COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 38 (2014) 1286–1315

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12149

Connecting Conscious and Unconscious Processing

Axel Cleeremans

Université Libre de Bruxelles

Received 21 February 2011; received in revised form 3 April 2014; accepted 3 April 2014

Abstract

Consciousness remains a mystery—"a phenomenon that people do not know how to think about—yet" (Dennett, 1991, p. 21). Here, I consider how the connectionist perspective on information processing may help us progress toward the goal of understanding the computational principles through which conscious and unconscious processing differ. I begin by delineating the conceptual challenges associated with classical approaches to cognition insofar as understanding unconscious information processing is concerned, and to highlight several contrasting computational principles that are constitutive of the connectionist approach. This leads me to suggest that conscious and unconscious processing are fundamentally connected, that is, rooted in the very same computational principles. I further develop a perspective according to which the brain continuously and unconsciously learns to redescribe its own activity itself based on constant interaction with itself, with the world, and with other minds. The outcome of such interactions is the emergence of internal models that are metacognitive in nature and that function so as to make it possible for an agent to develop a (limited, implicit, practical) understanding of itself. In this light, plasticity and learning are constitutive of what makes us conscious, for it is in virtue of our own experiences with ourselves and with other people that our mental life acquires its subjective character. The connectionist framework continues to be uniquely positioned in the Cognitive Sciences to address the challenge of identifying what one could call the "computational correlates of consciousness" (Mathis & Mozer, 1996) because it makes it possible to focus on the *mechanisms* through which information processing takes place.

Keywords: Consciousness; Learning; Connectionist modeling; Metacognition

1. Introduction

About 25 years ago, as an undergraduate student at the Université Libre de Bruxelles, I was lucky to attend a lecture that Donald Broadbent delivered in the stately, wood-paneled

Correspondence should be sent to Axel Cleeremans, Consciousness, Cognition & Computation Group, Université Libre de Bruxelles CP 191, 50 ave. F.-D. Roosevelt, B1050 Bruxelles, Belgium. E-mail: axcleer@ulb.ac.be

hall of the University Foundation in the center of Brussels. There, Broadbent presented his latest work on "implicit learning," a phenomenon first explored by Arthur Reber in 1967 whereby people are shown to be able to learn about novel information without intention to do so and without awareness of the underlying regularities. Broadbent's ingenious experiments propelled the field in hitherto unexplored directions and proved a seminal source of inspiration for my own work. Thus, I quickly set out to conduct a series of experiments modeled after Berry and Broadbent (1984).

Because I also had an interest in thinking about experimental phenomena through the lens of computational modeling, I began exploring how one could conceive of a theory of implicit learning that was amenable to computational instantiation. This proved to be surprisingly challenging. Indeed, when Reber (1967) proposed that learning can proceed without intention and without awareness, theorizing in the psychological sciences was dominated by the so-called classical models of information processing (e.g., Newell & Simon, 1972). Such models, up until the early 80s, all seemed to begin with the idea that learning is driven by hypothesis testing. Likewise, such models assumed that knowledge always consists of abstract, declarative, propositional-like representations. Implicit learning seemed to be irreconcilable with such assumptions: Not only does it often fail to result in propositional, verbalizable knowledge, but its central characteristic is probably its incidental nature, that is, the fact that one becomes sensitive to novel information merely through processing the material. Unlike what is the case when one learns about new facts or new procedures by being told about them, knowledge acquisition in implicit learning situations is not driven by consciously held hypotheses (though, of course, participants in psychology experiments will always formulate and test hypotheses when asked to participate in a study: consciousness cannot be "turned off"). Thus, the notion that learning can proceed unintentionally seemed to present a singular challenge for traditional perspectives on how change occurs in cognitive systems.

When I learned about the ideas that the PDP group (McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986) were pioneering at the time, I experienced an epiphany of sorts: There was now a clear, coherent framework with which to think about learning in a manner that was finally divorced from the classical assumptions, at least insofar as they were held with respect to human learning. Connectionism seemed to provide a solid alternative theory through which to understand how knowledge may accrue in cognitive systems through little more than mere exposure, almost as a side effect of information processing. The fact that connectionism seemed to provide the appropriate foundations with which to think about implicit learning is what prompted me to go work at Carnegie Mellon, first with Lynne Reder and then with Jay McClelland. The 4 years that followed proved incredibly exciting, for the work we carried out clearly demonstrated how one could build the lineaments of a theory of implicit learning based on the novel conceptual foundations provided by connectionism.

Here, I would like to visit these issues again and consider how our perspective has changed over the last 20 years. I begin by spelling out the conceptual challenges associated with accounting for unconscious cognition and implicit learning based on classical assumptions. Next, I assess the implications of a connectionist approach to the phenomena of

implicit cognition and attempt to highlight a few of the most important principles that underpin this endeavor. In the third section, I speculate on how connectionism can help us understand not only cognition without consciousness but also cognition *with* consciousness, and we propose the idea that the latter crucially depends on *learned redescriptions of a system's own activity*—a proposal that I have elsewhere dubbed "the Radical Plasticity Thesis" (Cleeremans, 2008).

2. The trouble with classical approaches

In Cleeremans (1997) and also in Cleeremans and Jiménez (2002), I suggest that the central reason why dissociations between conscious awareness and behavior remain so controversial, even today, is fundamentally a conceptual one—namely that the phenomena of implicit cognition cannot be reconciled with classical perspectives on information processing.

Empirically, the central characteristic of unconscious processing is the observation that an agent's behavior is influenced by knowledge of which it remains unaware. In Cleeremans (1997), I define implicit knowledge as follows:

> At a given time, knowledge is implicit when it can influence processing without possessing in and of itself the properties that would enable it to be an object of representation.

Thus, unconscious knowledge is knowledge that is causally efficacious yet unavailable to form the contents of conscious experience. Now, consider the manner in which knowledge is represented in classical models of cognition (Anderson, 1983; Newell, 1990). Such models—roughly speaking, the "Computational Theory of Mind" (see Fodor, 1975, 1983; Fodor & Pylyshyn, 1988, Pylyshyn, 1980, 1984) take it as a starting point that cognition consists of symbol manipulation. The flow of information processing in classical models goes roughly like this: There is a central processor that fetches or stores information in knowledge bases and processes it. The processor interacts with the world through input/output systems. Knowledge (either "programs" or "data") is represented symbolically. Bates and Elman (1993) dubbed this perspective on cognition "The First Computer Metaphor of Cognition" and characterized it as follows (p. 630):

> At its core, the serial digital computer is a machine that manipulates symbols. It takes individual symbols (or strings of symbols) as its input, applies a set of stored algorithms (a program) to that input, and produces more symbols (or strings of symbols) as its output. These steps are performed one at a time (albeit very quickly) by a central processor. Because of this serial constraint, problems to be solved by the First Computer Metaphor must be broken down into a hierarchical structure that permits the machine to reach solutions with maximum efficiency (e.g., moving down a decision tree until a particular subproblem is solved, and then back up again to the next step in the program).

In such systems, thus, knowledge always takes the form of symbolic propositions stored in a mental database (i.e., productions rules or declarative statements). There are two important and problematic features about such representations. First, they remain causally inert until activated or otherwise accessed by the processor. Second, their shape (symbolic propositions) makes their contents immediately accessible by the processor. The conjunction of these two features renders the entire approach incapable of accounting for unconscious cognition, for it entails that representations cannot influence processing independently of being accessed (activated, manipulated) by the processor. However, this property—causal influence without access—is precisely what one means when one says that knowledge is unconscious. A production rule, for instance, cannot influence ongoing processing unless the algorithm that drives the entire system has established that the rule's preconditions match current input and that the rule could now be applied. Intuitively, this is akin to a human participant who figures out, when confronted with a mathematical problem, for instance, that an arithmetical expression can be simplified in certain ways described by a rule that he has learned. But this process is clearly a conscious process. Now consider what happens when a chess expert intuitively decides to move a particular chess piece. One could claim that the same process described above in the case of the arithmetic problem now takes place: A heuristic rule is identified as being relevant to the particular situation at hand and applied. However, the chess expert claims that he is unable to justify his choice: The move he made is just what came to mind. Perhaps he could explain the specific reasons why he chose that particular move given sufficient time and effort, but the move itself simply appeared to pop in his mind. The difference between the arithmetical problem and the chess move is one of consciousness: One seems to have access to the relevant knowledge in the first instance, but not in the second.

Now here is the key argument: If one assumes, as do thoroughly classical approaches to cognition, that the mechanisms involved in each case always entail accessing and activating the relevant rule, then one is left with no principled difference between cognition with and without awareness, for in both cases, the very same mechanisms (specifically: access to the relevant knowledge) are involved.

More formally, the argument could be spelled out in this way:

1. Awareness of some knowledge entails access to the relevant representations.
2. In classical models, representations take the form of symbolic propositions.
3. Symbolic propositions cannot be causally efficacious unless they are accessed.

Therefore, in classical models, causally efficacious representations are necessarily conscious. Briefly put thus, the argument I introduced in Cleeremans (1997) is this: If you believe that cognition consists exclusively of manipulating structured, symbolic, propositional representations, then you only have two possibilities of accounting for the phenomena of implicit cognition. You can either (1) ascribe them to a separate "psychological unconscious" (Kihlstrom, 1987, 1990) that is capable of performing exactly the same sorts of computations as your conscious system is (specifically: access to the relevant knowledge), only minus consciousness (Searle, 1992), or (2) explain them away by reject-

ing existing evidence for implicit cognition altogether and claim that all of cognition involves conscious knowledge (e.g., Shanks & St. John, 1994).

There is also a third possibility, which consists of rejecting the idea that unconscious cognition always involves symbol manipulation. This is the "third way" that connectionism has made so salient over the past 20 years. While the classical perspective takes it as a starting point that information processing involves operations modeled after conscious cognition, connectionism turns this perspective on its head and proposes that information processing begins with unconscious cognition. It is worth pointing out here that many contemporary approaches rooted in symbolic processing (e.g., ACT-R, CLARION) have evolved to the point that they share many features more typically associated with connectionist models, such as associative processing.

Once we eliminate the idea that *all* of cognition, be it with or without consciousness, involves symbol manipulation, we can then focus on exploring what we can do without symbols. We are then facing the great challenge of figuring out how we can get symbols in the game after all, but at least we begin with more plausible assumptions. In the next section, I briefly overview how connectionism has changed our understanding of unconscious cognition.

3. What have we learned from connectionism?

Connectionist models have provided genuine insights into how knowledge can influence processing without access—a hallmark of unconscious processing—and of how change can accrue as a result of mere information processing—a hallmark of the phenomena of implicit learning. Numerous models of implicit learning based on connectionist models have now been proposed (see Cleeremans & Dienes, 2008; for a recent review), and it is fair to say that such models have been very successful in accounting for the mechanisms that subtend performance in a wide range of relevant empirical paradigms (see Cleeremans, Destrebecqz, & Boyer, 1998; for an overview), from artificial grammar learning (e.g., Dienes, 1992) and sequence learning (Cleeremans & McClelland, 1991) to process control (Gibson, Fichman, & Plaut, 1997) or priming (Mathis & Mozer, 1996).

The first fully implemented connectionist models of implicit learning are found in the early efforts of Dienes (1992) and of Cleeremans and McClelland (1991). While authors such as Brooks (1978) and Berry and Broadbent (1984) had already suggested that performance in implicit learning tasks such as artificial grammar learning or process control may be based on retrieving exemplar information stored in memory arrays, such models have in general been more concerned with accounting for performance at retrieval rather than on accounting for learning itself. The connectionist approach, by contrast, has been centrally concerned with the mechanisms involved during learning since its inception, and therefore constitutes an excellent candidate framework with which to think about the processes involved in implicit learning.

My purpose here is not to review these developments in detail (see Cleeremans & Dienes, 2008), but rather to focus on how several fundamental principles that characterize

the connectionist approach are relevant to our understanding of the differences between conscious and unconscious processing. In the following, I discuss each in turn.

3.1. Active representation

As discussed above, this first principle highlights a fundamental difference between classical and connectionist representations, namely that the former are inherently passive whereas the latter are continuously active. Indeed, the symbolic, propositional representations characteristic of classical models of cognition (i.e., production rules and declarative knowledge) are intrinsically passive: They are objects (data structures) stored in mental databases and can only influence ongoing processing when an algorithm (i.e., an inference engine) has determined that certain trigger conditions are met. Thus, for a classical representation to be causally efficacious, it first needs to be accessed or otherwise made active in some way. But, as discussed above, this necessary link between causal efficacy and access is immediately problematic for our conceptualization of the differences between information processing with and without awareness. The difficulty stems from the (tacitly) assumed equivalence between causal efficacy, access, and consciousness. This equivalence in turn stems from the fact that in classical perspectives on cognition, there is a complete separation between representation and processing. Connectionism solves this quandary very elegantly by proposing that access is not necessary to drive information processing. Nothing "accesses" anything in a connectionist network. Instead, connectionist models assume that all the long-term knowledge accrued over experience is embedded in the very same structures that support information processing, that is, the connection weights between processing units. Such knowledge therefore does not need to be accessed in any way to be causally efficacious; it simply exerts its influence automatically whenever the units whose activation propagates through the relevant connections are active. Thus, knowledge in connectionist networks is active *in and of itself*, and fundamental phenomena such as priming are accounted for naturally without the need to postulate additional mechanisms.

An important consequence of the fact that long-term knowledge in connectionist networks accrues in connection weights as a mandatory consequence of information processing is that connectionist models capture, without any further assumptions, two of the most important characteristics of implicit learning, namely (1) the fact that learning is incidental and mandatory, and (2) the fact that the resulting knowledge is difficult to express. A typical connectionist network, indeed, does not have direct access to the knowledge stored in connection weights. Instead, this knowledge can only be expressed through the influence that it exerts on the model's representations, and such representations may or may not contain readily accessible information, that is, information that can be retrieved with no or low computational cost (see Kirsh, 1991). Arguably, symbolic approaches may capture the same distinction through the difference between *compiled* and *interpreted* code. It would be too long to discuss the finer issues raised by this possibility here, but two points are worth mentioning. First, all compiled code necessarily existed as interpreted code before compilation took place. This makes the strong

prediction, under the assumption that compiled code corresponds to unconscious knowledge and that interpreted code corresponds to conscious knowledge, that all the unconscious knowledge we possess at some point previously existed as conscious knowledge. Whether this holds true or not is a matter for empirical investigation, but there is evidence that we are sensitive to regularities that were never made explicit (see Pacton, Perruchet, Fayol, & Cleeremans, 2001, for an example in the domain of learning orthographic regularities). Second, whether compiled or interpreted, symbolic computer code always needs a processor to execute it (and hence access it) for it to be causally efficacious. This stands in sharp contrast with the patterns of connection weights that drive processing in connectionist networks, which exert their effects directly, merely as a result of transmitting activation.

3.2. Emergent representation

The second principle simply states the following: Sensitivity to some regularity does not necessarily imply that the regularity is itself represented as an object of representation. What I mean by this is the following: It is not because you observe that the actions of an agent indicate that it is sensitive to certain regularities (such as in implicit learning situations) that you can conclude that these regularities are represented in its cognitive system as objects of representation that the agent can manipulate intentionally. There are so many examples of the importance of this principle that entire books have been written about it—see, for instance, the nice popularized treatment of this issue by Steven Johnson (2002), simply titled "Emergence." Thus, bees construct complex nests and perfectly regular hexagonal cells without any evidence that they even have simple representations of the overall structure of the nest. It is hard not be reminded of behaviorism in this context, but this is certainly one thing behaviorism got right: You do not always need internal representations to account for complex behavior. Of course, one must always be careful not to throw away the baby with the bathwater, to revisit an old cliché: We undeniably entertain systems of complex representations that we can access, manipulate, ponder about, and so on—just not always, and just not for anything.

In cognitive psychology, this principle of "Emergent Representation" has been expressed most clearly through dynamical approaches to Cognitive Science (van Gelder, 1998; Port & van Gelder, 1995) and through the connectionist approach (McClelland, 2010). To illustrate, consider the fact that connectionist networks can be often described as obeying rules without possessing anything like rule-like representations. A very well-known example is Rumelhart and McClelland's (1986a,b) model of the acquisition of the past tense morphology. In the model, not only are regular verbs processed in just the same way as exceptions, but they are not learned through anything like processes of rule acquisition.

Another example that attracted considerable attention when it was first reported is Hinton's (1986) "family-trees" demonstration that a back-propagation network can, through training, become sensitive to the structure of its stimulus environment in such a way that this sensitivity is clearly removed from the surface features of the stimulus material. In

Hinton's words, "The structure that must be discovered in order to generalize correctly is not present in the pairwise correlations between input units and output units" (p. 9). The model thus exhibits sensitivity to functional similarity based on the distributional information present in the input, and, as a result, develops abstract knowledge of the relevant dimensions of the domain. Because it is so illustrative of the points I wish to make here, it is worth going over this simulation in some detail.

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 was 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 rather than by their mere similarity expressed, for instance, in terms of how many features (input units) they share—a property that characterizes sensitivity to *functional* rather than *physical* similarity. 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 knowledge can simply emerge as a by-product of processing in structured domains. This introduces a crucial distinction, one that I will return to later, between *sensitivity* and *awareness.*

As a final example, consider also that a Simple Recurrent Network (Elman, 1990) trained on only some of the strings that may possibly be generated from a finite-state grammar will generalize to the infinite set of all possible grammatical instances (Cleeremans, Servan-Schreiber, & McClelland, 1989; Servan-Schreiber et al., 1991), thus demonstrating perfect, rule-like generalization based only on the processing of a necessarily finite set of exemplars. Interestingly, the representations developed by the network when trained on such material exhibited, under certain conditions, the remarkable property of corresponding almost perfectly with the nodes of the grammar: Cluster analyses indeed showed that the similarity structure of the learned internal representations that the network has developed about the relationships between each sequence element and its possible successors reflects the structure of the very grammar the network had been trained

on. Again, and crucially, such structure simply emerges out of exposure to relevant stimuli.

3.3. Graded processing

The third principle states that information processing as carried out by the brain (i.e., neural computation) is inherently graded (see Munakata, 2001 for an excellent overview). Note that this is not incompatible with the observation of all-or-none outputs. In fact, the logistic function that is so central to many neural network models demonstrates how the relationship between two quantities can be simultaneously graded and dichotomous, just as continuous variations in the temperature of a body of water can make it change state (i.e., freeze) at a critical point. Again, the connectionist literature is replete with striking demonstrations of this principle (see Elman et al., 1996). One of the clearest is perhaps McClelland's (Schapiro & McClelland, 2009) model of the balance scale problem, in which continuous, incremental learning nevertheless produces both the plateaus and the abrupt, stage-like changes in performance characteristics of many aspects of cognitive development. Another potent illustration of how graded representations can nevertheless produce complex patterns of associations and dissociations between several aspects of behavior is provided by the work of Munakata, McClelland, Johnson, and Siegler (1997) on object permanence, in which a Simple Recurrent Network was used to model children's ability to keep active representations of hidden objects. In both cases, the graded nature of the underlying representations is crucial in producing the observed effects; that is, it is precisely in virtue that representations are graded that such models are successful in accounting both for the steady changes characteristic of plateaus and for the abrupt changes characteristic of stage-like transitions. Again, while the implications of graded processing are perhaps clearest in the case of development, they are just as relevant to our understanding of the differences between conscious and unconscious processing for they highlight the fact that qualitative differences can accrue from purely quantitative changes. Whether consciousness is graded or all-or-none is both an important empirical debate (Sandberg et al., 2010; Sergent & Dehaene, 2004; Windey et al., 2013) as well as challenging conceptual issue, for it is the case that graded output can be obtained based on the operation of all-or-none computing elements, and that all-or-none output can be obtained based on the operation of graded computing elements. Connectionism, in many cases, has given us new conceptual tools with which to think about the distinction between graded and all-or-none processing.

3.4. Mandatory plasticity

This final principle states that learning is a mandatory consequence of information processing. Thus, the brain is inherently plastic. Every experience leaves a trace in many neural pathways. William James stated that "Every impression which impinges on the incoming nerves produces some discharge down the outgoing ones, whether we be aware of it or not" (James, 1890, vol. 2, p. 372). Donald Hebb (1949) later operationalized this

idea in the form of what is now known as the Hebb rule, which simply states that activity between two neurons will tend to increase whenever they are simultaneously active. The Hebb rule, unlike other learning procedures, actually forms the basis for elementary mechanisms of plasticity in the brain, namely long-term potentiation (see Bliss & Lomo, 1973) and depression.

O'Reilly and Munakata (2000) proposed an interesting distinction between what they called "model learning" (Hebbian learning) and "task learning" (error-driven learning). Their argument is framed in terms of the different computational objectives that 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. Thus, as made clear by the manner in which information processing is construed in the connectionist framework, (1) representations are dynamical, constantly causally efficacious objects, and (2) change occurs as soon as information processing takes place.

The fact that learning is almost viewed as a by-product of information processing networks accounts very naturally (that is, without requiring further assumptions) for a host of phenomena associated with unconscious cognition, and in particular with implicit learning.

To summarize, these four connectionist principles—active representation, emergent representation, graded processing, and mandatory plasticity—help us recast the differences between conscious and unconscious cognition in a manner that is strikingly different from thoroughly classical approaches. Instead of assuming that representations take the form of inert symbolic propositions that cannot be active unless they are somehow accessed, we now have a constantly causally efficacious network of subsymbolic computational elements (units, neurons). These features make it easy to understand how knowledge can influence behavior in a way that does not entail that the relevant representations be accessed as objects of representation, which is precisely what happens in the many phenomena characteristic of implicit cognition, such as priming, implicit learning, and implicit memory.

However, we now face the even greater challenge of understanding how such systems can also account for consciousness. What are the computational principles through which one can characterize the differences between conscious and unconscious representations? This is the question that I attempt to sketch an answer to in the next section.

4. Consciousness

Numerous theories of consciousness have been proposed over the last 20 years (see Atkinson, Thomas, & Cleeremans, 2000)—each author in this burgeoning domain seems to have his or her own theory of consciousness. While it would take an entire book to

attempt to summarize the state-of-the-art in this respect, it is probably sufficient for the purposes of this text to point out that most theories fall into two (broadly defined) camps: global workspace theories and higher order theories.

Global workspace theory (GWT, see Baars, 1988; Dehaene, Kerszberg, & Changeux, 1998) is currently the most consensual account of the functional characteristics of consciousness. According to GWT, conscious representations are globally accessible in a manner that unconscious representations are not. Global accessibility, that is, the capacity for a given representation to influence processing on a global scale (supporting, in particular, verbal report), is achieved by means of "the neural workspace," a large network of high-level neural "processors" or "modules" linked to each other by long-distance cortico-cortical connections emanating from layer 5 of the cortex. Thus, while information processing can take place without awareness in any given specialized module, once the contents processed by that module enter in contact with the neural workspace, "ignition" occurs and the contents are "broadcast" to the entire brain, so achieving what Dennett (2001) has dubbed "fame in the brain." In this respect, it is interesting to note that in some ways, early connectionist models such as the interactive activation model (McClelland, 1981) already contain the lineaments of GWT.

GWT thus solves the quandary spelled out in the introduction (i.e., which computational principles differentiate between conscious and unconscious cognition) by distinguishing between causal efficacy and conscious access through architecture: On one hand, knowledge embedded in peripheral modules can bias and influence processing without entering the global workspace, and so remain unconscious. On the other hand, knowledge that is sufficiently supported both by bottom-up factors such as stimulus strength and by top-down factors such as attention can "mobilize" the neural workspace, resulting in "ignition" and so become conscious and available for the global control of action.

Higher order thought (HOT) theories of consciousness (Lau & Rosenthal, 2011; Rosenthal, 1997) have a very different flavor. According to HOT, a mental state is conscious when the agent entertains, in a non-inferential manner, thoughts to the effect that it currently is in that mental state. Importantly, for Rosenthal, it is in virtue of occurrent HOTs that the target first-order representations become conscious. In other words, a particular representation, say, a representation of the printed letter "J," will only be a conscious representation to the extent that there exists another (unconscious) representation (in the same brain) that indicates the fact that a (first-order) representation of the letter "J" exists at time *t*. Dienes and Perner (1999) have elaborated this idea by analyzing the implicit-explicit distinction as reflecting a hierarchy of different manners in which a given representation can be explicit. Thus, a representation can explicitly indicate a property (e.g., "yellow"), predication to an individual (the flower is yellow), factivity (it is a fact and not a belief that the flower is yellow), and attitude ("I know that the flower is yellow"). Fully conscious knowledge is thus knowledge that is "attitude-explicit." A conscious state is thus necessarily one that the subject if conscious of. While this sounds highly counter-intuitive to some authors (most notably Ned Block, see e.g., Block, 2011), it captures the central intuition that it is precisely the fact that I know (that I experience the fact, that I feel) that I possess some knowledge that makes this knowledge conscious.

HOT thus solves the problem of distinguishing between conscious and unconscious cognition in a completely different manner, specifically by assuming the involvement of specific kinds of representations the function of which it is to denote the existence of and to qualify target first-order representations. Such HOTs, or metarepresentations, need not be localized in any particular brain region, but of course the densely interconnected prefrontal cortex is a good candidate for such metarepresentations to play out their functions.

Regardless of whether one takes GWT or HOT to best characterize the differences between conscious and unconscious cognition, one question that connectionist thinking about this issue prompts us to ask is: *How do we get there*? How do we *build* the global workspace? Where do metarepresentations come form?

Considering existing theories of consciousness through a connectionist lens offers the tantalizing possibility not only of unifying the two accounts but also of rooting them both in mechanisms of learning. On this view, unconscious representations constantly compete with each other to capture the best interpretation of the input (Maia & Cleeremans, 2005). This competition is biased by further representations that capture the system's high-level, learned knowledge (its expectations and its goals). The "winning coalitions" come to dominate processing as the result of prior learning, and hence afford the global availability claimed to be constitutive of consciousness by GWT. Global availability is not sufficient, however, for one can perfectly imagine all of the aforementioned to take place without consciousness (as any interactive neural network readily demonstrates). What I surmise to be also necessary, congruently with the assumptions of HOT, is that the winning representations *be known as objects of representation* by the system that possesses them. In other words, that first-order representations be redescribed by other representations in such a way as to make the former be identified or recognized by the system as familiar states of knowledge, that is, "attitude-explicit" in the terminology of Dienes and Perner.

In the following, I first attempt to flesh out the main computational principles that differentiate GW-like theories from HOT theories of consciousness. Next, I describe recent simulation work in which we specifically explore how it may be possible to build connectionist models that capture the central intuition of HOT, namely that knowledge is conscious when it is appropriately redescribed by means of metarepresentations.

4.1. Computational principles to distinguish conscious from unconscious representations

A salient point of agreement shared by most GW-like contemporary theories of consciousness is 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, all of which enable such representations to exert global influence on ongoing processing. Interestingly, Rumelhart, Smolensky, McClelland, and Hinton (1986) had already characterized consciousness as involving a trajectory through a sequence of stable states. I have proposed 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 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 (Frank, Loughry, & O'Reilly, 2001).

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 an entire processing pathway, as in the Stroop model of Cohen, Dunbar, and McClelland (1990). Likewise, in attractor networks (Mozer, 2009), strength refers to how well a system has been tuned through training to a particular representation, that is, to how easy it is for such a system to produce an appropriate response to the representation. Strong activation patterns exert more influence on ongoing processing than weak patterns.

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, McNaughton, & O'Reilly, 1995), 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 dogs all share a number of characteristic features. This latter knowledge can be made explicit when the task at hand requires it, but it is normally only conveyed implicitly (as a presupposition) by statements about or by actions directed toward dogs.

Importantly, 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" (Tononi & Edelman, 1998), or from contact with the neural workspace, brought about by "dynamic mobilization" (Dehaene & 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, in precisely the way attractor basins instantiate in dynamical connectionist networks (Mathis & Mozer, 1996).

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

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

While 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, Arnell, & Raymond, 1997), inattentional blindness (Mack & Rock, 1998) as well as some instances of change blindness (Simons & Levin, 1997), for instance, all suggest that quality of representation alone does not suffice, for even strong representations 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, an assumption that is at the core of HOT theories of consciousness (Rosenthal, 1997). This analysis thus suggests that a further important principle that differentiates between conscious and unconscious cognition is the extent to which a given representation endowed with the proper properties (stability, strength, and distinctiveness) is itself the target of meta-representations. Note that metarepresentations 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 metarepresentations that take first-order representations as their input.

Thus, 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. There are at least four possibilities. First, knowledge that is embedded in the connection weights within and between processing modules can never be directly available to conscious awareness and control. This is sim-

ply a consequence of the fact that consciousness, by assumption, necessarily involves explicit representations (i.e., patterns of activation over processing units). This is equivalent to Dehaene and Changeux (2004)'s principle of "active firing." 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. Low-quality representations (i.e., weak representations, fleeting representations, or representations that fail to be sufficiently distinct from other 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.

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 into other conscious contents.

Finally, a representation can be so strong that its influence can no longer be controlled, as is the case when a behavior has become automatic. In these cases, it is debatable whether the knowledge should be taken as genuinely unconscious, 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. In such cases, consciousness has become optional, in a way.

Strong, stable, and distinctive representations are thus *explicit* representations 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). 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.

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.

Beyond giving a cognitive system the ability to learn about its own representations, there is another important function that metarepresentations may play: They can also be used to anticipate the future occurrences of first-order representations (see Bar, 2009; on the human brain as a prediction machine). Thus for instance, if my brain learns that the Supplementary Motor Area (SMA) is systematically active before the Primary Motor Area (MA) then it can use SMA representations to explicitly represent their consequences downstream, that is, M1 activation, and ultimately, action. If neurons in SMA systematically become active before an action is carried out, a metarepresentation can link the two and represent this fact explicitly in a manner that will be experienced as intention; that is, when neurons in the SMA become active, I experience the feeling of intention *because* my brain has learned, unconsciously, that such activity in SMA precedes action. It is this knowledge that gives qualitative character to experience, for, as a result of learning, each stimulus that I see, hear, feel, or smell is now not only represented but also re-represented through independent metarepresentations that enrich and augment the original representation(s) with knowledge about (1) how similar the manner in which the stimulus's representation is with respect to that associated with other stimuli, (2) how similar the stimulus's representation is now with respect to what it was before, (3) how consistent is a stimulus's representation with what it typically is, (4) what other regions of my brain are active at the same time that the stimulus's representation is, etc. This perspective is akin to the sensorimotor perspective (O'Regan & Noë, 2001) in the sense that awareness is linked with knowledge of the consequences of our actions, but, crucially, the argument is extended inwards, that is, to the entire domain of neural representations (it can also be extended further outwards—this is what I take Theory of Mind to be—but that story is too long to tell here).

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*" (Cleeremans, 2011), 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. How can we begin to explore this line of thought using computational modeling? In the following, I present an overview of the recent work we have carried out in attempting to do just so.

5. Metacognitive networks

In a strikingly insightful article titled "The Cognizer's innards, A psychological and philosophical perspective on the development of thought," Clark and Karmiloff-Smith (1993) wrote:

> [. . .] genuine thinkers, we submit, are endowed with an internal organization which is geared to the repeated redescription of its own stored knowledge. This organization is one in which information already stored in an organism's special-purpose responses to the environment is subsequently made available, by the RR [Representation Redescription] process, to serve a much wider variety of ends. Thus knowledge that is initially embedded in special-purpose effective procedures subsequently becomes a data structure available to other parts of the system. (p. 488)

We have recently begun exploring these ideas, focusing on the following question: What kind of mechanism may enable the sort of redescription processes envisioned by Clark and Karmiloff-Smith? First, enabling redescription of one's own internal states minimally requires such internal states to be *available* to redescription, where *availability* is contingent, as discussed above, on such internal states being *patterns of activation* endowed with certain characteristics such as their strength, their stability in time, and their distinctiveness. Note that these assumptions rule out many potential sources of internal knowledge. For instance, the sort of weak, fleeting representations presumably result-

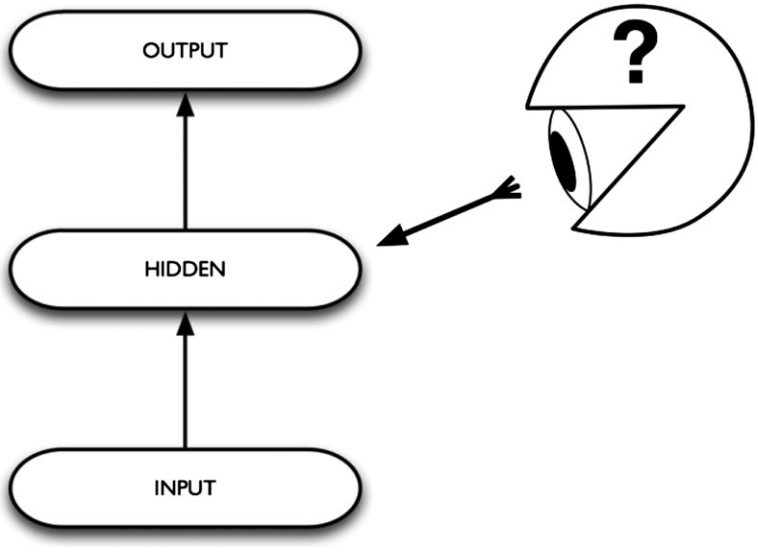

Fig. 1. What kind of mechanism would make it possible for a network to find out about and learn something about its own internal states?

ing from the presentation of a brief stimulus would be poor candidates to be available to further processing. Likewise, the associative links that exist between representations, if implemented through patterns of connectivity between groups of units, as they would be in connectionist networks, would remain inaccessible to further processing.

Second, those representations that meet the requirements for redescription need to be accessed by another part of the system whose function it is to redescribe them. This requires the existence of monitoring or observing systems, such as depicted in Fig. 1. *Something* needs to be able to observe, from the outside so to speak, the states of the system. Note that this is precisely what modelers are doing when poring over the activation patterns learned by a trained network. Here, however, we would like this process to be integrated into the *system itself*, so that the *system itself* redescribes its own activity in the service of further tasks. There are several possible approaches to this problem.

An important point worth highlighting right away is that any network that contains one or multiple layers of hidden units is in a sense already redescribing its own activity to itself: Each layer of hidden units constitutes a redescription of its inputs. But there is an important difference between redescriptions of this kind and what one could call independent redescriptions, that is, redescriptions that lie outside the causal chain that links input and output. I shall return to this difference in the discussion.

I suggest that the general form of such redescription mechanisms is something similar to what is depicted in Fig. 2. Two independent networks (the first-order network and the second-order network) are connected to each other in such a way that the entire first-order network is input to the second-order network. Both networks are simple feedforward back-propagation networks. The first-order network consists of thee pools of units: a pool of input units, a pool of hidden units, and a pool of output units. Let us further imagine that this network is trained to perform a simple discrimination task, that is, to produce what is named Type I response in the language of signal detection theory. My claim is that there is nothing in the computational principles that characterize how this network performs its task that is intrinsically associated with awareness. The network simply performs the task. While it will develop knowledge of the associations between its inputs and outputs over its hidden units, and while this knowledge may be in some cases very sophisticated, it will forever remain knowledge that is "in" the network as opposed to being knowledge "for" the network.

A trained network of this kind could thus properly be described as being *sensitive* to its inputs, but there is no sense in which it can be described as *being aware* of what it has learned. In other words, such a (first-order) network can never know *that* it knows: It simply lacks the appropriate machinery to do so. Likewise, in signal detection theory, while Type I responses always reflect sensitivity to some state of affairs, this sensitivity may or may not be conscious sensitivity; that is, a participant may be successful in discriminating one stimulus from another, yet fail to be aware *that* he is able to do so and thus claim, if asked, that he is merely guessing or responding randomly.

Enabling such metacognitive judgements about one's own performance thus appears to require the involvement of a second-order network, the task of which consists of learning about the internal states of the first-order network in such a way as to make it possible

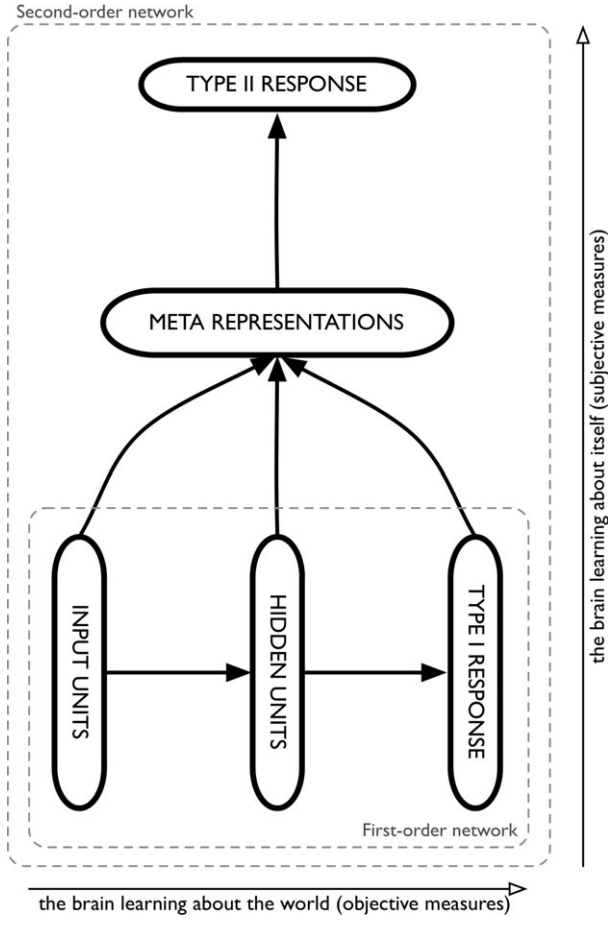

Fig. 2. General architecture of a metacognitive network. A first-order network, consisting for instance of a simple three-layer backpropagation network, has been trained to perform a simple classification task and thus contains knowledge that links inputs to outputs in such a way that the network can produce Type I responses. By design, this entire first-order network then constitutes the input to a second-order network, the task of which consists of redescribing the activity of the first-order network in some way. Here, the task that this second-order network is trained to perform is to issue Type II responses, that is, judgments about the extent to which the first-order network has performed *its* task correctly. One can think of the first-order network as instantiating cases where the brain learns about the world, and of the second-order network as instantiating cases where the brain learns about itself.

for it to make decisions about when the first-order network is correct or not. In its more general form, as depicted in Fig. 2, such an architecture would also be sufficient for the second-order network to perform other judgements, such as distinguishing between a hallucination and a veridical perception, or developing other kinds of knowledge about the overall geography of the internal representations developed by the first-order network.

Can we use such architectures to account for relevant data? That is the question we set out to answer in recent work (e.g., Cleeremans, Timmermans, & Pasquali, 2007) aimed at

exploring different facets of the overall challenge that the relationships between performance and awareness represents.

In different simulations I overview below, we have chosen to focus on exploring the relationships between performance and post-decision wagering. Post-decision wagering was introduced by Persaud, McLeod, and Cowey (2007) as a measure of awareness through which participants are required, on a trial-by-trial basis, to place a high or a low wager on their decisions, such as relative to stimulus identification for example. The intuition behind this measure is that people will place a high wager when they have conscious knowledge that their decision was correct, and a low wager when they are uncertain of their decisions. In this, wagering is thus similar to other subjective measures of awareness such as confidence judgments (Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008). According to Persaud et al., wagering provides an incentive for participants not to withhold any conscious information, as well as not to guess, making it a more objective measure of awareness than confidence judgments. Despite recent criticism of Persaud et al.'s claims (Dienes & Seth, 2010), wagering certainly reflects the extent to which an agent is sensitive to its own internal states.

Beginning with Cleeremans et al. (2007), we therefore focused on creating "wagering networks," for wagering affords easy quantification and thus appeared more readily amenable to computational simulation than other metacognitive measures such as confidence.

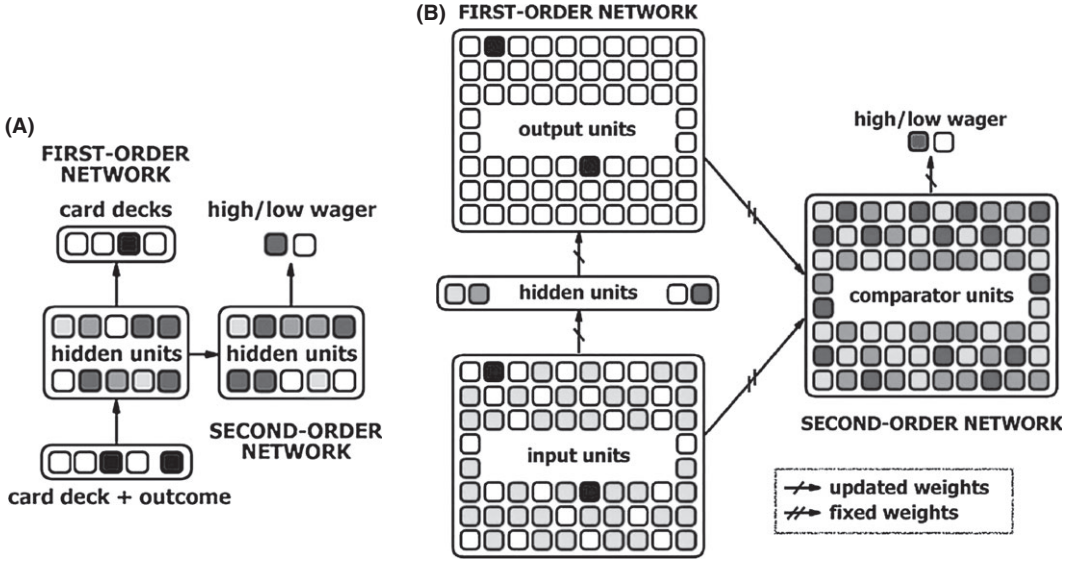

Fig. 3. Architectures for wagering networks. (A) Network architecture for the Iowa gambling task simulation (see Pasquali et al., 2010, simulation 3). The network consists of a first-order feedforward backpropagator, of which the hidden units feedforward into a set of second-order hidden units, which in turn feed forward into two wagering units. (B) Network architecture for the Blindsight and AGL simulations (see Pasquali et al., 2010, simulations 1 and 2). The network consists of a first-order feedforward backpropagation autoassociator, of which the input and output units are connected through fixed weights to a second-order comparator, which in turn feeds forward into two wagering units.

We have found that different approaches to instantiating the general principles we have described so far are required to capture empirical findings.

In one, as hinted above, the first-order and the second-order network stand in a hierarchical relationship and are thus part of the same causal chain, but they are trained on different tasks, one corresponding to first-order decisions and the second corresponding to metacognitive decisions, that is, decisions about the first-order network's performance (see Fig. 3A). Such networks are best described as hierarchical metacognitive networks, since the sensory input needs to be fully processed by the first-order network before it becomes available to the second-order network. Further, the information contained in the second-order network is directly dependent on the information contained in the first-order network in that the hidden unit patterns predict both the first-order and the second-order responses.

In a second approach (Fig. 3B), the two networks are truly independent. Here, the first-order network again consists of a simple feedforward back-propagation network, trained for instance on performing auto-association on its inputs. Unlike hierarchical models, however, here, the second-order network uses comparator units to assess the difference between first-order input and output so as to make a decision about whether the first-order network was correct or not in its decision. Thus, in such networks, the second-order network lies outside of the first-order causal chain, because the information used by the first-order network to execute its task is not the information used by the second-order network to place a high or a low wager. Thus, such networks are in principle what has been called the "dual-channel" models. Nevertheless, since both networks "plug into" the same basic knowledge (first-order performance), this type of model is effectively a hybrid between hierarchical and dual-route models. Note that in either case, our assumptions are

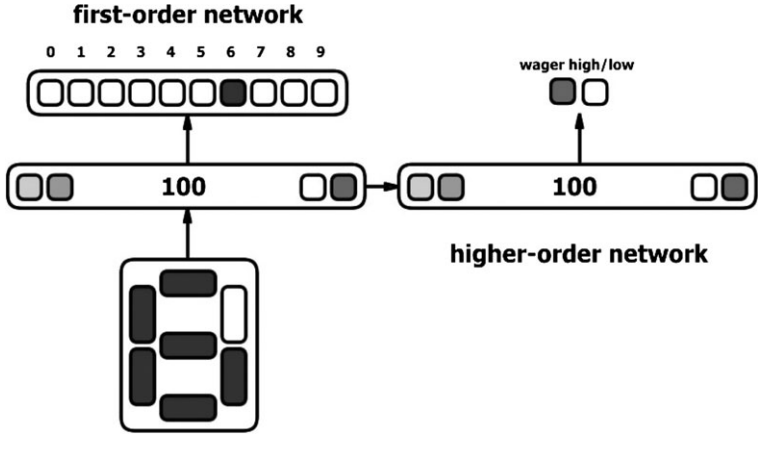

Fig. 4. Architecture for the digit classification metacognitive network. A first-order network instantiates a simple feedforward backpropagator trained to classify "visual" input patterns representing the shapes of digits 0–9 in 10 categories. A second-order network is assigned the task of wagering on the first-order network's performance based on the latter's internal representations of the stimulus. The second-order network thus performs judgements about the extent to which the first-order network is correct in its own decisions.

oversimplified, for a complete implementation of the theory would require overcoming the limitation that the second-order network cannot influence processing as it takes place in the first-order network.

In one of our first simulations, which I will describe in more detail here, the first-order feedforward backpropagation network (see Fig. 4) consisted of 7 input units representing digit shapes (as on a digital watch), 100 hidden units, and 10 output units for the 10 digits. The task of the first-order network is a simple one: It consists of identifying the "visual" representations of the digits 0–9. This is achieved by training the first-order network to respond to each input by activating one of its 10 output units. The 100 first-order hidden units connected to a different pool of 100 hidden units of the second-order feedforward network, with 2 output units representing a high and a low wager, as shown in Fig. 3.

The task of the higher order network consisted of wagering on the first-order network's decisions. It was trained to place a high wager if the first-order network had provided a correct answer (correct identification of the digit), and to wager low when the first network had given an incorrect answer (misidentification of the digit). Both networks were trained simultaneously. Importantly, this implies that the second-order network is trained on a continuously changing training set, since the first-order network patterns of activation over its hidden units are themselves changing as a result of training.

A learning rate of 0.15 and a momentum of 0.5 were used during training of the first-order network. The second-order network was simultaneously and independently trained to wager high or low on the performance of the first-order network. This was achieved simply by training the network to activate one output unit when the first-order network had produced the correct response and a second output unit when it had not, using a simple winner-take-all approach. In an attempt to explore differences in the learning regime of the second-order network, we also contrasted a condition where the second-order network was trained with a learning rate of 0.1 and a condition where it was trained with a much lower learning rate of 10^{-7}. Because this can (admittedly very crudely) be taken as reflecting different degrees of metacognitive awareness; we dubbed the first condition "high awareness" and the second "low awareness." Ten such networks were trained to perform their tasks concurrently throughout 200 epochs of training and their performance averaged. The performance of all three networks (the first-order network; the second-order network trained with a low learning rate, and the second-order network trained with a higher learning rate) is depicted in Fig. 5.

Chance level for the first-order network is 10% (there is one chance of out 10 of correctly identifying one digit among ten); it is 50% for the second-order network (one chance out of two of placing a correct wager). The figure shows that the first-order network simply gradually learns to improve its classification performance continuously until it achieves 100% correct responses at the end of training. The performance of the "high awareness" second-order network, however, exhibits a completely different pattern. Indeed, one can see that the second-order network initially performs quite well, only to show decreasing performance up until about epoch 40, at which point its performance has sagged to chance level. From epoch 40 onwards, the second-order network's perfor-

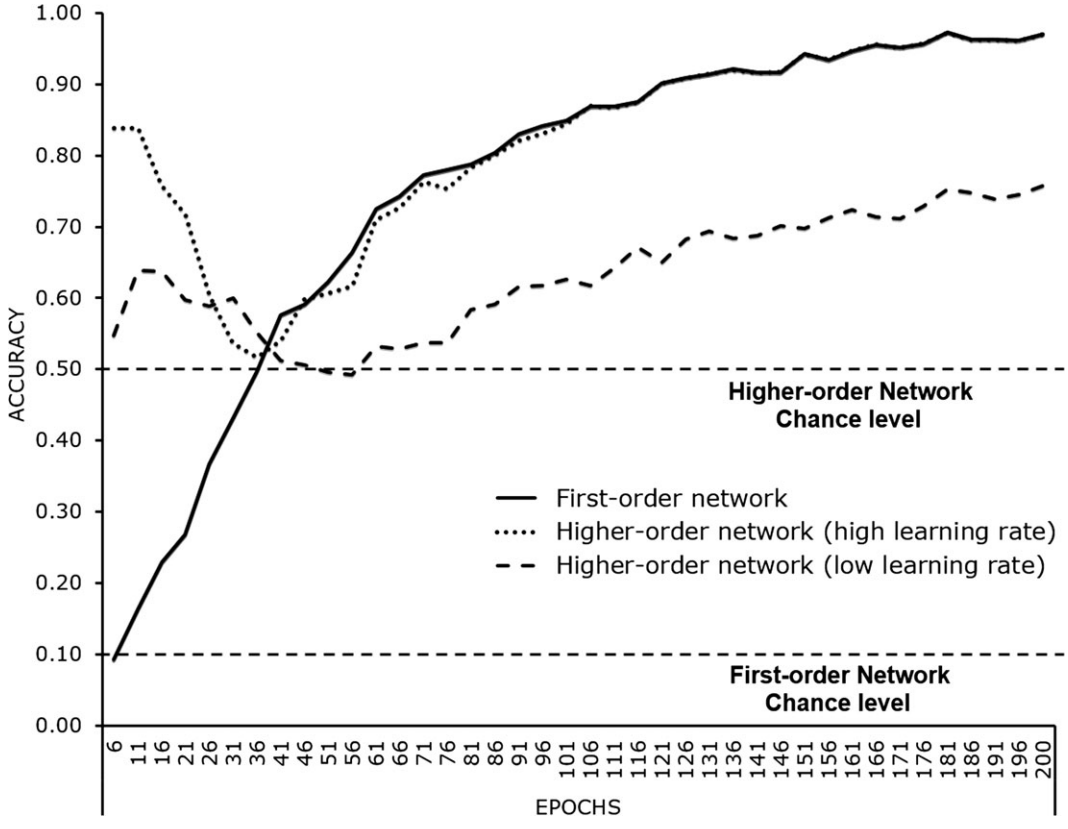

Fig. 5. Performance of the first-order and second-order networks digit-classification network, as a function of training expressed as number of epochs.

mance increases in parallel with that of the first-order network. This U-shaped performance pattern is replicated, to a lesser degree and with slightly different dynamics, in the "low awareness" second-order network.

One can understand this performance pattern as follows. Initially, the second-order network quickly learns that the first-order network is systematically incorrect in classifying the digits (which is expected since it has not begun to learn how to perform the task). The safest response (i.e., the response that minimizes error) is thus to always bet low. This, incidentally, is what any rational agent would do. However, as the first-order network quickly begins to exceed chance level performance on its digit classification task, the performance of the second-order network begins to decrease. This corresponds to a stage where the second-order network is beginning to bet "high" on some occasions as it learns to categorize states of the first-order network that are predictive of a correct classification. An interesting pattern of dissociation then occurs, for the second-order network is performing rather poorly just when the first-order network is beginning to truly master

its own digit classification task. One can think of that stage as corresponding to a point in training where the system as a whole is essentially acting based on unconscious knowledge: First-order performance on the digit classification task is well above chance level, yet, wagering by the second-order network is close to chance, and is at chance on epoch 40. Intuitively, at that point in time, the networks are performing just like a participant who is able to correctly make decisions yet claims to be guessing—precisely the dissociation one observes in many implicit learning and subliminal priming paradigms. Thus, epoch 40 corresponds to the second-order network's "most doubtful moment." One could view this as the moment at which the higher order network abandons a simple "safe" strategy of low wagers and explores the space of first-order hidden unit representations, looking for a criterion that will allow it to separate good from bad identifications.

Later on, after epoch 40, the second-order network has learned enough about when the first-order network will be correct versus incorrect to begin attempting to maximize its own wagering performance. As the two networks simultaneously learn to perform their respective tasks, one then sees the entire system shifting from a situation where there is no relationship between first- and second-order performances to a situation where the two are correlated. This transition reflects, under our assumptions, a shift between unconscious versus conscious processing.

In later work (Pasquali, Timmermans, & Cleeremans, 2010), we have explored similar models based on germane or identical architectures and shown that they are capable of accounting for the data reported by Persaud et al. (2007) in three different domains: artificial grammar learning, blindsight, and the Iowa gambling task (IGT).

In all three cases, our simulations were successful in duplicating the patterns of associations and dissociations observed in human participants with respect to the relationship between task performance and wagering. For instance, Fig. 6 shows the results of a simulation study of the IGT (Bechara, Damasio, Damasio, & Anderson, 1994), which requires participants to choose, on each trial, a card that they may select from one of four decks. Unknown to them, two of the decks are advantageous in the long term for they yield modest wins, but wins that ultimately exceed the modest losses also associated with those decks. The other two decks are initially enticing because they yield substantial wins early on, but ultimately disadvantageous for they also contain cards associated with severe losses. Because participants are initially ignorant of the reward structure of the decks each of their choices is, at least initially, ambiguous with respect to the outcome. In Persaud's adaptation of the IGT, participants additionally placed wagers on whether each card would be winning or losing. The wager is placed after deck selection, but before turning over the card (revealing how much was won or lost). Participants typically manage to improve deck selection well before they start wagering advantageously, suggesting implicit knowledge. However, when participants are made more aware of their strategy to determine deck relative pay-offs by being asked specific questions regarding their strategy such as "What would you expect your average winning amount to be by picking 10 cards from deck 1?," wagering follows performance more closely (Maia & McClelland, 2004).

In our simulation (using the architecture depicted in Fig. 3A), as in the digit task simulation, we captured the difference between "low awareness" and "high awareness"

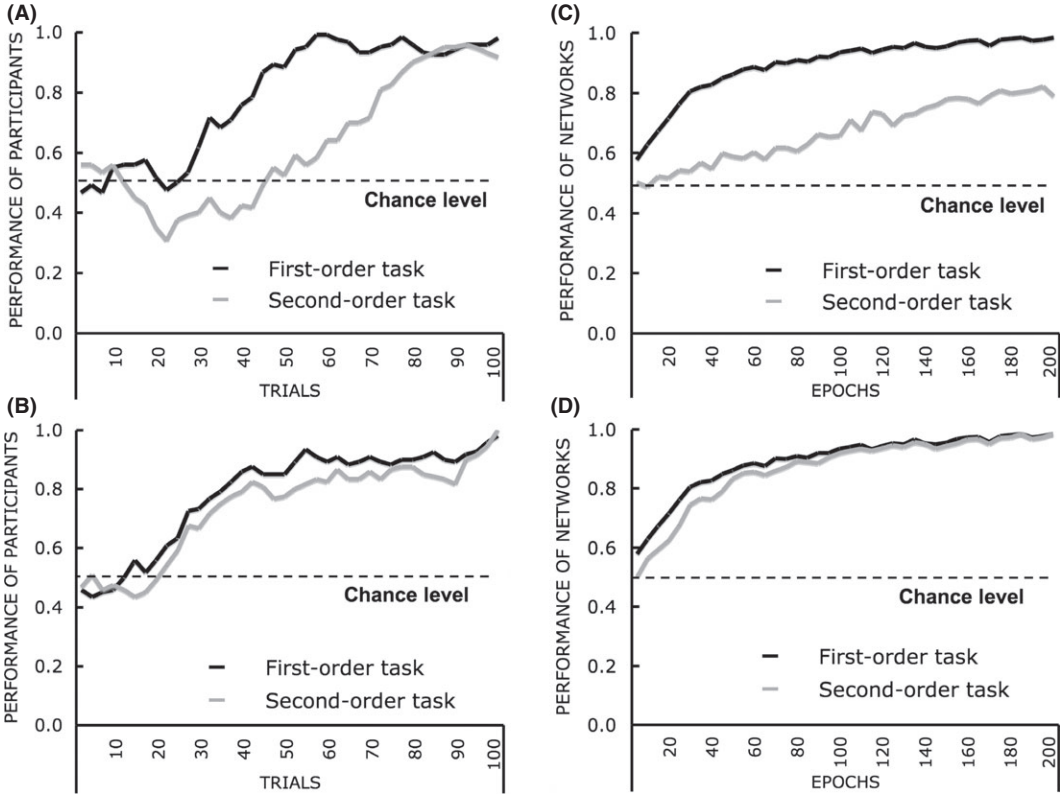

Fig. 6. Results for the Iowa gambling task simulation. Network performance is plotted across time (epochs) for (C) "low awareness" and (D) "high awareness" conditions. Persaud et al.'s results are reproduced (with permission) for comparison purposes (A and B for low and high awareness conditions, respectively).

conditions by the simple expedient of modulating the learning rate of the second-order network accordingly. The results show a good fit with the human data and are thus suggestive that differences in the extent to which the second-order is allowed to learn about the first-order information is instrumental in defining metacognitive access (and hence, awareness, by this account). This is borne out by a recent study of Fleming, Weil, Nagy, Dolan, and Rees (2010) which indicates large individual differences in people's ability to judge their own performance. Strikingly, the authors found that differences in metacognitive ability were subtended not only by differences in the activity of anterior prefrontal cortex but also by structural differences in the white matter of these regions.

It may seem that the proposed mechanisms are identical with signal-detection accounts of metacognition (e.g., Scott & Dienes, 2008). However, there is a crucial difference. Signal detection accounts typically make the second-order distinction between confidence and guessing (high vs. low wagers) on the very signal that is used for first-order classifications by setting two boundaries on the signal: one boundary that accounts for the first-order classification and a second boundary (on either side of the first-order boundary) that

distinguishes between guessing (cases that fall within the area defined by the second boundaries) and cases that fall outside of these boundaries (on the extremes of the distribution). In such an account, confidence thus depends directly on first-order signal strength. However, in the hybrid models we have also proposed (Fig. 3B), the second-order classification does not depend on the same signal as the first-order task. Indeed, instead of wagering high or low based on signal strength, the second-order network re-represents the first-order error as a new pattern of activation. Thus, before it can wager correctly, the second-order network, like the first-order network, has to learn to make a new, single-boundary classification based on this second-order representation (the error representation). Thus, the second-order network actually learns to judge the first-order network's performance independently of the first-order task itself. The difference between our model and signal detection theory is substantial, for it impinges on whether one considers Type I and Type II performance; that is, first-order and second-order judgments about these decisions entertain hierarchical or parallel relationships with each other. This issue is currently being debated, with some authors defending a dual-route model (Dehaene & Charles, 2010) and others (Lau, 2010) defending hierarchical models. The simulation work described in Pasquali et al. (2010) is suggestive that the former may be more fruitful in that they afford additional flexibility and generality.

6. Conclusions

Information processing, be it conscious or not, necessarily takes place in brains interacting with their environment. Because brains consist of large-scale, interacting neural networks, it must be the case that the difference between conscious and unconscious amounts to functional differences between the way such networks are organized. In other words, conscious and unconscious processing are fundamentally connected, that is, rooted in the very same principles of information processing—a point already made forcefully by Searle (1992). The singular challenge we are thus faced with is to understand how the symbolic representations characteristic of conscious information processing can emerge out of the subsymbolic representations characteristic of unconscious information processing. One possibility to address this challenge, outlined here, is that the brain continuously and unconsciously learns to redescribe its own activity to itself based on constant interaction with itself, with the world, and with other minds. The outcome of such interactions is the emergence of internal models that are metacognitive in nature and that function so as it make it possible for an agent to develop a (limited, implicit, practical, embodied) understanding of itself. In this light, plasticity and learning are constitutive of what makes us conscious, for it is in virtue of our own experiences with ourselves and with other people that our mental life acquires its subjective character. The connectionist framework continues to be uniquely positioned in the Cognitive Sciences to address the challenge of identifying what one could call the "computational correlates of consciousness" (Cleeremans, 2005; Mathis & Mozer, 1996), both because it makes it possible to focus

on the *mechanisms* through which information processing takes place and because its fundamental principles are inspired by the manner in which the brain computes.

Acknowledgments

Axel Cleeremans is a research director with the National Fund for Scientific Research (Belgium). This work benefited from funding under the "Interuniversity Poles of Attraction" Progam by the Belgian Science Policy Office (PAI 7/33). I thank Jay McClelland, Rick Cooper, Mike Mozer, Tim Rogers, Darren Abramson, Antoine Pasquali, and Bert Timmermans for insightful and often stimulatingly challenging comments about this article. Sections of this article were adapted from Pasquali et al. (2010) and from Timmermans, Schilbach, Pasquali, and Cleeremans (2012).

References

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Atkinson, A. P., Thomas, M. S. C., & Cleeremans, A. (2000). Consciousness: Mapping the theoretical landscape. *Trends in Cognitive Sciences*, *4*(10), 372–382.

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge, UK: Cambridge University Press.

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

Bates, E. A., & Elman, J. L. (1993). Connectionism and the study of change. In M. Johnson (Ed.), *Brain development and cognition: A reader* (pp. 623–642). Oxford, UK: Blackwell.

Bechara, A., Damasio, A., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damade to human prefrontal cortex. *Cognition*, *50*, 7–15.

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology*, *36A*, 209–231.

Bliss, T. V. P., & Lomo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology (London)*, *232*, 331–356.

Block, N. (2011). The higher-order approach to consciousness is defunct. *Analysis*, *71*(3), 419–431.

Brooks, L. R. (1978). Non-analytic concept formation and memory for instances. In E. Rosch & B. Lloyd (Eds.), *Cognition and Concepts* (pp. 16–211). Mahwah, NJ: Lawrence Erlbaum.

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

Cleeremans, A. (1997). Principles for implicit learning. In D. C. Berry (Ed.), *How implicit is implicit learning?* (pp. 195–234). Oxford, UK: Oxford University Press.

Cleeremans, A. (2005). Computational correlates of consciousness. In S. Laureys (Ed.), *Progress in brain research* (vol. 150, pp. 81–98). Amsterdam: Elsevier.

Cleeremans, A. (2008). Consciousness: The radical plasticity thesis. *Progress in Brain Research*, *168*, 19–33.

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

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: News from the front. *Trends in Cognitive Sciences*, *2*, 406–416.

Cleeremans, A., & Dienes, Z. (2008). Computational models of implicit learning. In R. Sun (Ed.), *The Cambridge handbook of computational modeling* (pp. 396–421). Cambridge, UK: Cambridge University Press.

Cleeremans, A., & Jiménez, L. (2002). Implicit learning and consciousness: A graded, dynamic perspective. In R. M. French & A. Cleeremans (Eds.), *Implicit learning and consciousness: An empirical, computational and philosophical consensus in the making?* (pp. 1–40). Hove, UK: Psychology Press.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*, 235–253.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation*, *1*, 372–381.

Cleeremans, A., Timmermans, B., & Pasquali, A. (2007). Consciousness and metarepresentation: A computational sketch. *Neural Networks*, *20*(9), 1032–1039.

Cohen, A., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*, 332–361.

Dehaene, S., & Changeux, J.-P. (2004). Neural mechanisms for access to consciousness. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (3rd ed., pp. 1145–1157). New York: W.W. Norton.

Dehaene, S., & Charles, L. (2010). A dual-route theory of evidence accumulation during conscious access. Paper presented at the 14th Annual meeting of the Association for the Scientific Study of Consciousness, Toronto, Canada

Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences of the U.S.A.*, *95* (24), 14529–14534.

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

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

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

Dienes, Z. (1992). Connectionist and memory-array models of artificial grammar learning. *Cognitive Science*, *16*, 41–79.

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

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

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating instrospective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–1543.

Fodor, J. A. (1975). *The language of thought*. New York: Harper & Row.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective & Behavioral Neuroscience*, *1*(2), 137–160.

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, *21*(5), 1–14.

Gibson, F., Fichman, M., & Plaut, D. C. (1997). Learning in dynamic decision task: Computational models and empirical evidence. *Organizational Behavior and Human Decision Processes*, *71*, 1–35.

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

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the Cognitive Science Society* (pp. 1-12), Hillsdale, NJ: Lawrence Erlbaum Associates.

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

James, W. (1890). *The principles of psychology*. New York: H. Holt and Company.

Johnson, S. (2002). *Emergence: The connected lives of ants, brains, cities, and software*. London: Penguin Books.

Kihlstrom, J. F. (1987). The cognitive unconscious. *Science*, *237*(1445–52).

Kihlstrom, J. F. (1990). The psychological unconscious. In L. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 445–464). New York: Guilford.

Kirsh, D. (1991). When is information explicitly represented? In P. P. Hanson (Ed.), *Information, language, and cognition*. New York: Oxford University Press.

Koch, C. (2004). *The quest for consciousness. A neurobiological approach*. Englewood, CO: Roberts & Company Publishers.

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

Lau, H. (2010). Comparing different signal processing architectures that support conscious reports. Paper presented at the 14th annual meeting of the Association for the Scientific Study of Consciousness, Toronto, Canada.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of consciousness. *Trends in Cognitive Sciences*, *15*(8), 365–373.

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

Maia, T. V., & Cleeremans, A. (2005). Consciousness: Converging insights from connectionist modeling and neuroscience. *Trends in Cognitive Sciences*, *9*(8), 397–404.

Maia, T. V., & McClelland, J. L. (2004). A reexamination of the evidence for the somatic marker hypothesis: What participants really know in the Iowa gambling task. *Proceedings of the National Academy of Sciences USA*, *101*(45), 16075–16080.

Mathis, W.D., & Mozer, M.C. (1996). Conscious and unconscious perception: A computational theory. In G. Cottrelll (Ed.), *Proceedings of the eighteenth annual conference of the cognitive science society* (pp. 324–328). Hillsdale, NJ: Lawrence Erlbaum Associates.

McClelland, J. L. (1981). Retrieving general and specific information from stored knowledge of specifics. *Proceedings of the Third Annual Cognitive Science Society*, *17*, 0–172.

McClelland, J. L. (2010). Emergence in cognitive science. *Topics in Cognitive Science*, *2*(4), 751–770.

McClelland, J. L., McNaughton, B. L., & 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. *Psychological Review*, *102*, 419–457.

McClelland, J.L., & Rumelhart, D.E. (1986). *Parallel distributed processing. Explorations in the microstructure of cognition. Volume 2: Psychological and biological models*. Cambridge, MA: MIT Press.

Mozer, M. C. (2009). Attractor Networks. In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 86–89). Oxford, UK: Oxford University Press.

Munakata, Y. (2001). Graded representations in behavioral dissociations. *Trends in Cognitive Sciences*, *5*(7), 309–315.

Munakata, Y., McClelland, J. L., Johnson, M. H., & Siegler, R. S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review*, *10*(4), 686–713.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, UK: Harvard University Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

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

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

Pacton, S., Perruchet, P., Fayol, M., & Cleeremans, A. (2001). Implicit learning out of the lab: The case of orthographic regularities. *Journal of Experimental Psychology: General*, *130*(3), 401–426.

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

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, *10*, 257–261.

Port, R. F., & van Gelder, T. (1995). *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, *3*, 111–169.

Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge, MA: MIT Press.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *5*, 855–863.

Rosenthal, D. (1997). A theory of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates*. Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986a). On learning the past tense of english verbs. In J. L. McClelland, & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (vol. 2, pp. 216–271), Cambridge, MA: MIT Press.

Rumelhart, D.E., & McClelland, J.L. (1986b). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (vol. 2, pp. 7–57). Cambridge, MA: MIT Press.

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

Schapiro, A. C., & McClelland, J. L. (2009). A connectionist model of a continuous developmental transition in the balance scale task. *Cognition*, *110*(1), 395–411.

Scott, R. B., & Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *34*(5), 1264–1288.

Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.

Sergent, C., & Dehaene, S. (2004). Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychological Science*, *15*(11), 720–728.

Servan-Schreiber, D., Cleeremans, A., & McClelland, J. L. (1991). Graded State Machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning*, *7*, 161–193.

Seth, A., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: Relating behavioural and neuropsychological approaches. *Trends in Cognitive Science*, *12*(8), 314–321.

Shanks, D. R., & St. John, M.F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, *17*, 367–447.

Shapiro, K. L., Arnell, K. M., & Raymond, J. E. (1997). The Attentional Blink. *Trends in Cognitive Sciences*, *1*, 291–295.

Simons, D. J., & Levin, D. T. (1997). Change Blindness. *Trends in Cognitive Sciences*, *1*, 261–267.

Timmermans, B., Schilbach, L., Pasquali, A., & Cleeremans, A. (2012). Higher order thoughts in action: Consciousness as a unconscious re-description process. *Philosophical Transcations of the Royal Society B*, *367*, 1412–1423.

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

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

Windey, B., Gevers, W., & Cleeremans, A. (2013). Subjective visibility depends on level of processing. *Cognition*, *129*, 404–409.