R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 3

Consciousness: the radical plasticity thesis

Axel Cleeremans*

Cognitive Science Research Unit, Université Libre de Bruxelles CP 191, 50 ave. F.-D. Roosevelt, B1050 Brussels, Belgium

Abstract: In this chapter, I sketch a conceptual framework which takes it as a starting point that conscious and unconscious cognition are rooted in the same set of interacting learning mechanisms and representational systems. On this view, the extent to which a representation is conscious depends in a graded manner on properties such as its stability in time or its strength. Crucially, these properties are accrued as a result of learning, which is in turn viewed as a mandatory process that always accompanies information processing. From this perspective, consciousness is best characterized as involving (1) a graded continuum defined over "quality of representation", such that availability to consciousness and to cognitive control correlates with quality, and (2) the implication of systems of metarepresentations. A first implication of these ideas is that the main function of consciousness is to make flexible, adaptive control over behavior possible. A second, much more speculative implication, is that we learn to be conscious. This I call the "radical plasticity thesis" — the hypothesis that consciousness emerges in systems capable not only of *learning* about their environment, but also about their own internal representations of it.

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

Information processing can undoubtedly take place without consciousness, as abundantly demonstrated by empirical evidence, 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, 2007b) that comes over and beyond mere computation.

With this premise in mind — a premise that just restates Chalmers' *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 (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

*Corresponding author. Tel.: +32 2 650 32 96; Fax: +32 2 650 22 09; E-mail: axcleer@ulb.ac.be

DOI: 10.1016/S0079-6123(07)68003-0

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, *contra* Chalmers, 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), or the involvement of higher order representations (Rosenthal, 1997), to name just a few.

Another take on this most difficult question 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 Chalmer's 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, 2007a, 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.

However, what 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 (2007b) 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. It is interesting to note that Indian philosophical traditions have placed similar emphasis on the role that emotion plays in shaping conscious experience (Banerjee, in press).

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 is 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, a first point about what we mean by ''experience'' is that 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 does 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 in which it is embedded. In other words, it lacks knowledge *that* it can find itself in different states; it is but a mere object that responds to inputs in certain ways that one can fully describe by the laws of physics. 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.

This point can also 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 a simple back-propagation network can learn about abstract dimensions of the training set. Hinton's 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 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 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 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 *aware* of nationality, in the sense that it is sensitive to the concept. 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 or awareness of 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*.

What would it take for a network like Hinton's 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's own 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, in press), or to come up with judgments about the performance of the first-order network (Persaud et al., 2007; Dienes, in press).

To address the second question (what difference would representational redescription make in terms of consciousness), if you think this is starting to sound like a higher order thought theory of consciousness (Rosenthal, 1997), you may be right. While I do not feel perfectly happy with all aspects of Higher-Order Thought Theory, I do believe, however, that higher order representations (I will call them metarepresentations 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 metarepresentations? After all, metarepresentations are representations also. Yes indeed, but with a crucial difference. Metarepresentations 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. To put this claim even more provocatively: consciousness is the brain's theory about itself, gained through experience interacting with the world, 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. How so? The short answer, as hinted above, is that consciousness involves not only knowledge about the world, but crucially, knowledge about our own internal states, or mental representations. When I claim to be conscious of a stimulus, I assert my ability to discriminate cases where the stimulus is present from cases where it is not. But what is the basis of this ability, given that I have no direct access to the stimulus? The answer is obvious: some neural states correlate with the presence or absence of the stimulus, and I make judgments about these states to come to a decision.

Note that this is the way in which *all* information processing takes place, with or without consciousness. After all, we *never* have direct access to anything that is part of the world in which we are embedded; any perception necessarily involves mediation through neural states, which in this sense are appropriately characterized as internal representations of external states of affairs.

What, then, differentiates cases where one is conscious of a state of affairs from cases where one remains unaware of it? It is obvious that in the first case, the relevant representations are accompanied by subjective experience whereas in the second, they are not.

This difference is in fact what motivates Baars' "contrastive approach", through which one seeks to identify differences between information processing with and without consciousness by "treating consciousness as a variable", that is, by designing experimental paradigms in which the only difference of interest is one of conscious awareness. The same idea underpins what neuroscientists call the "search for the neural correlates of consciousness" (Frith et al., 1999). Here, the goal is to identify cerebral regions, neural processes, or processing pathways where one finds activity that correlates not with some objective state of affairs (i.e., a stimulus), but rather with people's own subjective reports that they are conscious of that state of affairs.

As Lau (this volume) points out, however, things are not so simple, for this approach rests on the premise that one can indeed design an experimental situation in which consciousness is the *only* difference. This, as it turns out, is extremely difficult to achieve, precisely because consciousness does make a difference! In other

words, performance at a given task will also be different depending on whether the subject is conscious or not of the relevant state of affairs.

In the following, I would now like to present 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, 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, strength, or distinctiveness (Cleeremans, 2005, 2006). 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 repertoire of mental states; such representations are then, and only then, recognized as playing the function of representing some other (external) state of affairs.

A learning-based account of consciousness

I would now like to introduce the set of assumptions that together form the core of the framework (see Cleeremans and Jiménez, 2002; Cleeremans, in preparation, for more detailed accounts). It is important to keep it in mind that the framework is based on the connectionist framework (e.g., 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). 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" (see also 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.

Having put in place assumptions about representations and learning, the central ideas that I would now like to explore are (1) that the extent to which a particular representation is available to consciousness depends on its quality, (2) that learning produces, over time, higher quality (and therefore adapted) representations, and (3) that the function of consciousness is to offer necessary control over those representations that are strong enough to influence behavior, yet not sufficiently adapted that their influence does not require control anymore.

Figure 1 aims to capture these ideas by representing the relationships between quality of representation (X-axis) on the one hand and (1) potency, or the extent to which a representation can influence behavior, (2) availability to control, (3) availability to subjective experience. I discuss the figure at length in the following section. Let us simply note here that the X-axis represents a continuum between weak, poor-quality representations on the left

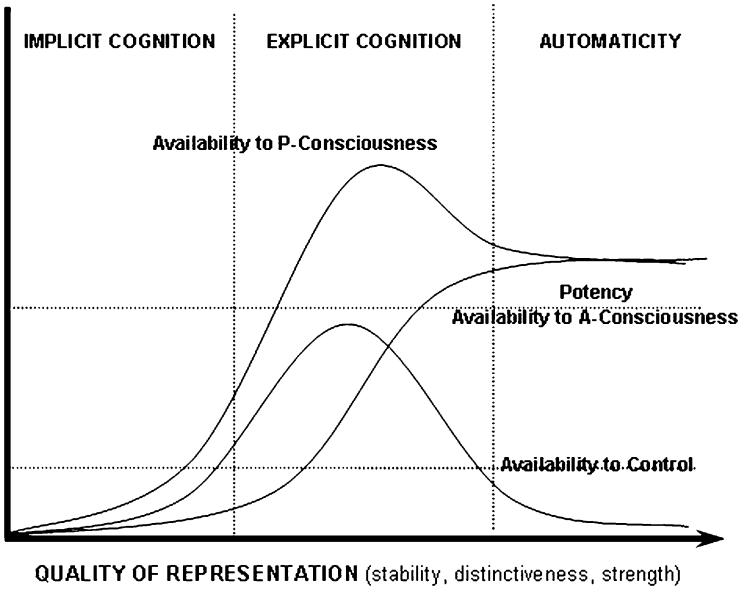

Fig. 1. Graphical representation of the relationships between quality of representation (X-axis) and (1) potency, (2) availability to control, (3) availability to subjective experience. See text for further details.

and very strong, high-quality representations on the right.

Two further points are important to be kept in mind with respect to Fig. 1. First, the relationships depicted in the figure are intended to represent *availability* to some dimension of behavior or consciousness independently of other considerations. Many potentially important modulatory influences on the state of any particular module are thus simply not meant to be captured neither by Fig. 1, nor by the framework presented here. Second, the figure is intended to represent what happens in *each* of the many processing modules involved in any particular cognitive task. Thus, at any point in time, there will be many such modules active, each contributing to some extent to behavior and to conscious experience; each modulating the activity of other modules. With these caveats in mind, let me now turn to four assumptions about consciousness and its relationship with learning:

Assumption C1: *Consciousness involves two dimensions: subjective experience and control*

As argued by Block (1995, 2005) and even though there is continuing debate about this issue, consciousness involves at least two separable aspects, namely access consciousness (A-consciousness) and phenomenal consciousness (P-consciousness). According to Block (1995), "A perceptual state is access-conscious roughly speaking if its content — what is represented by the perceptual state — is processed via that information processing function, that is, if its content gets to the Executive system, whereby it can be used to control reasoning and behavior" (p. 234). In other words, whether a state is A-conscious is defined essentially by the causal efficacy of that state; the extent to which it is available for global control of action. Control refers to the ability of an agent to control, to modulate, and to inhibit the influence of particular representations on processing. In this framework, control is simply a function of potency, as described in assumption C3. In contrast, P-consciousness refers to the phenomenal aspects of subjective experience discussed in the introduction: a state is P-conscious to the extent that there is something it is like to be in that state: I am currently experiencing a pain, hearing a beautiful piece of music, entertaining the memory of a joyful event. While the extent to which potency (i.e., availability to access consciousness) and

control on the one hand, and subjective experience (i.e., availability to phenomenal consciousness) on the other, are dissociable is debatable, the framework suggests that potency, control, and phenomenal experience are closely related to each other.

Assumption C2: *Availability to consciousness correlates with quality of representation*

This assumption is also a central one in this framework. It states that explicit, conscious knowledge involves higher quality memory traces than implicit knowledge. "Quality of representation" designates several properties of memory traces, such as their relative strength in the relevant information-processing pathways, their distinctiveness, or their stability in time. The assumption is consistent with the theoretical positions expressed by several different authors over the last few years. O'Brien and Opie (1999) have perhaps been the most direct in endorsing a characterization of phenomenal consciousness in terms of the properties of mental representations in defending the idea that "consciousness equals stability of representation", that is, that the particular mental contents that one is aware of at some point in time correspond to those representations that are sufficiently stable in time. Mathis and Mozer (1996) have also suggested that consciousness involves stable representations, specifically by offering a computational model of priming phenomena in which stability literally corresponds to the state that a dynamic "attractor" network reaches when the activations of a subset of its units stops changing and settle into a stable, unchanging state.

A slightly different perspective on the notion of "quality of representation" is offered by authors who emphasize not stability, but strength of representation as the important feature by which to characterize availability to consciousness. One finds echoes of this position in the writings of Kinsbourne (1997), for whom availability to consciousness depends on properties of representations such as duration, activation, or congruence.

In Fig. 1, I have represented the extent to which a given representation is available to the different components of consciousness (phenomenal consciousness, access-consciousness/potency, and control) as functions of a single underlying dimension expressed in terms of the quality of this representation. Availability to access-consciousness is represented by the curve labeled "potency", which expresses the extent to which representations can influence behavior as a function of their quality: high-quality, strong, distinctive representations, by definition, are more potent than weaker representations and hence more likely to influence behavior. "Availability to control processes" is represented by a second curve, so labeled. We simply assume that both weak and very strong representations are difficult to control, and that maximal control can be achieved on representations that are strong enough that they can begin to influence behavior in significant ways, yet not so strong that have become utterly dominant in processing. Finally, availability to phenomenal experience is represented by the third curve, obtained simply by adding the other two. The underlying intuition, discussed in the context of assumption C4, is that which contents enter subjective experience is a function of both availability to control and of potency.

Assumption C3: *Developing high-quality representations takes time*

This assumption states that the emergence of high quality representations in a given processing module takes time, both over training or development, as well as during processing of a single event. Figure 1 can thus be viewed as representing not only the relationships between quality of representation and their availability to the different components of consciousness, but also as a depiction of the dynamics of how a particular representation will change over the different time scales corresponding to development, learning, or within-trial processing (see Destrebecqz and Cleeremans, 2001, 2003; Destrebecqz et al., 2005, for further developments of this specific idea; Cleeremans and Sarrazin, 2007).

Both skill acquisition and development, for instance, involve the long-term progressive emergence of high-quality, strong memory traces based on early availability of weaker traces. Likewise, the extent to which memory traces can influence

performance at any moment (e.g., during a single trial) depends both on available processing time, as well as on overall trace strength. These processes of change operate on the connection weights between units, and can involve either task-dependent, error-correcting procedures, or unsupervised procedures such as hebbian learning. In either case, continued exposure to exemplars of the domain will result in the development of increasingly congruent and strong internal representations that capture more and more of the relevant variance. Although I think of this process as essentially continuous, three stages in the formation of such internal representations (each depicted as separate regions in Fig. 1) can be distinguished: implicit representations, explicit representations, and automatic representations.

The first region, labeled ''Implicit Cognition'' in Fig. 1, is meant to correspond to the point at which processing starts in the context of a single trial, or to some early stage of development or skill acquisition. In either case, this stage is characterized by weak, poor-quality representations. A first important point is that representations at this stage are already capable of influencing performance, as long as they can be brought to bear on processing together with other sources of constraints, that is, essentially through mechanisms of associative priming and constraint satisfaction. A second important point is that this influence is best described as ''implicit'', because the relevant representations are too weak (i.e., not distinctive enough) for the system as a whole to be capable of exerting control over them: you cannot control what you cannot identify as distinct from something else.

The second region of Fig. 1 corresponds to the emergence of explicit representations, defined as representations over which one can exert control. In the terminology of attractor networks, this would correspond to a stage during learning at which attractors become better defined — deeper, wider, and more distinctive, so corresponding to the best ''constraint-satisfaction'' interpretation of a state of affairs (Maia and Cleeremans, 2005). It is also at this point that the relevant representations acquire generative capacity, in the sense that they now have accrued sufficient strength to have the potential to determine appropriate responses when their preferred stimulus is presented to the system alone. Such representations are also good candidates for redescription and can thus be recoded in various different ways, for instance, as linguistic propositions.

The third region involves what I call automatic representations, that is, representations that have become so strong that their influence on behavior can no longer be controlled (i.e., inhibited). Such representations exert a mandatory influence on processing. Importantly, however, one is aware both of possessing them (i.e., one has relevant metaknowledge) and of their influence on processing (see also Tzelgov, 1997), because availability to conscious awareness depends on the quality of internal representations, and that strong representations are of high quality. In this perspective then, one can always be conscious of automatic behavior, but not necessarily with the possibility of control over these behaviors.

In this framework, skill acquisition and development therefore involve a continuum at both ends of which control over representations is impossible or difficult, but for very different reasons: implicit representations influence performance but cannot be controlled because they are not yet sufficiently distinctive and strong for the system to even know it possesses them. This might in turn be related to the fact that, precisely because of their weakness, implicit representations cannot influence behavior on their own, but only in conjunction with other sources of constraints. Automatic representations, on the other hand, cannot be controlled because they are too strong, but the system is aware both of their presence and of their influence on performance.

Assumption C4: *The function of consciousness is to offer flexible, adaptive control over behavior*

The framework gives consciousness a central place in information processing, in the sense that its function is to enable flexible control over behavior. Crucially, however, consciousness is not necessary for information processing, or for adaptation in general, thus giving a place for implicit learning in cognition. I believe this

perspective to be congruent with theories of adaptation and optimality in general.

Indeed, another way to think about the role of learning in consciousness is to ask: "When does one need control over behavior?" Control is perhaps not necessary for implicit representations, for their influence on behavior is necessarily weak (in virtue of the fact that precisely because they are weak, such representations are unlikely to be detrimental to the organism even if they are not particularly well-adapted). Likewise, control is not necessary for automatic representations, because presumably, those representations that have become automatic after extensive training should be adapted (optimal) as long as the processes of learning that have produced them can themselves be assumed to be adaptive. Automatic behavior is thus necessarily optimal behavior in this framework, except, precisely, in cases such as addiction, obsessive–compulsive behavior, or laboratory situations where the automatic response is manipulated to be non-optimal, such as in the Stroop situation. Referring again to Fig. 1, my analysis therefore suggests that the representations that require the most control are the explicit representations that correspond to the central region of Fig. 1: representations that are strong enough that they have the potential to influence behavior in and of themselves (and hence that one should really care about, in contrast to implicit representations), but not sufficiently strong that they can be assumed to be already adapted, as is the case for automatic representations. It is for those representations that control is needed, and, for this reason, it is these representations that one is most aware of.

Likewise, this analysis also predicts that the dominant contents of subjective experience at any point in time consist precisely of those representations that are both strong enough that they can influence behavior, yet weak enough that they still require control. Figure 1 reflects these ideas by suggesting that the contents of phenomenal experience depend both on the potency of currently active representations as well as on their availability to control. Since availability to control is inversely related to potency for representations associated with automatic behavior, this indeed predicts weaker availability to phenomenal experience of "very strong" representations as compared to "merely strong" representations. In other words, such representations can become conscious if appropriate attention is directed towards their contents — as in cases where normally automatic behavior (such as walking) suddenly becomes conscious because the normal unfolding of the behavior has been interrupted (e.g., because I've stumbled upon something) — but they are not normally part of the central focus of awareness nor do they require cognitive control. It is interesting to note that these ideas are roughly consistent with Jackendoff's (1987) and Prinz's (2007) "Intermediate Level Theory of Consciousness".

The framework thus leaves open four distinct possibilities for knowledge to be implicit. 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 I assume that consciousness 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 be detectable — but only indirectly, and only to the extent that these effects are sufficiently marked in the corresponding representations.

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.

Third, a representation can be strong enough to enter conscious awareness, but fail to be associated with relevant metarepresentations. 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. Dienes and Perner (2003) offer an insightful analysis of

the different ways in which what I have called high-quality representations can remain implicit. Likewise, phenomena such as inattentional blindness (Mack and Rock, 1998) or blindsight (Weiskrantz, 1986) also suggest that high-quality representations can nevertheless fail to reach consciousness, not because of their inherent properties, but because they fail to be attended to or because of functional disconnection with other modules (see Dehaene et al., 2006).

Finally, a representation can be so strong that its influence can be no longer be controlled. In such 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 it, 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.

Metarepresentation

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.

Once a representation has accrued sufficient strength, stability, and distinctiveness, it may be the target of metarepresentations: 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 metarepresentations are learned in just the same way as first-order representations, that is, by virtue of continuously operating learning mechanisms. Because metarepresentations are also representations, the same principles of stability, strength, and distinctiveness therefore apply. An important implication of this observation is that activation of metarepresentations can become automatic, just as it is the case for first-order representations.

What might be the function of such metarepresentations? One intriguing 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.

However, there is another important function that metarepresentations may play: they can also be used to anticipate the future occurrences of first-order representations. 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 metarepresentations 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. This perspective is akin to the sensorimotor perspective (O'Regan and Noë, 2001) in the sense that awareness is linked with knowledge of the consequences of our actions, but, crucially, the argument is extended to the entire domain of neural representations.

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, but about its own representations of it. The result of this unconscious learning is conscious experience, in virtue of the fact that each representational state is now accompanied by (unconscious learnt) metarepresentations that convey the mental attitude with which these 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; Dienes, in press), but also with those of Lau (in press) who characterizes consciousness as "signal detection on the mind". Finally, one dimension that I feel is sorely missing from contemporary discussion of consciousness is emotion (see Damasio, 1999; LeDoux, 2002; Tsuchiya and Adolphs, 2007). Conscious experience would not exist without experiencers who *care* about their experiences!

Acknowledgments

A.C. 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 A.C. 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). Substantial portions of this article were adapted from the following publication: Cleeremans (2006). I would like to thank the organizers of the *Models of Brain and Mind: Physical, Computational and Psychological Approaches* workshop, and Rahul Banerjee in particular, for inviting me to contribute this piece.

References

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

Banerjee, R. (in press). Buddha and the bridging relations. Prog. Brain Res.

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

Block, N. (2005) Two neural correlates of consciousness. Trends Cogn. Sci., 9(2): 46–52.

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

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

Chalmers, D.J. (2007a) Naturalistic dualism. In: Velmans M. and Schneider S. (Eds.), The Blackwell Companion to Consciousness. Blackwell Publishing, Oxford, UK, pp. 359–368.

Chalmers, D.J. (2007b) The hard problem of consciousness. In: Velmans M. and Schneider S. (Eds.), The Blackwell Companion to Consciousness. Blackwell Publishing, Oxford, UK, pp. 225–235.

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. (2005) Computational correlates of consciousness. In: Laureys S. (Ed.), Progress in Brain Research, Vol. 150. Elsevier, Amsterdam, pp. 81–98.

Cleeremans, A. (2006) Conscious and unconscious cognition: a graded, dynamic perspective. In: Jing Q., Rosenzweig M.R., d'Ydewalle G., Zhang H., Chen H.-C. and Zhang C. (Eds.), Progress in Psychological Science around the World. Vol. 1:

Neural, Cognitive, and Developmental IssuesPsychology Press, Hove, UK, pp. 401–418.
Cleeremans, A. (in preparation) Being Virtual. Oxford University Press, Oxford, UK.
Cleeremans, A. and Jiménez, L. (2002) Implicit learning and consciousness: a graded, dynamic perspective. In: French R.M. and Cleeremans A. (Eds.), Implicit Learning and Consciousness: An Empirical, Computational and Philosophical Consensus in the Making? Psychology Press, Hove, UK, pp. 1–40.
Cleeremans, A. and Sarrazin, J.-C. (2007) Time, action, and consciousness. Hum. Mov. Sci., 26(2): 180–202.
Damasio, A. (1999) The Feeling of What Happens: Body and Emotion in the Making of Consciousness. Harcourt Brace & Company, New York, NY.
Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J. and Sergent, C. (2006) Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends Cogn. Sci., 10(5): 204–211.
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(24): 14529–14534.
Dennett, D.C. (1991) Consciousness Explained. Little, Brown & Co., Boston, MA.
Dennett, D.C. (2001) Are we explaining consciousness yet? Cognition, 79: 221–237.
Destrebecqz, A. and Cleeremans, A. (2001) Can sequence learning be implicit? New evidence with the Process Dissociation Procedure. Psychon. Bull. Rev., 8(2): 343–350.
Destrebecqz, A. and Cleeremans, A. (2003) Temporal effects in sequence learning. In: Jiménez L. (Ed.), Attention and Implicit Learning. John Benjamins, Amsterdam, pp. 181–213.
Destrebecqz, A., Peigneux, P., Laureys, S., Degueldre, C., Del Fiore, G. Aerts, J. et al. (2005) The neural correlates of implicit and explicit sequence learning: interacting networks revealed by the process dissociation procedure. Learn. Mem., 12(5): 480–490.
Dienes, Z. (in press). Subjective measures of unconscious knowledge. Prog. Brain Res.
Dienes, Z. and Perner, J. (1999) A theory of implicit and explicit knowledge. Behav. Brain Sci., 22: 735–808.
Dienes, Z. and Perner, J. (2003) Unifying consciousness with explicit knowledge. In: Cleeremans A. (Ed.), The Unity of Consciousness: Binding, Integration, and Dissociation. Oxford University Press, Oxford, UK, pp. 214–232.
Farah, M.J. (1994) Visual perception and visual awareness after brain damage: a tutorial overview. In: Umiltà C. and Moscovitch M. (Eds.), Attention and Performance XV: Conscious and Nonconscious Information Processing. MIT Press, Cambridge, MA, pp. 37–76.
Frith, C.D., Perry, R. and Lumer, E. (1999) The neural correlates of conscious experience: an experimental framework. Trends Cogn. Sci., 3: 105–114.
Hinton, G.E. (1986) Learning distributed representations of concepts. In: Proceedings of the Eighth Annual Conference of the Cognitive Science Society (Amherst, MA), Hillsdale, Erlbaum, pp. 1–12.
Humphrey, N. (1971) Colour and brightness preferences in monkeys. Nature, 229: 615–617.
Humphrey, N. (2006) Seeing Red. Harvard University Press, Cambridge, MA.
Jackendoff, R. (1987) Consciousness and the Computational Mind. MIT Press, Cambridge, MA.
Karmiloff-Smith, A. (1992) Beyond Modularity: A Developmental Perspective on Cognitive Science. MIT Press, Cambridge.
Kinsbourne, M. (1997) What qualifies a representation for a role in consciousness? In: Cohen J.D. and Schooler J.W. (Eds.), Scientific Approaches to Consciousness. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 335–355.
Kirsh, D. (1991) When is information explicitly represented? In: Hanson P.P. (Ed.), Information, Language, and Cognition. Oxford University Press, New York, NY.
Kirsh, D. (2003) Implicit and explicit representation. In: Nadel L. (Ed.), Encyclopedia of Cognitive Science, Vol. 2. Macmillan, London, UK, pp. 478–481.
Koch, C. (2004) The Quest for Consciousness. A Neurobiological Approach. Roberts & Company Publishers, Englewood, CO.
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., 7(1): 12–18.
Lau, H. (in press). A higher-order Bayesian Decision Theory of consciousness. Prog. Brain Res.
LeDoux, J. (2002) Synaptic Self. Viking Penguin, Harmondsworth, UK.
Mack, A. and Rock, I. (1998) Inattentional Blindness. MIT Press, Cambridge, MA.
Maia, T.V. and Cleeremans, A. (2005) Consciousness: converging insights from connectionist modeling and neuroscience. Trends Cogn. Sci., 9(8): 397–404.
Mathis, W.D. and Mozer, M.C. (1996) Conscious and unconscious perception: a computational theory. In: Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 324–328.
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.
Persaud, N., McLeod, P. and Cowey, A. (2007) Post-decision wagering objectively measures awareness. Nat. Neurosci., 10: 257–261.
Prinz, J.J. (2007) The intermediate level theory of consciousness. In: Velmans M. and Schneider S. (Eds.), The Blackwell Companion to Consciousness. Oxford University Press, Oxford, UK, pp. 248–260.

O'Regan, J.K. and Noë, A. (2001) What it is like to see: a sensorimotor theory of visual experience. Synthèse, 129(1): 79–103.

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

Rosenthal, D. (1997) A theory of consciousness. In: Block N., Flanagan O. and Güzeldere G. (Eds.), The Nature of Consciousness: Philosophical Debates. MIT Press, Cambridge, MA.

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

Tononi, G. (2003) Consciousness differentiated and integrated. In: Cleeremans A. (Ed.), The Unity of Consciousness: Binding, Integration, and Dissociation. Oxford University Press, Oxford, UK, pp. 253–265.

Tononi, G. (2007) The information integration theory. In: Velmans M. and Schneider S. (Eds.), The Blackwell Companion to Consciousness. Blackwell Publishing, Oxford, UK, pp. 287–299.

Tsuchiya, N. and Adolphs, R. (2007) Emotion and consciousness. Trends Cogn. Sci., 11(4): 158–167.

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

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