Progress in Brain Research, Vol. 150
ISSN 0079-6123

CHAPTER 7

Computational correlates of consciousness

Axel Cleeremans*

Cognitive Science Research Unit, Université Libre de Bruxelles CP 122, Avenue F.-D. Roosevelt, 50 1050 Brussels, Belgium

Abstract: Over the past few years numerous proposals have appeared that attempt to characterize consciousness in terms of what could be called its computational correlates: Principles of information processing with which to characterize the differences between conscious and unconscious processing. Proposed computational correlates include architectural specialization (such as the involvement of specific regions of the brain in conscious processing), properties of representations (such as their stability in time or their strength), and properties of specific processes (such as resonance, synchrony, interactivity, or information integration). In exactly the same way as one can engage in a search for the neural correlates of consciousness, one can thus search for the computational correlates of consciousness. The most direct way of doing is to contrast models of conscious versus unconscious information processing. In this paper, I review these developments and illustrate how computational modeling of specific cognitive processes can be useful in exploring and in formulating putative computational principles through which to capture the differences between conscious and unconscious cognition. What can be gained from such approaches to the problem of consciousness is an understanding of the function it plays in information processing and of the mechanisms that subtend it. Here, I suggest that the central function of consciousness is to make it possible for cognitive agents to exert flexible, adaptive control over behavior. From this perspective, consciousness is best characterized as involving (1) a graded continuum defined over quality of representation, such that as availability to consciousness and to cognitive control correlates with properties of representation, and (2) the implication of systems of meta-representations.

Introduction

In a surprisingly lucid passage, Sigmund Freud (1949), reflecting on the prospects of developing a scientific approach to psychological phenomena, wrote the following:

> We know two kinds of things about what we call our psyche (or mental life): firstly, its bodily organ and scene of action, the brain (or nervous system) and, on the other hand, our acts of consciousness, which are immediate data and cannot be further explained by any sort of description. Everything that lies in between is unknown to us, and the data do not include any direct relation between these two terminal points of our knowledge. If it existed, it would at the most afford an exact localization of the processes of consciousness and would give us no help towards understanding them.

Freud's insightful but rather pessimistic thoughts about the possibility of developing a "Science of Consciousness" thus illustrates the most fundamental problem that cognitive neuroscience must confront in this context: That of establishing caus-

---

*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(05)50007-4

al relationships between fundamentally private, subjective states (what Freud calls "our acts of consciousness") on the one hand, and objective, observable states (e.g., behavioral and neural states) on the other hand.

This program of establishing direct correspondences between subjective and objective states now finds a contemporary echo in the unfolding search for the "Neural Correlates of Consciousness (NCC)." The expression "Neural Correlates of Consciousness" was first used by Crick and Koch (1990) and has since attracted, as an empirical program, the attention of a large community of researchers — from scientists to philosophers alike (see Metzinger, 2000, for an extensive collection of relevant contributions).

According to Chalmers (2000, p. 31), a "neural correlate of consciousness" is "a minimal neural system N such that there is a mapping from states of N to states of consciousness, where a given state of N is sufficient, under conditions C, for the corresponding state of consciousness".

Candidate's NCC, to mention just a few of those listed in Chalmers (2000), include, for instance, 40-Hz oscillations in the cerebral cortex (Crick and Koch, 1990; also Ribary, this volume; John, this volume), reentrant loops in thalamocortical systems (Edelman, 1989; also see Tononi, this volume), neural assemblies bound by NMDA (Flohr, 1985; also see Greenfield, this volume), or extended reticular-thalamic activation systems (Newman and Baars, 1993, also see Baars, this volume).

Chalmers (2000) is quick to point out several potential shortcomings of this definition, such as the facts that there might not be a single NCC, NCCs might not consist of circumscribed regions of the brain, or it might be the case that some aspects of consciousness simply fail to correlate in some sense with brain activity (a view to which few would subscribe). Noë and Thompson (2004) likewise critique — but in a somewhat different direction — what they call the "matching-content doctrine," that is, the idea that the representation of a particular content in a neural system is sufficient for representation of that same content in consciousness. Specifically, Noë and Thompson aim to suggest that the search for the NCC might be misguided to the extent that it eschews the fact that conscious states cannot be analyzed independent of the environment with which the agent interacts constantly (also see O'Regan et al., this volume).

In a rather pessimistic article, Haynes and I raised similar points about the possibility of developing a "science of consciousness" (Cleeremans and Haynes, 1999). How are we to proceed, we asked, given not only that one has no clear idea of what it is exactly that one is measuring when using methods such as fMRI, but also, and perhaps more importantly, that we lack the conceptual tools that would be necessary to develop a scientific approach to phenomenology? I do not have direct access to your mental states, and, some would argue, neither do I have perfect access to my own mental states (or if I do, I am likely to be mistaken in different ways, see Nisbett and Wilson, 1977; Dennett, 1991; Wegner, 2002).

This assessment will strike many as overly grim, and yet, the challenges are both substantial and numerous. In this respect, it is worth pointing out that renewed interest in consciousness has triggered rather unrealistic expectations in the community. Somehow, many continue to expect that there will be a single "aha" moment when an obscure neuroscientist suddenly comes up with "the" mechanism of consciousness. Needless to say, this is not going to happen: functional accounts of consciousness that take it as a starting point that it is a single, static property associated with some mental states and not with others are doomed to fail, for consciousness is neither "a single thing" nor is it static. Instead, consciousness refers to several, possibly dissociable, aspects of information processing, and it is a fundamentally dynamic, graded, process.

Despite these caveats, many have now rightfully opted for a pragmatic approach focused on the following simple assumption, namely that "for any mental state (state of consciousness) there is an associated neural state; it is impossible for there to be a change of mental state without a corresponding change in neural state" (Frith et al., 1999, p. 105).

On the basis of this rather non-controversial assumption (for materialists, at least), Frith et al. (1999, p. 107) continue by offering a straightfor-

Table 1. Characterization of different experimental paradigms Frith (1999) through which to study differences between conscious and unconscious cognition in normal (clear cells) and abnormal (shaded cells) cases (see text for details)

	Perception	Memory	Action
Subjective experience change, stimulation and/or behavior remains constant	Binocular rivalry	Episodic recall	Awareness of intention
	Hallucinations	Confabulation	Delusion of control
Stimulation changes, subjective experience remains constant	Stimulation changes without awareness	Unrecognized “old” items	Stimuli eliciting action without awareness
	Blindsight	Unrecognized items in amnesia	Stimuli eliciting unintended action
Behavior changes, subjective experience remains constant	Correct guessing without awareness	Implicit learning	Implicit motor behavior
	Correct reaching in form-agnosia	Implicit learning in amnesia	Unintended action

ward canvas with which to guide the search for the neural correlates of consciousness:

> A major part of the program for studying the neural correlates of consciousness must be to investigate the difference between neural activities that are associated with awareness and those that are not.

This contrastive approach to consciousness (see Baars, 1988, 1994) now constitutes the core of many current efforts to understand the neural bases of consciousness. Frith et al., in their superb review, usefully propose an analysis of the different paradigms through which one can pursue this contrastive approach. Table 1 summarizes the different possibilities delineated by Frith and colleagues, who suggested to organize paradigms to study the “neural correlates of consciousness” in nine groups resulting from crossing two dimensions: (1) three classes of psychological processes involving knowledge of the past, present, and future — memory, perception, and action — and (2) three types of cases where subjective experience is incongruent with the objective situation — cases where subjective experience fails to reflect changes in either (a) the stimulation or (b) behavior, and (c) cases where subjective experience changes, whereas stimulation and behavior remain constant. This approach can be further applied to either normal or pathological cases.

The paradigmatic example of a situation where one seeks to identify the neural correlates of perception is binocular rivalry (see e.g., Lumer et al., 1998; Logothethis and Schall, 1989; Naccache, this volume), in which an unchanging compound stimulus consisting of two elements presented separately and simultaneously to each eye produces spontaneously alternating complete perceptions of each element. By asking participants (or certain animals) to indicate which stimulus they perceive at any moment, one can then strive to establish which regions of the brain exhibits activity that correlates with subjective experience and which do not, in a situation where the actual stimulus remains unchanged. Research on the neural correlates of implicit learning, in contrast, instantiates the reverse situation, where people’s subjective experience fails to reflect the fact that they are becoming increasingly sensitive to novel information they are learning about over the course of practicing a task such as sequence learning (Cleeremans et al., 1998). Here again, by contrasting cases where learning is accompanied by conscious awareness with cases where it is not, one can strive to explore which regions of the brains subtend implicit and explicit learning, and to what degree (Destrebecqz et al., 2003; Destrebecqz and Peigneux, this volume). Literally, dozens of other studies have now followed the same logic in varied domains, as illustrated in Table 1.

However, there are reasons to claim that the search for the NCC should now be (and indeed, is) augmented by similar efforts aimed at unraveling what one could call, on the one hand, the *behavioral correlates of consciousness* (BCC), and, on the other hand, the *computational correlates of consciousness* (CCC). One could thus paraphrase Frith et al.'s quote in the following manner:

> A major part of the program for studying the behavioral correlates of consciousness must be to investigate the difference between behaviors that are associated with awareness and those that are not.

and:

> A major part of the program for studying the computational correlates of consciousness must be to investigate the difference between computations that are associated with awareness and those that are not.

While what I have called the ''search for the behavioral correlates of consciousness'' is nothing new, the search for the computational correlates of consciousness is barely beginning. There is, however, a small community of scientists specifically interested in pursuing the goal of building ''conscious machines'' (Holland, 2003; Aleksander, this volume) through the development of implemented computational models aimed either at fleshing out broad theories of consciousness (Cotterill, 1998; Dehaene et al., 1998; Franklin and Graesser, 1999; Taylor, 1999; Aleksander, 2000; Sun, 2001; Perruchet and Vinter, 2003) or at providing detailed accounts of the difference between conscious and unconscious cognition (Farah et al., 1994; Mathis and Mozer, 1996; Dehaene et al., 2003; Fragopanagos and Taylor, 2003; Colagrosso and Mozer, in press). Also relevant is the growing computationally oriented literature dedicated to the phenomena of implicit learning (Cleeremans et al., 1998).

A joint search for the NCC, BCC, and CCC sets up a clear multidisciplinary program for the scientific study of consciousness — one that involves systematically manipulating variables that will result in producing differences between conscious and unconscious neural states, behaviors, or computations. The latter contrast is in my view particularly important, for it may result in the identification of *computational principles* that differentiate between cognition with and without consciousness. This is the issue that I will focus on in the rest of this chapter. To do so, I will first briefly overview different existing, broad proposals with the goal of establishing how they differ from each other and on which information-processing principles they rely to account for differences between conscious and unconscious cognition. Next, I will suggest that, from a computational point of view, consciousness can be analyzed as involving two central aspects.

The first is what one could call ''quality of representation'' (see also Farah, 1994) — properties associated with representations in the brain or in artificial systems, such as their strength, their stability in time, or their distinctiveness. Quality of representation, by this account, determines, in a graded manner, the extent to which a particular representation becomes available to conscious experience and to cognitive control, and is viewed as a necessary condition for a particular representation to become available to consciousness. The second is the extent to which a given representation is accompanied by further (re-)representation of itself — in other words, whether the system is capable of meta-representation.

Finally, I will close with a brief discussion of a novel class of computational models, — the so-called ''forward models,'' — and their potential in capturing many insights into the computational correlates of consciousness within a single broad computational framework. Before undertaking this analysis, however, it seems important to reflect upon the functions of consciousness. Indeed, as Taylor (1999) points out, ''... without a function for consciousness, we have no clue as to a mechanism for it. Scientific modeling cannot even begin in this case; it has nothing to get its teeth into'' (p. 49).

The functions of consciousness

Analyzing consciousness in terms of its underlying mechanisms first requires us to identify the functions that it may play within a cognitive system. There are several different manners in which this question can be approached depending on which aspect of consciousness one focuses on. The fact that consciousness is not a unitary concept (Zeman, this volume) is important, particularly because many recent experiments tend to treat it as though it were a ''single thing'', whereas it is neither a thing nor a unitary concept.[1] Block's (1995) well-known analysis is useful here as a starting point. Block distinguishes between access consciousness, phenomenal consciousness, monitoring consciousness, and self-consciousness.

Access consciousness (A-consciousness) refers to our ability to report and act on our experiences. For a person to be in an A-conscious state entails that there is a representation in that person's brain whose content is available for verbal report and for high-level processes such as conscious judgment, reasoning, and the planning and guiding of action. There is wide agreement around the idea that conscious representations differ from unconscious ones in terms of such global accessibility: Conscious representations are informationally available to multiple systems in a manner that unconscious representations are not. Accessibility is in turn viewed as serving the function of making it possible for an agent to exert flexible, adaptive control over action. Tononi (Tononi and Edelman, 1998, 2003, this volume) proposes that the main function of consciousness is to rapidly integrate a lot of information — a function that would clearly endow agents who possess this ability with an evolutionary advantage over others who lack it. In a recent overview article, Dehaene and Naccache (2001) state that ''The present view associates consciousness with a unified neural workspace through which many processes can communicate. The evolutionary advantages that this system confers to the organism may be related to the increased independence that it affords.'' (p. 31). Dehaene and Naccache thus suggest that consciousness allows organisms to free themselves from acting out their intentions in the real world, relying instead on less hazardous simulation made possible by the neural workspace. Most existing computational models of consciousness are explicitly targeted toward capturing the computational consequences of A-consciousness rather than the phenomenal qualities associated with conscious states — Block's second concept of phenomenal consciousness.

Phenomenal consciousness (P-consciousness) refers to the qualitative nature of subjective experience: What it is like to smell a particular scent, to feel a particular pain, to remember the emotions associated with a particular event, to be a bat chasing insects at nightfall. There is no agreement concerning the putative functions of P-consciousness. Some authors argue that there is nothing to be explained, that qualia are illusory, or that they are purely epiphenomenal and hence play no causal role in information processing. For instance, O'Regan and Noë (2001) hold that qualia reflect nothing more than mastery of learned sensory–motor contingencies: What it means to consciously experience something is simply to know about the consequences of one's actions (O'Regan et al., this volume). For Dennett (1991, 2001), conscious contents merely reflect the dominance of some representations over others at some point in time — ''fame in the brain'', as he calls it. Others have proposed that conscious experience might serve error-correcting functions. For instance, Gray's ''comparator hypothesis'' (2004) states that the function of P-consciousness is to make it possible for the agent to rehearse and deliberate upon the conditions under which something unexpected happened (such as the consequences of an error). Koch proposes that the function of P-consciousness is to provide an ''executive summary'' to those parts of the brain involved in planning and deliberation (Crick and Koch, 1995; Koch, 2004). This executive summary is assumed to be the result of constraint satisfaction processes, and reflects the best interpretation of the current situation. Another interesting hypothesis concerning the function of conscious experience was put forward by Gregory (2003), according to whom P-consciousness might serve the function of ''flagging

---

[1] Contrast, for instance, cases where one asks whether a subject is conscious of a single stimulus presented to her to cases where one asks what is it is like to walk in the Alps or to sample an excellent wine. Our concept of consciousness is radically different in each case.

the present'', so making it possible for the agent to distinguish between actual, remembered, and anticipated states. More generally, perhaps the function of conscious experience is to associate emotional valence to the consequences of one's actions. If nothing ever is done to an agent, there seems to be little basis for learning and adapting behavior in general. On the other hand, one might also argue that it is simply misguiding to look for putative functional accounts of phenomenal consciousness since, by definition, it is what is ''left over'' once all functional aspects of consciousness have been accounted for.

Monitoring consciousness refers to thoughts about or awareness of one's sensations and percepts, as distinct from those sensations and percepts themselves. Functionally, some form of monitoring consciousness appears to be necessary to support adapted control over behavior, through appraisal of one's internal states and metacognition in general.

Finally, *self-consciousness* refers to thoughts about or awareness of oneself. Studying the self is a huge undertaking in and of itself, and the domain is currently witnessing fascinating developments (see e.g., Knoblich et al., 2003 for a review). It would be too long to develop this aspect of consciousness in this chapter, but a basic fact about conscious experience is simply that it would not make any sense unless there was a self-aware agent experiencing the experience. Hence, consciousness of self is clearly a very important component of what it means to be conscious (Damasio, 1999).

Having delineated a few possible functions for consciousness in its different aspects, we can now ask the following questions: What sorts of mechanisms have been proposed to fulfill these functions? What are the computational correlates of consciousness? These will be the object of the next section.

The search for the CCC

Computational models of the differences between conscious and unconscious information processing are few and far between. This is not surprising, for the challenge of exploring the mechanisms of something as complex and ill-defined as consciousness is enormous. This is also the main reason why most existing computational models of consciousness have been directed at accounting for A-consciousness as opposed to P-consciousness: The former at least receives some sort of functionalist interpretation, while the functions of the latter, if any, clearly remain controversial at this point. Monitoring- and self-consciousness, on the other hand, require accounts that necessarily involve a great deal of complexity before they can even get off the ground, and are hence challenging to explore from a computational point of view.

This being said, existing models generally fall into two classes: Overarching models — often only partially implemented — that aim to offer a general blueprint for information processing with or without consciousness on the one hand, and very specific models of particular empirical situations on the other. Each suffers from its own set of limitations (which they share with computational models in general). Overarching models are often difficult to compare with existing data because they often fail to make testable predictions. Specific models, on the other hand, can always be dismissed as convincing accounts of the mechanisms of consciousness precisely because of their limited scope. In either case, one could question the extent to which such modeling efforts are worth it, though this would clearly invalidate any scientific approach to the problem. For instance, if you assume that consciousness crucially includes properties that can never be amenable to functionalist and cognitive analyses — Chalmers' (1996) ''hard problem'' — then clearly such models are doomed to fail, and so would the possibility of understanding conscious experience from a third-person perspective. Some authors have also pointed out that while it might be possible to build conscious machines, we would never be able to decide whether such machines actually have experiences of any kind (Prinz, 2003).

Nevertheless, both types of models can play a substantial role in helping us converge onto a set of computational principles to characterize the differences between conscious and unconscious cognition. Identifying such principles is an important endeavor, for it would clearly make it possible to go beyond establishing mere relationships between conscious states and their neural or behavioral correlates. In other words, if we are able to define such principles, we would be in a position to address the mechanisms through which consciousness is achieved in cognitive systems.

Current theories of consciousness sometimes make very different assumptions about its underlying mechanisms. Farah (1994) distinguishes between three types of neuroscientific/computational accounts of consciousness: “privileged role” accounts, “integration” accounts, and “quality of representation” accounts. “Privileged role” accounts take their roots in Descartes’ thinking and assume that consciousness depends on the activity of specific brain systems whose function it is to produce subjective experience. “Integration” accounts, in contrast, assume that consciousness only depends on processes of integration through which the activity of different brain regions can be synchronized or made coherent. Finally, “quality of representation” accounts assume that consciousness depends not on particular processes, but on particular properties of neural representations, such as their strength or their stability in time.

In a recent overview article (see also O’Brien and Opie, 1999; Atkinson et al., 2000), my co-authors and I proposed to organize computational theories of consciousness along two dimensions, as depicted in Fig. 1[2]: A process versus vehicle dimension, which opposes models that characterize consciousness in terms of specific processes operating over mental representations to models that characterize consciousness in terms of intrinsic properties of mental representations, and a specialized versus non-specialized dimension, which contrasts models that posit information-processing systems dedicated to consciousness with models for which consciousness can be associated with any information-processing system as long as this system has the relevant properties.

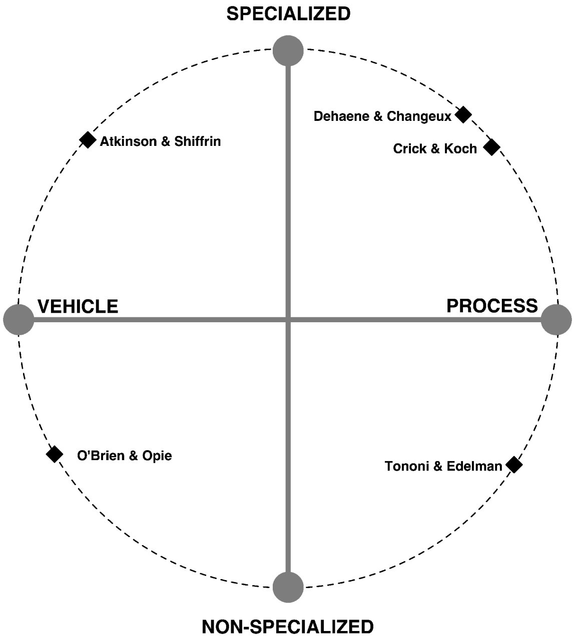

Fig. 1. A conceptual 2-D map in which to locate theories of consciousness. (Adapted from Atkinson et al., 2000.) The map is defined by two dimensions relevant to computational theories of consciousness: Whether the theory assumes the involvement of specialized structures or not (Y-axis), and whether the theory assumes that consciousness depends on properties associated with representational vehicles or with processes (X-axis).

Farah’s three categories can be subsumed in this analysis in the following manner: “privileged role” models, which assume that some brain systems play a specific role in subtending consciousness, are specialized models that can be instantiated either through “vehicle” or through “process” principles. “Quality of representation”, models, on the other hand, are typical vehicle theories in that they emphasize that what makes some representations available to conscious experience are properties of those representations rather than their functional role. Finally, Farah’s “integration” models are examples of non-specialized theories, which can again be either instantiated in terms of the properties of the representations involved or in terms of the processes that engage these representations. Atkinson et al.’s analysis thus offers four broad

---

[2] Figure 1 is aimed at providing a few illustrative examples and is by no means intended to be exhaustive. Your favorite theory (or your own theory!) may thus not be on the map, which I urge you not to interpret as a suggestion that it is not important.

categories of computational accounts of consciousness.

1. *Specialized vehicle theories* assume that consciousness depends on the properties of the representations that are located within a specialized system in the brain. An example of such accounts is Atkinson and Shiffrin's (1971) model of short-term memory, which specifically assumes that representations contained in the short-term memory store (a specialized system) only become conscious if they are sufficiently strong (a property of representations).
2. *Specialized process theories* assume that consciousness arises from specific computations that occur in a dedicated mechanism, as in Schacter's (1989) Conscious Awareness System (CAS) model. Schacter's model assumes that the CAS's main function is to integrate inputs from various domain specific modules, and to make this information available to executive systems. It is therefore a specialized model in that it assumes that there exist specific regions in the brain whose function is to make its contents available to conscious awareness. It is a process model to the extent that any representation that enters the CAS will become available to conscious awareness in virtue of the processes that manipulate these representations, and not in virtue of properties of those representations themselves. More recent computational models of consciousness also fall into this category, most notably Dehaene and colleagues' (1998) neural workspace model and Crick and Koch's (2003) framework, both of which assume, albeit somewhat differently, that the emergence of consciousness depends on the occurrence of specific processes in specialized systems.
3. *Non-specialized vehicle theories* include any model that posits that availability to consciousness only depends on properties of representations, regardless of where in the brain these representations exist or of which processes engage these representations. O'Brien and Opie's (1999) "connectionist theory of phenomenal experience" is the prototypical example of this category, to the extent that it specifically assumes that any stable neural representation will both be causally efficacious and form part of the contents of phenomenal experience. Mathis and Mozer (1995) likewise propose to associate consciousness with stable states in neural networks, though Mozer's more recent PIT framework (Colagrosso and Mozer, in press) also puts emphasis on the existence of functional connectivity between different modules as critical for A-consciousness Zeki's notion of "micro-consciousness" is also an example of this type of perspective (Zeki and Bartels, 1998).
4. *Non-specialized process theories* finally, are theories which assume that representations become conscious whenever they are engaged by certain specific processes, regardless of where these representations exist in the brain. Many recent proposals fall into this category. Examples include Tononi and Edelman's (1998) "dynamic core" model; Crick and Koch's (1995) idea that synchronous firing constitutes the primary mechanisms through which disparate representations become integrated as part of a unified conscious experience or Grossberg's (1999) characterization of consciousness as involving processes of "adaptive resonance" through which representations that simultaneously receive bottom-up and top-down activation become conscious because of their stability and strength.

There are two important caveats to this analysis. Firstly, the taxonomy is defined by how specific computational theories of consciousness characterize the difference between conscious and unconscious cognition rather than by a sharp distinction between vehicles versus processes on the one hand, and specialized versus non-specialized systems on the other. Thus, it should be clear that representation and process cannot be considered independently from each other, to the extent that the effects of particular processes will necessarily result in changes in the nature of the representations involved. For instance, processes like resonance, amplification, or reentrant processing (Lamme, 2004), all of which basically involve constraint satisfaction processes as they occur in interactive networks, will all result in stabilizing and in strengthening specific patterns of activity in the corresponding neural pathways. The distinction between specialized and non-specialized models similarly fails to be as sharp as depicted above, for there are multiple ways in which a system can be described as specialized. For instance, a system can be specialized to the extent that it involves a single "box" or cerebral region whose function it would be to make whatever contents are represented in that system conscious (no current neuroscientific theory of consciousness adopts this assumption this bluntly). On the other hand, a system can be specialized to the extent that it involves specific connectivity between different cerebral regions. Dehaene and Changeux's (in press) notion that the neural workspace relies on specific long-distance cortico-cortical connections is an example of the latter case of specialization, and so contrasts with other proposals that put less emphasis on the involvement of dedicated systems (Tononi and Edelman, 1998).

Secondly, several proposals also tend to be somewhat more hybrid, instantiating features and ideas from several of the categories described by Atkinson et al. Baars' influential "global workspace" model (Baars, 1988, this volume), for instance, incorporates features from specialized process models as well as from non-specialized vehicles theories, to the extent that the model assumes that consciousness involves a specialized system (the global workspace), but also characterizes conscious states in terms of the properties associated with their representations (i.e., global influence and widespread availability) rather than in terms of the processes that operate on these representations. Likewise, Dehaene et al. (1998) assume that consciousness depends on (1) *active firing*, which can be construed as a property of representation, (2) *long-distance connectivity* (a specialized system), and (3) dynamic mobilization, a selective process depending on simultaneous bottom-up and top-down activation of the representations contained in the linked modules. Thus, this model acknowledges both the existence of specific, dedicated mechanisms to support consciousness as well as specific properties of representations brought about by particular processes (e.g., dynamic mobilization).

Lastly, Tononi and Edelman's (1998) analysis recognizes the importance of the thalamo-cortical system in subtending consciousness (and could hence be viewed as specialized theory), but reaches this conclusion based on computational principles that are explicitly non-specialized to the extent that they could occur in any system properly structured.

A final comment on this analysis is that pure vehicle theories of consciousness remain problematic from a computational point of view, for they fail to make it clear how any aspect of consciousness could be produced exclusively by properties of the representational vehicles involved in information processing. Simply equating consciousness with stability in time (see, e.g., O'Brien and Opie, 1999), for instance, would not only force us to consider many physical systems to be conscious to some degree (thus raising the specter of panpsychism), but also appears to eschew any sort of computational explanation short of resorting to hitherto unknown causal properties of neural patterns of activity.

Toward computational principles for the distinction between conscious and unconscious cognition

What can we conclude from this brief overview of current computational approaches to consciousness? A salient point of agreement shared by several of the most popular current theories is that all such models, regardless of whether they assume specialized or non-specialized mechanisms, and regardless of whether they focus primarily on vehicles or on processes, converge toward assuming the following: Conscious representations differ from unconscious representations in that the former are endowed with certain properties such as their stability in time, their strength, or their distinctiveness. Cleeremans (Cleeremans and Jiménez, 2002; forthcoming) proposes the following definitions for these properties:

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, the prefrontral cortex, which plays a central role in working memory (Baddeley, 1986), is widely assumed to involve circuits specialized in the formation of the enduring representations needed for the active maintenance of task-relevant information (Frank et al., 2001; Norman and O'Reilly, 2001). Stability of representation is clearly related to availability to consciousness, to the extent that consciousness takes time. For instance, the brief stimuli associated with subliminal presentation will result in weaker representations than supraliminal presentation does.

Strength of representation simply refers to how many processing units are involved in a given representation, and to how strongly activated these units are. Strength can also be used to characterize the efficiency of a an entire processing pathway, as in the Stroop model of Cohen et al. (1990). Strong activation patterns exert more influence on ongoing processing than weak patterns, and are most clearly associated with automaticity, to the extent that they dominate ongoing processing.

Finally, *distinctiveness* of representation refers to the extent of overlap that exists between representations of similar instances. Distinctiveness, or discreteness, 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 context of the terminology associated with attractor networks, this contrast would thus be captured by the difference between attractors with a wide basin of attraction, which will tend to respond to a large number of inputs, and attractors with a narrow basin of attraction, which will only tend to respond to a restricted range of inputs. The notion also overlaps with the difference between episodic and semantic memory, that is, the difference between knowing that Brutus the dog bit you yesterday and knowing that all dogs are mammals: There is a sense in which the distinctive episodic trace, because it is highly specific to one particular experience, is more accessible and more explicit than the semantic information that all dogs share a number of characteristic features. This latter knowledge can be made explicit when the task at hand requires it, but is only normally conveyed implicitly (as a presupposition) by statements about or by actions directed toward dogs.

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.

The analysis presented above resonates well with recent computational models of overall cerebral function. O'Reilly and colleagues (McClelland et al., 1995; O'Reilly and Munakata, 2000; Atallah et al., 2004), for instance, have recently proposed that different regions of the brain have evolved to solve different — and incompatible — computational problems by using different representational formats and different learning regimes (McClelland et al., 1995). In their "tripartite" proposal, the brain is organized in three broad interacting systems: The hippocampus (HC), prefrontal cortex/basal ganglia (FC), and posterior cortex (PC). In this framework, each system uses similar, but not identical learning mechanisms and representational formats. The main function of HC is to rapidly learn about specific novel facts (episodic memory). Function of PC, in contrast, is to learn about the statistical regularities shared by many exemplars of a given domain (semantic memory). Finally, the main function of FC is to maintain information in an active state (active maintenance, subtending working memory) and to rapidly switch between active representations. Achieving each of these functions require different (but germane) learning mechanisms and different representational formats. Thus, HC uses the sparse, conjunctive representations necessary to avoid catastrophic interference, and a high learning rate that makes it possible to rapidly bind together the various elements of the current percept. PC, in contrast, slowly accumulates information over largely overlapping, distributed representations, so that broad semantic knowledge can progressively emerge over learning and development. Finally, FC is characterized by self-sustaining representational systems involving the recurrent connectivity necessary for active maintenance as well as the gating mechanisms necessary for rapid switching.

The three systems also differ from each other in terms of processing and learning mechanisms. Thus, O'Reilly and Munakata (2000) argue that the functions typically attributed to FC (i.e., working memory, inhibition, executive control, and monitoring or evaluation of ongoing behavior) require "activation-based processing", characterized by mechanisms of active maintenance through which representations can remain strongly activated for long periods of time as well as rapidly updated so as to make it possible for these representations to modulate processing elsewhere in the brain. Note how this is consistent with Crick and Koch's (2003) notion that "the front of the brain is looking at the back." Because of these properties, frontal representations are thus more accessible to verbalization and other reporting systems.[3] To this, they oppose "weight-based processing", characteristic of PC, in which knowledge is encoded directly by the pattern of connectivity between processing units and hence tends to remain tacit to the extent that this knowledge only manifests itself through the effects it exerts on ongoing processing rather than through the form of representations themselves.

In terms of learning mechanisms, O'Reilly and Munakata (2000) also propose an interesting distinction between model learning (Hebbian learning) and task learning (error-driven learning). Again, their argument is framed in terms of the different computational objectives each of these types of learning processes fulfills: Capturing the statistical structure of the environment so as to develop appropriate models of it on the one hand, and learning specific input–output mappings so as to solve specific problems (tasks) in accordance with one's goals on the other hand. There is a very nice mapping between this distinction — expressed in terms of the underlying biology and a consideration of computational principles — and the distinction between incidental learning and intentional learning on the other hand.

It is tempting to relate the different aspects of the quality of a representation delineated earlier with the functions of each system identified by O'Reilly and colleagues (McClelland et al., 1995; O'Reilly and Munakata, 2000; Atallah et al., 2004). Stability in time is what most saliently characterizes FC representations. Distinctiveness is a property most clearly associated with HC. Finally, PC representations are best characterized by their strength. Importantly, in this computational framework, there is no single system that is uniquely associated with the occurrence of conscious representations. Rather, conscious representations emerge as a result of the joint involvement of each system in ongoing processing.

Stability, strength, or distinctiveness can be achieved by different means. They can result, for instance, from the simultaneous top-down and bottom-up activation involved in the so-called "reentrant processing" (Lamme, 2004), 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.

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

---

[3] In this respect, O'Reilly and Munakata (2000) rightfully point out that a major puzzle is to understand how the FC comes to develop what they call a "rich vocabulary of frontal activation-based processing representations with appropriate associations to corresponding posterior-cortical representations" (p. 382).

> “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 high-quality representation thus appears to be a necessary condition for their availability to consciousness, one should ask, however, whether it is a sufficient condition. Cases such as hemineglect, blindsight (Weiskrantz, 1986), or, in normal subjects, attentional blink phenomena (Shapiro et al., 1997), or some instances of change blindness (Simons and Levin, 1997), for instance, suggest that quality of representation alone does not suffice, for even strong patterns can fail to enter conscious awareness unless they are somehow attended. 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 the so-called higher order thought (HOT) theories of consciousness (Rosenthal, 1997), according to which a mental state is conscious when the agent entertains, in a non-inferential manner, thoughts to the effect that it currently is in that mental state. Importantly, for Rosenthal, it is in virtue of current 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 a belief that the flower is yellow) and attitude (I know that the flower is yellow). Fully conscious knowledge is thus knowledge that is “attitude-explicit”.

This analysis suggests that another 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. Note that meta-representations are *de facto* assumed to play an important role in any theory that assumes interactivity. Indeed, for processes such as resonance, amplification, integration, or dynamic mobilization to operate, one minimally needs to assume two interacting components: A system of first-order representations, and a system of meta-representations that take first-order representations as their input.

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, forthcoming). There are at least four possibilities. Firstly, 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 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 in-directly, and only to the extent that these effects are sufficiently marked in the corresponding representations. This is equivalent to Dehaene's principle of "active firing" (Dehaene and Changeux, in press).

Secondly, 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 subthreshold effects, including subliminal priming, for instance.

Thirdly, a representation can be strong enough to enter conscious awareness, but fail 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. 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.

Finally, a representation can be so strong that its influence can no longer be controlled — automaticity. In these 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, 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.

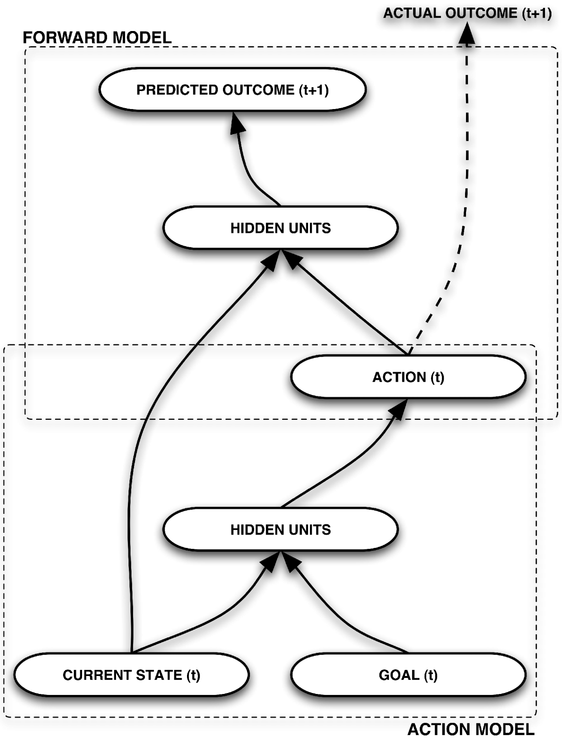

Fig. 2. A Forward Model. Two interconnected networks interact continuously: The action (inverse) model, the task of which is to produce appropriate actions given a representation of the current state and a goal (an intention), and the forward model, the task of which is to anticipate the sensory consequences (the next state) resulting from the model's actions.

Forward models

How might one go about capturing intuitions about the importance of both quality of representation and of meta-representations in the form of a computational model? There is an extremely interesting class of models that might provide a good starting point for exploring the computational principles described above (Fig. 2). These models are called "forward models" (Jordan and Rumelhart, 1992) and have been applied mostly in the domain of motor control so far (Miall and Wolpert, 1996; Jordan and Wolpert, 1999). Many control problems (and acting adaptively is the control problem per excellence) are difficult because they require solving two separate problems: (1) learning about the effects of particular actions on the environment, that is, developing a model of the system one is attempting to control (the ‘‘forward’’ model), and (2) learning which particular actions to take so as to achieve a desired goal, that is, learning how to control the system (the ‘‘inverse’’ problem). Forward models make it possible to solve both problems simultaneously. To do so, they generally consist of two interconnected networks. The first takes as input a goal and a description of the current state as input, and produces actions. The second, that is the forward model, takes the response of the first network (an action) and a description of the current state as input, and produces a prediction of how the to-be-controlled system (the ‘‘plant’’, in control theory parlance) would change if the produced action were carried out.

Crucially, the forward component of the model necessarily turns, as a result of training, into an internal model of the environment with which the network as a whole interacts. This sort of model can thus form the basis for a complex system of meta-representations that takes perceptual states and self-produced actions as input. It is also interesting to note that consistently with enactive and embodied perspectives on consciousness (Varela et al., 1991; O’Regan and Noë, 2001; Clark, 2002; Noë, in press; O’Regan et al., this volume), this model is totally dependent on action: Not only will it be shaped by the sorts of actions the model can enact on its environment, but it would not even be able to bootstrap itself were the system as a whole unable to act.

The fact that sophisticated internal models emerge as a result of the perception–action–anticipation loop that the system implements becomes particularly interesting when one additionally considers (1) that socialized agents not only interact with physical environments, but also with other agents, and (2) that agents also interact with themselves by recycling their expectations about the consequences of their own actions as perceptual input. The main implication of the first point is that a forward model that interacts with other agents will end up developing a model of the internal states of those agents (their ‘‘state of mind’’, so to speak). The main implication of the second point is that we now have a mechanism through which to flesh out the idea that thought is simulation (Hesslow, 2002; Grush, in press). When combined, however, the implications of these two points become particularly stimulating, for they suggest a mechanism through which representations of self could emerge out of an agent’s understanding of the internal states of other agents (Cleeremans, forthcoming) — an idea already hinted at by Rumelhart et al. (1986).

Several authors have recently begun to use such models as the cornerstone of theories in rather disparate domains ranging from motor behavior to cultural cognition and the development of theory of mind (Wolpert et al., 1998; Frith et al., 2000; Grush, in press; Hesslow, 2002; Holland and Goodman, 2003; Taylor, 2003; Wolpert et al., 2004). Frith and colleagues (2000), for instance, have proposed to analyze some of the symptoms of schizophrenia (i.e., delusions of control) or autism through lesions at various sites in the different components of forward models. Taylor’s (1999) CODAM model is built around the same assumptions (also see Aleksander, this volume). Miall (2003) noted the connection between such models and the mirror system discovered by Rizzolati and colleagues (1996). Forward models thus appear to be one of the most promising avenues for further exploration of the CCC, for they suggest a possible integrated functional account of different aspects of conscious experience — both low-level and high-level — as they occur in a system that is tightly coupled with its environment and with other agents.

Discussion and conclusions

In this paper, I have offered a survey of some recent computational models of consciousness, with the overall goal of suggesting that the unfolding search for the NCC should be augmented by a search for the CCC. I have suggested that whether a representation becomes available to consciousness depends on both properties associated with the representation (strength, stability, distinctiveness) and properties associated with the mechanisms through which the representation is redescribed in further, meta-representational systems.

An important benefit of engaging in a search for the CCC is that traditional dichotomies in the cognitive neurosciences (declarative versus procedural memory; implicit versus explicit learning; conscious versus unconscious perception, and so on) are now progressively replaced by accounts that take it as a starting point that such distinctions, rather than being set in stone and subtended by dedicated systems, instead emerge out of the interactions between different regions of the brain that have evolved to solve particular computational problems characterized by the fact that they are incompatible with each other. This focus on function and on mechanisms will undoubtedly contribute to naturalize consciousness. Architectures such as the forward models described in the previous section, while they remain very abstract, offer an intriguing avenue for further research in this direction.

In conclusion, a few pending issues relevant to the search for the CCC:

1. *Should consciousness be viewed as a graded or as an all-or-none phenomenon?* Some computational theories of consciousness, in particular global workspace models, assume that once a representation has entered the workspace, it is fully conscious. Dehaene specifically refers to this process as ''ignition'', and accordingly predicts that all measures of conscious awareness should systematically be strongly associated with each other (Dehaene et al., 1998, 2003; Dehaene and Naccache, 2001; Dehaene and Changeux, in press). In this view, consciousness is thus an all-or-none phenomenon. Other frameworks, in contrast, predict that consciousness is fundamentally graded (Cleeremans and Jiménez, 2002; Moutoussis and Zeki, 2002; Lamme, 2004). While there is a clear sense in which one is either aware or unaware of a stimulus (i.e., I perceive the stimulus or I do not), there are also other cases where there is a clear sense of gradedness in conscious experience (e.g., ambient noises, for instance, or perhaps chronic pains). Perceptual awareness also seems to depend in a graded manner on action systems; Marcel (1993) likewise suggests that it is far from being all-or-none. Note that it might also be the case that consciousness is both graded and all-or-none: Any complex system will exhibit non-linearities, and the physical word is replete with cases where continuous, graded changes in some dimension result in abrupt changes in some other dimension (e.g., continuous changes in the temperature of a body of water result in a change of state, say from liquid to solid).
2. *What is the relationship between attention and conscious awareness? What is the nature of the distinction between phenomenal and access consciousness?* Whether attention is necessary for consciousness or not remains a point of debate. Note that this debate is really one about how we should think about what best characterizes conscious states. Some authors take it that unattended perceptual states should simply be considered as unconscious (Dehaene and Changeux, in press), whereas others consider that such states can form part of the global phenomenology of a conscious subject even when unattended (O'Brien and Opie, 1999; Lamme, 2004). Defenders of the first perspective put more emphasis on the processes (access by systems of meta-representations), while defenders of the second put more emphasis on properties of representational vehicles themselves (strength, stability, distinctiveness). This is related to the distinction between A- and P-consciousness, which Block (1997) describes as involving a battle between biological and computational approaches to the mind. Whether A- and P-consciousness should be taken as different kinds of consciousness or whether they constitute points on a continuum thus remains an object of debate.
3. *What is the function of meta-representational systems?* While some functions of meta-representations are clear (e.g., monitoring and control), it is nevertheless challenging to build computational models that develop ''interesting'' (i.e., rich, structured) meta-representations. As suggested by the discussion of forward models, the difficulty arises likely from the fact that computational models are

often developed in isolation rather than in interaction with other agents. However, one probable function of meta-representations is that they are necessary to communicate one's internal states to others, and to infer internal states from the observation of others' behavior. Building models that acknowledge this extended character of consciousness is certainly one of the promising avenues of research in the context of the search for the CCC.

Uncited Reference

Tononi (2003).

Acknowledgments

Axel Cleeremans is a Senior Research Associate with the National Fund for Scientific Research (Belgium). This work was supported by an institutional grant from the Université Libre de Bruxelles to Axel Cleeremans and by grant HPRN-CT-1999-00065 from the European Commission. I thank an anonymous referee for very useful and detailed comments on a previous version of this article. I am afraid that lack of time prevented me from following all of this referee's suggestions, but they are duly noted.

References

Aleksander, I. (2000) How to Build a Mind. weidenfeld and Nicolson, London, UK.

Atallah, H., Frank, M.J. and O'Reilly, R.C. (2004) Hippocampus, cortex, and basal ganglia: insights from computational models of complementary learning systems. Neurobiology of Learning and Memory, 82: 253–267.

Atkinson, A.P., Thomas, M.S.C. and Cleeremans, A. (2000) Consciousness: mapping the theoretical landscape. Trends Cogn. Sci., 4(10): 372–382.

Atkinson, R.C. and Shiffrin, R.M. (1971) The control of short-term memory. Sci. Am., 224: 82–90.

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

Baars, B. J. (1994) A thoroughly empirical approach to consciousness, from http://psyche.cs.monash.edu.au/v1/psyche-1-06-baars.html

Baddeley, A.D. (1986) Working Memory. Oxford University Press, New York, NY.

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

Block, N. (1997) Biology versus computation in the study of consciousness. Behav. Brain Sci., 20(1): 159.

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

Chalmers, D.J. (2000) What is a neural correlate of consciousness? In: Metzinger T. (Ed.), Neural Correlates of Consciousness. Empirical and Conceptual Questions. MIT Press, Cambridge, MA, pp. 17–39.

Clark, A. (2002) Being There: Putting Brain, Body, and World Together Again. MIT Press, Cambridge, MA.

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. (forthcoming) Being Virtual. Oxford University Press, Oxford, UK. QA :3

Cleeremans, A., Destrebecqz, A. and Boyer, M. (1998) Implicit learning: news from the front. Trends Cogn. Sci., 2: 406–416.

Cleeremans, A. and Haynes, J.-D. (1999) Correlating consciousness: a view from empirical science. Rev. Int. Philos., 53: 387–420.

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.

Cohen, A., Dunbar, K. and McClelland, J.L. (1990) On the control of automatic processes: a parallel distributed processing account of the Stroop effect. Psych. Rev., 97: 332–361.

Colagrosso, M. D. and Mozer, M. C. (in press) Theories of access consciousness. Paper presented at the Neural Information Processing Systems 17. QA :4

Cotterill. (1998) Enchanted Looms. Conscious Networks in Brains and Computers. Cambridge University Press, Cambridge, UK. QA :5

Crick, F.H.C. and Koch, C. (1990) Towards a neurobiological theory of consciousness. Semin. Neuros., 2: 263–275.

Crick, F.H.C. and Koch, C. (1995) Are we aware of neural activity in primary visual cortex? Nature, 375: 121–123.

Crick, F.H.C. and Koch, C. (2003) A framework for consciousness. Nat. Neuros., 6(2): 119–126.

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

Dehaene, S. and Changeux, J.-P. (in press) Neural mechanisms for access to consciousness. In: Gazzaniga M. (Ed.), The Cognitive Neurosciences. QA :6

Dehaene, S., Kerszberg, M. and Changeux, J.-P. (1998) A neuronal model of a global workspace in effortful cognitive tasks. Proc. Natl. Acad. Sci. USA, 95(24): 14529–14534.

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

Dehaene, S., Sergent, C. and Changeux, J.-P. (2003) A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proc. Natl. Acad. Sci. USA, 100(14): 8520–8525.

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

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

Destrebecqz, A., Peigneux, P., Laureys, S., Degueldre, C., Del Fiore, G., Aerts, J., et al. (2003) Cerebral correlates of explicit sequence learning. Cognitive Brain Res., 16(3): 391–398. QA :7

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.

Edelman, G.M. (1989) The Remembered Present: A Biological Theory of Consciousness. Basic Books, New York, NY.

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

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.

Farah, M.J., O'Reilly, R.C. and Vecera, S.P. (1994) Dissociated overt and covert recognition of as an emergent property of a lesioned neural network. Psych. Rev., 100: 571–588.

Flohr, H. (1985) Sensations and brain processes. Behav. Brain Res., 71: 157–161.

Fragopanagos, N. and Taylor, J. G. (2003) A computational model of the attentional blink. Paper presented at the IJCNN. QA :8

Frank, M.J., Loughry, B. and O'Reilly, R.C. (2001) Interactions between frontal cortex and basal ganglia in working memory: a computational model. Cognitive, Affect. Behav. Neuros., 1(2): 137–160.

Franklin, S. and Graesser, A.C. (1999) A software agent model of consciousness. Conscious. Cogn., 8: 285–305.

Freud, S. (1949) An Outline of Psychoanalysis (J. Strachey, Trans.). Hogarth Press, London.

Frith, C.D., Blakemore, S.-J. and Wolpert, D.M. (2000) Explaining the symptoms of schizophrenia: abnormalities in the awareness of action. Brain Res. Rev., 31: 357–363.

Frith, C.D., Perry, R. and Lumer, E. (1999) The neural correlates of conscious experience: an experimental framework. Trends Cogn. Sci., 3: 105–114.

Gray, J. (2004) Consciousness: Creeping up on the Hard Problem. Oxford University Press, Oxford.

Gregory, R.L. (Ed.). (2003) The Oxford Companion to the Mind (2nd ed.). Oxford University Press, Oxford, UK.

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

Grush, R. (in press) The emulation theory of representation: motor control, imagery, and perception. Behav. Brain Sci. QA :9

Hesslow, G. (2002) Conscious thought as simulation of behaviour and perception. Trends Cogn. Sci., 6(6): 242–247.

Holland, O. (Ed.). (2003) Machine Consciousness. Imprint Academic, Exeter, UK.

Holland, O. and Goodman, R. (2003) Robots with internal models. A route to machine consciousness? In: Holland O. (Ed.), Machine Consciousness. Imprint Academic, Exeter, UK, pp. 77–109.

Jordan, M.I. and Rumelhart, D.E. (1992) Forward models: supervised learning with a distal teacher. Cogn. Sci., 16: 307–354.

Jordan, M.I. and Wolpert, D.M. (1999) Computational motor control. In: Gazzaniga M. (Ed.), The Cognitive Neurosciences. MIT Press, Cambridge, MA.

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.

Knoblich, G., Elsner, B., Aschersleben, G. and Metzinger, T. (Eds.). (2003) Self and Action. Special Issue of Consciousness and Cognition. (Vol. 12, 4). QA :10

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

Lamme, V.A.F. (2004) Separate neural definitions of visual consciousness and visual attention; a case for phenomenal awareness. Neural Networ., 17(5–6): 861–872.

Logothethis, N. and Schall, J. (1989) Neuronal correlates of subjective visual perception. Science, 245: 761–763.

Lumer, E.D., Friston, K.J. and Rees, G. (1998) Neural correlates of perceptual rivalry in the human brain. Science, 280: 1931–1934.

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

Marcel, A.J. (1993) Slippage in the unity of consciousness. In: Bock G.R. and Marsh J. (Eds.), Experimental and Theoretical Studies of Consciousness (Ciba Foundation Symposium 174). John Wiley and Sons, Chichester, pp. 168–186.

Mathis, W.D. and Mozer, M.C. (1995) On the computational utility of consciousness. In: Tesauro G. and Touretzky D.S. (Eds.) Advances in Neural Information Processing Systems, Vol. 7. MIT Press, Cambridge, pp. 10–18.

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, N.J., pp. 324–328.

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.

Metzinger, T. (2000) Neural Correlates of Consciousness. Empirical and Conceptual Questions. MIT Press, Cambridge, MA.

Miall, R.C. (2003) Connecting mirror neurons and forward models. Neuroreport, 14(16): 1–3.

Miall, R.C. and Wolpert, D.M. (1996) Forward models for physiological motor control. Neural Networ, 9(8): 1265–1279.

Moutoussis, K. and Zeki, S. (2002) The relationship between cortical activation and perception investigated with invisible stimuli. Proc. Natl. Acad. Sci. USA, 99(4): 9527–9532.

Newman, J. and Baars, B.J. (1993) A neural attentional model of access to consciousness: a global workspace perspective. Conc. Neurosci., 4: 255–290.

Nisbett, R.E. and Wilson, T.D. (1977) Telling more than we can do: verbal reports on mental processes. Psychol. Rev., 84: 231–259.

Noë, A. (in press) Action in Perception. MIT Press, Boston, MA.

Noë, A. and Thompson, E. (2004) Are there neural correlates of consciousness? J. Conscious. Stud., 11(1): 3–28.

Norman, K. and O'Reilly, R. C. (2001) Modeling Hippocampal and Neocortical Contributions to Recognition Memory: A Complementary Learning Systems Approach (No. Technical Report 01-02): Institute of Cognitive Science — University of Colorado, Boulder.

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(5): 883–917.

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

Perruchet, P. and Vinter, A. (2003) The self-organizing consciousness. Behav. Brain Sci., 25(3): 297.

Prinz, J.J. (2003) Level headed mysterianism and artificial experience. In: Holland O. (Ed.), Machine Consciousness. Imprint Academic, Exeter, UK, pp. 111–132.

Rizzolati, G. (1996) Premotor cortex and the recognition of motor actions. Cogn. Brain Res., 3: 131–141.

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., Smolensky, P., McClelland, J.L. and Hinton, G.E. (1986) Schemata and sequential thought processes in PDP models. In: McClelland J.L. and Rumelhart D.E. (Eds.) Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models. MIT Press, Cambridge, MA, pp. 7–57.

Schacter, D.L. (1989) On the relations between memory and consciousness: dissociable interactions and conscious experience. In: H.L.R. III and Craik F.I.M. (Eds.), Varieties of Memory and Consciousness: Essays in Honour of Endel Tulving. Lawrence Erlbaum Associates, Mahwah, NJ.

Shapiro, K.L., Arnell, K.M. and Raymond, J.E. (1997) The Attentional Blink. Trends Cogn. Sci., 1: 291–295.

Simons, D.J. and Levin, D.T. (1997) Change Blindness. Trends Cogn. Sci., 1: 261–267.

Sun, R. (2001) Duality of the Mind. Lawrence Erlbaum Associates, Mahwah, NJ.

Taylor, J.G. (1999) The Race for Consciousness. MIT Press, Cambridge, MA.

Taylor, J.G. (2003) Paying attention to consciousness. Prog. Neurobiol., 71: 305–335.

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. (in press) Consciousness and the brain: some theoretical considerations. Prog. Brain Res. xx(xx), xx.

Tononi, G. and Edelman, G.M. (1998) Consciousness and complexity. Science, 282(5395): 1846–1851.

Varela, F.J., Thompson, E. and Rosch, E. (1991) The Embodied Mind: Cognitive Science and Human Experience. MIT Press, Cambridge, MA.

Wegner, D.M. (2002) The Illusion of Conscious Will. Bradford Books, MIT Press, Cambridge, MA.

Weiskrantz, L. (1986) Blindsight: A case study and implications. Oxford University Press, Oxford, England.

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

Wolpert, D.M., Miall, R.C. and Kawato, M. (1998) Internal models in the cerebellum. Trends Cogn. Sci., 2: 338–347.

Zeki, S. and Bartels, A. (1998) The asynchrony of consciousness. Proc. Roy. Soc. B, 265: 1583–1585.

Zeman, A. (in press) The concept of consciousness. Prog. Brain Res., xx(xx), xx–xx.

AUTHOR QUERY FORM

ELSEVIER

Progress in Brain Research

JOURNAL TITLE:	**PBR**
ARTICLE NO:	**50007**

Queries and / or remarks

Query No	Details required	Author's response
AQ1	Please provide the expansion of the acronym NMDA.	
AQ2	Please provide the expansion of the acronym fMR.	
AQ3	Please update the ref. for Cleeremans.	
AQ4	Please update and provide place of publication for Colegrosso and Mozer.	
AQ5	Please provide the initials for Cotterill (1998).	
AQ6	Please update the ref. for Dehaene and Changeux.	
AQ7	Please provide all authors names for Destrebecqz et al. (2003).	
AQ8	Please provide place for Fragopanagos and Taylor (2003).	
AQ9	Please update the ref. for Grush.	
AQ10	Please provide publisher and place of publication for Knoblich et al. (2003).	
AQ11	Please update the ref. for Noe.	
AQ12	Please update the ref. for Tononi.	
AQ13	Please update the ref. for Zeman.	