performance independently of the first-order task itself. The difference between our model and Signal-Detection Theory is substantial, for it impinges on whether one considers Type I and Type II performance, that is, first-order and second-order judgments about these decisions entertain hierarchical or parallel relationships with each other. This issue is currently being debated, with some authors defending a dual-route model (Del Cul et al., 2009; Dehaene and Charles, 2010) and others (Lau, 2010; Maniscalco and Lau, 2010) defending hierarchical models. The simulation work described in Pasquali et al. (2010) explored both approaches by means of distinct architectures. Clearly, additional research is necessary to clarify the predictions of each approach and to further delineate their mechanisms.

Beyond giving a cognitive system the ability to learn about its own representations, there is another important function that meta-representations may play: They can also be used to anticipate the future occurrences of first-order representations (see Bar, 2009, on the human brain as a prediction machine). Thus for instance, if my brain learns that SMA is systematically active before M1, then it can use SMA representations to explicitly represent their consequences downstream, that is, M1 activation, and ultimately, action. If neurons in SMA systematically become active before an action is carried out, a metarepresentation can link the two and represent this fact explicitly in a manner that will be experienced as intention. That is: When neurons in the SMA become active, I experience the feeling of intention *because* my brain has learned, unconsciously, that such activity in SMA precedes action. It is this knowledge that gives qualitative character to experience, for, as a result of learning, each stimulus that I see, hear, feel, or smell is now not only represented, but also re-represented through independent meta-representations that enrich and augment the original representation(s) with knowledge about (1) how similar the manner in which the stimulus' representation is with respect to that associated with other stimuli, (2) how similar the stimulus' representation is now with respect to what it was before, (3) how consistent is a stimulus' representation with what it typically is, (4) what other regions of my brain are active at the same time that the stimulus' representation is, etc.

To see how this is different from mere first-order knowledge, consider what happens in the case of hallucination. Imagine a simple three-layers network akin to those described above in which a first layer of units receives perceptual input and is connected to a second layer of internal ("hidden") units that are in turn connected to response units. One can easily train such a simple system to produce specific outputs in response to specific inputs (i.e., activating the "9" unit when presented with the visual pattern corresponding to the digit "9"). After training, each input will cause the emergence of a specific (learned) pattern of action over the network's hidden units, and this will in turn cause a specific response. Crucially, one can now induce a specific response by either presenting a familiar pattern over the network's input units (as it would be in the case of a genuine perception) or by directly activating the network's hidden units with the learned pattern corresponding to that same input (as it could be, for instance, in the case of a memory retrieval whereby the pattern is reinstated by means of other pathways). The point is that the network would respond in exactly the same way in both cases for it simply lacks the ability to identify whether its response was caused by the activation of its input units or by the activation of its hidden units in the absence of any input. In other words, such a network is unable to distinguish between a veridical perception and an hallucination. Doing so would require the existence of another, independent network, whose task it is to learn to associate specific input patterns with specific patterns of activity of the first network's hidden units. That system would then be able to identify cases where the latter exists in the absence of the former, and hence, to learn to distinguish between cases of veridical perception and cases of hallucination. Such internal monitoring is viewed here as

constitutive of conscious experience: A mental state is a conscious mental state when the system that possesses this mental state is (at least non-conceptually) sensitive to its existence. Thus, and unlike what is assumed to be case in HOT Theory, meta-representations can be both subpersonal and non-conceptual.

Overall, this perspective is thus akin to the sensorimotor or enactive perspective (O'Regan and Noë, 2001) and to the general conceptual framework provided by forward modeling (e.g., Wolpert et al., 2004) in the sense that awareness is linked with knowledge of the consequences of our actions, but, crucially, the argument is extended inwards, that is, to the entire domain of neural representations. It can also be extended further outwards, specifically toward social cognition (see also Graziano and Karstner, in press). Our representations of ourselves are shaped by our history of interactions with other agents. Learning about the consequences of the actions that we direct toward other agents uniquely require more sophisticated models of such other agents than when interacting with objects, for agents, unlike objects can react to actions directed toward them in many different ways as a function of their own internal state. A further important point here is that caretakers act as external selves during development, interpreting what happens to developing children for them, and so providing meta-representations where they lack. In this light, theory of mind can thus be understood as rooted in the very same mechanisms of predictive redescriptions as involved when interacting with the world or with one self.

CONCLUSION

Thus we end with the following idea, which is the heart of the "Radical Plasticity Thesis": The brain continuously and unconsciously learns not only about the external world and about other agents, but also about its own representations of both. The result of this unconscious learning is conscious experience, in virtue of the fact that each representational state is now accompanied by (unconscious learnt) meta-representations that convey the mental attitude with which the first-order representations are held. From this perspective thus, there is nothing intrinsic to neural activity, or to information *per se*, that makes it conscious. Conscious experience involves specific mechanisms through which particular (i.e., stable, strong, and distinctive) unconscious neural states become the target of further processing, which I surmise involves some form of representational redescription in the sense described by Karmiloff-Smith (1992). These ideas are congruent both with higher-order theories in general (Rosenthal, 1997; Dienes and Perner, 1999), and with those of Lau (2008), who has characterized consciousness as "signal detection on the mind."

In closing, there is one dimension that I feel is sorely missing from contemporary discussion of consciousness: Emotion (but see, e.g., Damasio, 1999, 2010; LeDoux, 2002; Tsuchiya and Adolphs, 2007). Emotion is crucial to learning, for there is no sense in which an agent would learn about anything if the learning failed to *do something* to it. Conscious experience not only requires an experiencer who has *learned* about the geography of its own representations, but it also requires experiencers who *care* about their experiences.

ACKNOWLEDGMENTS

Axel Cleeremans is a Research Director with the National Fund for Scientific Research (FNRS, Belgium). This work was supported by an institutional grant from the Université Libre de Bruxelles to Axel Cleeremans and by Concerted Research Action 06/11-342 titled "Culturally modified organisms: What it means to be human in the age of culture," financed by the Ministère de la Communauté Française – Direction Générale l'Enseignement non obligatoire et de la Recherche scientifique (Belgium). Portions of this article were adapted from the following publication: Cleeremans (2008), Consciousness: The Radical Plasticity Thesis, In R. Banerjee and B.K. Chakrabati (Eds.), Progress in Brain Science, *168*, 19–33.

REFERENCES

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

Bar, M. (2009). Predictions: a universal principle in the operation of the human brain. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 364, 1181–1182.

Brunel, L., Oker, A., Riou, B., and Versace, R. (2010). Memory and consciousness: trace distinctiveness in memory retrievals. *Conscious. Cogn.* 19, 926–937.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York, NY: Oxford University Press.

Chalmers, D. J. (2007a). "The hard problem of consciousness," in *The Blackwell Companion to Consciousness*, eds M. Velmans and S. Schneider (Oxford: Blackwell Publishing), 225–235.

Chalmers, D. J. (2007b). "Naturalistic dualism," in *The Blackwell Companion to Consciousness*, eds M. Velmans and S. Schneider (Oxford: Blackwell Publishing), 359–368.

Clark, A., and Karmiloff-Smith, A. (1993). The cognizer's innards: a psychological and philosophical perspective on the development of thought. *Mind Lang.* 8, 487–519.

Cleeremans, A. (1994). Awareness and abstraction are graded dimensions. *Behav. Brain Sci.* 17, 402–403.

Cleeremans, A. (1997). "Principles for implicit learning," in *How Implicit is Implicit Learning?* ed. D. C. Berry (Oxford: Oxford University Press), 195–234.

Cleeremans, A. (2008). "Consciousness: the radical plasticity thesis," in *Models of Brain and Mind: Physical, Computational and Psychological Approaches. Progress in Brain Research*, Vol. 168, eds R. Banerjee and B. K. Chakrabarti (Amsterdam: Elsevier), 19–33.

Cleeremans, A., and Jiménez, L. (2002). "Implicit learning and consciousness: a graded, dynamic perspective," in *Implicit Learning and Consciousness: An Empirical, Computational and Philosophical Consensus in the Making?* eds R. M. French and A. Cleeremans (Hove: Psychology Press), 1–40.

Cleeremans, A., Timmermans, B., and Pasquali, A. (2007). Consciousness and metarepresentation: a computational sketch. *Neural Netw.* 20, 1032–1039.

Damasio, A. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York, NY: Harcourt Brace and Company.

Damasio, A. (2010). *Self Comes to Mind: Constructing the Conscious Brain*. London: William Heinemann.

Dehaene, S., and Changeux, J.-P. (2004). "Neural mechanisms for access to consciousness," in *The Cognitive Neurosciences*, 3rd Edn, ed. M. Gazzaniga (New York: W.W. Norton), 1145–1157.

Dehaene, S., and Charles, L. (2010). "A dual-route theory of evidence accumulation during conscious access," in *Paper Presented at the 14th Annual Meeting of the Association for the Scientific Study of Consciousness*, Toronto.

Dehaene, S., Kerszberg, M., and Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl. Acad. Sci. U.S.A.* 95, 14529–14534.

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

Del Cul, A., Dehaene, S., Reyes, P., Bravo, E., and Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain* 132, 2531–2540.

Dennett, D. C. (1991). *Consciousness Explained*. Boston, MA: Little, Brown and Co.

Dennett, D. C. (2001). Are we explaining consciousness yet? *Cognition* 79, 221–237.

Dienes, Z. (2008). "Subjective measures of unconscious knowledge," in

Models of Brain and Mind. Physical, computational and Psychological Approaches. Progress in Brain Research, Vol. 168, eds R. Banerjee and B. K. Chakrabarti (Amsterdam: Elsevier), 49–64.

Dienes, Z., and Perner, J. (1999). A theory of implicit and explicit knowledge. *Behav. Brain Sci.* 22, 735–808.

Dienes, Z., and Seth, A. (2010). Gambling on the unconscious: a comparison of wagering and confidence as measures of awareness in an artificial grammar task. *Conscious. Cogn.* 19, 674–681.

Edelman, G. M., and Tononi, G. (2000). *Consciousness. How Matter Becomes Imagination*. London: Penguin Books.

Farah, M. J. (1994). Neuropsychological inference with an interactive brain: a critique of the "locality" assumption. *Behav. Brain Sci.* 17, 43–104.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., and Rees, G. (2010). Relating instrospective accuracy to individual differences in brain structure. *Science* 329, 1541–1543.

Graziano, M., and Karstner, S. (in press). Human consciousness and its relationship to social neuroscience: a novel hypothesis. *Cogn. Neurosci.*

Grossberg, S. (1999). The link between brain learning, attention, and consciousness. *Conscious. Cogn.* 8, 1–44.

Hinton, G. E. (1986). "Learning distributed representations of concepts," in *Proceedings of the Eighth Annual Conference of the Cognitive Science Society,* Amherst, MA.

Humphrey, N. (1971). Colour and brightness preferences in monkeys. *Nature* 229, 615–617.

Humphrey, N. (2006). *Seeing Red*. Cambridge, MA: Harvard University Press.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge: MIT Press.

Kirsh, D. (1991). "*When is information explicitly represented?*" in *Information, Language, and Cognition*, ed. P. P. Hanson (New York, NY: Oxford University Press).

Kirsh, D. (2003). "Implicit and explicit representation," in *Encyclopedia of Cognitive Science*, Vol. 2, ed. L. Nadel (London: Macmillan), 478–481.

Koch, C. (2004). *The Quest for Consciousness. A Neurobiological Approach*. Englewood, CO: Roberts and Company Publishers.

Kouider, S., de Gardelle, V., Sackur, J., and Dupoux, E. (2010). How rich is consciousness: the partial awareness hypothesis. *Trends Cogn. Sci. (Regul. Ed.)* 14, 301–307.

Kreiman, G., Fried, I., and Koch, C. (2002). Single-neuron correlates of subjective vision in the human medial temporal lobe. *Proc. Natl. Acad. Sci. U.S.A.* 99, 8378–8383.

Lamme, V. A. F. (2003). Why visual attention and awareness are different. *Trends Cogn. Sci. (Regul. Ed.)* 7, 12–18.

Lamme, V. A. F. (2006). Toward a true neural stance on consciousness. *Trends Cogn. Sci. (Regul. Ed.)* 10, 494–501.

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*, Vol. 168, eds R. Banerjee and B. K. Chakrabarti (Amsterdam: Elsevier), 35–48.

Lau, H. (2010). "Comparing different signal processing architectures that support conscious reports," in *Paper Presented at the 14th Annual Meeting of the Association for the Scientific Study of Consciousness*, Toronto.

LeDoux, J. (2002). *Synaptic Self*. Harmondsworth: Viking Penguin.

Mack, A., and Rock, I. (1998). *Inattentional Blindness*. Cambridge, MA: MIT Press.

Maia, T. V., and Cleeremans, A. (2005). Consciousness: converging insights from connectionist modeling and neuroscience. *Trends Cogn. Sci. (Regul. Ed.)* 9, 397–404.

Maia, T. V., and McClelland, J. L. (2004). A reexamination of the evidence for the somatic marker hypothesis: what participants really know in the Iowa Gambling Task. *Proc. Natl. Acad. Sci. U.S.A.* 101, 16075–16080.

Maniscalco, B., and Lau, H. (2010). Comparing signal detection models of perceptual decision confidence. *J. Vis.* 10, 213.

McClelland, J. L. (1979). On the time-relations of mental processes: an examination of systems in cascade. *Psychol. Rev.* 86, 287–330.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* 102, 419–457.

Nagel, T. (1974). What is like to be a bat? *Philos. Rev.* 83, 434–450.

Nelson, T. O., and Narens, L. (1990). Metamemory: a theoretical framework and new findings. *Psychol. Learn. Motiv.* 26, 125–173.

O'Brien, G., and Opie, J. (1999). A connectionist theory of phenomenal experience. *Behav. Brain Sci.* 22, 175–196.

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

O'Reilly, R. C., and Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, MA: MIT Press.

Pasquali, A., Timmermans, B., and Cleeremans, A. (2010). Know thyself: metacognitive networks and measures of consciousness. *Cognition* 117, 182–190.

Persaud, N., McLeod, P., and Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nat. Neurosci.* 10, 257–261.

Pleskac, T. J., and Busemeyer, J. (2010). Two-stage dynamic signal detection: a theory of confidence, choice, and response time. *Psychol. Rev.* 117, 864–901.

Rosenthal, D. (1997). "A theory of consciousness," in *The Nature of Consciousness: Philosophical Debates*, eds N. Block, O. Flanagan, and G. Güzeldere (Cambridge, MA: MIT Press), 729–753.

Rosenthal, D. (2006). *Consciousness and Mind*. Oxford: Oxford University Press.

Rumelhart, D. E., and McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press.

Sandberg, K., Timmermans, B., Overgaard, M., and Cleeremans, A. (2010). Measuring consciousness: is one measure better than the other? *Conscious. Cogn.* 19, 1069–1078.

Scott, R. B., and Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *J. Exp. Psychol. Learn. Mem. Cogn.* 34, 1264–1288.

Seth, A., Dienes, Z., Cleeremans, A., Overgaard, M., and Pessoa, L. (2008). Measuring consciousness: relating behavioural and neuropsychological approaches. *Trends Cogn. Sci. (Regul. Ed.)* 12, 314–321.

Shapiro, K. L., Arnell, K. M., and Raymond, J. E. (1997). The attentional blink. *Trends Cogn. Sci. (Regul. Ed.)* 1, 291–295.

Simons, D. J., and Levin, D. T. (1997). Change blindness. *Trends Cogn. Sci. (Regul. Ed.)* 1, 261–267.

Tononi, G. (2003). "Consciousness differentiated and integrated," in *The Unity of Consciousness: Binding, Integration, and Dissociation*, ed. A. Cleeremans (Oxford: Oxford University Press), 253–265.

Tononi, G. (2007). "The information integration theory," in *The Blackwell Companion to Consciousness*, eds M. Velmans and S. Schneider (Oxford: Blackwell Publishing), 287–299.

Tsuchiya, N., and Adolphs, R. (2007). Emotion and consciousness. *Trends Cogn. Sci. (Regul. Ed.)* 11, 158–167.

Tzelgov, J. (1997). "Automatic but conscious: that is how we act most of the time," in *The Automaticity of Everyday life*, Vol. X, ed. R. S. Wyer (Mahwah, NJ: Lawrence Erlbaum Associates), 217–230.

Weiskrantz, L. (1986). *Blindsight: A Case Study and Implications*. Oxford: Oxford University Press.

Wolpert, D. M., Doya, K., and Kawato, M. (2004). "A unifying computational framework for motor control and social interaction," in *The Neuroscience of Social Interaction*, eds C. D. Frith and D. M. Wolpert (Oxford: Oxford University Press), 305–322.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 19 November 2010; paper pending published: 25 February 2011; accepted: 26 April 2011; published online: 09 May 2011.
Citation: Cleeremans A (2011) The radical plasticity thesis: how the brain learns to be conscious. Front. Psychology **2**:*86. doi: 10.3389/fpsyg.2011.00086*
This article was submitted to Frontiers in Consciousness Research, a specialty of Frontiers in Psychology.