Available online at www.sciencedirect.com

ScienceDirect

Neural Networks ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Neural Networks

www.elsevier.com/locate/neunet

2007 Special Issue

# Consciousness and metarepresentation: A computational sketch

Axel Cleeremans*, Bert Timmermans, Antoine Pasquali

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

**Abstract**

When one is conscious of something, one is also conscious *that* one is conscious. Higher-Order Thought Theory [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] takes it that it is in virtue of the fact that one is conscious of *being conscious*, that one is conscious. Here, we ask what the computational mechanisms may be that implement this intuition. Our starting point is Clark and Karmiloff-Smith's [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] point that knowledge acquired by a connectionist network always remains "knowledge *in* the network rather than knowledge *for* the network". That is, while connectionist networks may become exquisitely sensitive to regularities contained in their input–output environment, they never exhibit the ability to access and manipulate this knowledge *as* knowledge: The knowledge can only be expressed through performing the task upon which the network was trained; it remains forever embedded in the causal pathways that developed as a result of training. To address this issue, we present simulations in which two networks interact. The states of a first-order network trained to perform a simple categorization task become input to a second-order network trained either as an encoder or on another categorization task. Thus, the second-order network "observes" the states of the first-order network and has, in the first case, to reproduce these states on its output units, and in the second case, to use the states as cues in order to solve the secondary task. This implements a limited form of metarepresentation, to the extent that the second-order network's internal representations become re-representations of the first-order network's internal states. We conclude that this mechanism provides the beginnings of a computational mechanism to account for mental attitudes, that is, an understanding by a cognitive system of the manner in which its first-order knowledge is held (belief, hope, fear, etc.). Consciousness, in this light, thus involves knowledge of the geography of one own's internal representations — a geography that is itself learned over time as a result of an agent's attributing value to the various experiences it enjoys through interaction with itself, the world, and others.

*Keywords:* Consciousness; Representation; Higher-order thought; Neural networks

As abundantly demonstrated not only by empirical evidence but also by the very fact that extremely powerful information-processing machines, namely, computers, have now become ubiquitous, information processing can undoubtedly take place without consciousness. Only but a few would be willing to grant any quantum of conscious experience to contemporary computers, yet they are undeniably capable of sophisticated information processing — from recognizing faces to analyzing speech, from winning chess tournaments to helping prove theorems. Likewise, it is hard to discern anything intrinsic to neural activity that mandates that such activity be associated to or produce conscious experience. Thus, consciousness is not information processing *tout court*; experience is an "extra ingredient" (Chalmers, 2007) that comes over and beyond mere computation.

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

One is to simply state, as per Dennett (e.g., Dennett (1991, 2001)) that there is nothing more to explain. Experience is *just* (a specific kind of) information processing in the brain; the contents of experience are *just* whatever representations have come to dominate processing at some point in time ("fame in the brain"); consciousness is *just* a harmless

* Corresponding author. Tel.: +32 2 650 32 96; fax: +32 2 650 22 09.
*E-mail address:* axcleer@ulb.ac.be (A. Cleeremans).

0893-6080/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2007.09.011

illusion. From this perspective, it is easy to imagine that machines will be conscious when they have accrued sufficient complexity; the reason they are not conscious now is simply because they are not sophisticated enough: They lack the appropriate architecture perhaps, they lack sufficiently broad and diverse information processing abilities, and so on. Regardless of what is missing, the basic point here is that, *contra* Chalmers, there is no reason to assume that conscious experience is anything special. Instead, all that is required is one or several yet-to-be-identified functional mechanisms: Recurrence, perhaps (Lamme, 2003), stability of representation (O'Brien & Opie, 1999), global availability (Baars, 1988; Dehaene, Kerszberg, & Changeux, 1998), integration and differentiation of information (Tononi, 2003), or the involvement of higher-order representations (Rosenthal, 1997), to name just a few (see Atkinson, Thomas, and Cleeremans (2000), Maia and Cleeremans (2005), for reviews).

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

Both Chalmers' (somewhat paradoxically) and Tononi's analyses, however, describe conscious experience as a rather abstract dimension or aspect of information, whereas our intuition is that *what it feels like* is anything but abstract. On the contrary, what we mean when we say that seeing a patch of red elicits an "experience" is that the seeing *does something to us* — in particular, we might feel one or several emotions, and we may associate the redness with memories of red. Perhaps seeing the patch of red makes you remember the color of the dress that your prom night date wore 20 years ago. Perhaps it evokes a vague anxiety, which we now know is also shared by monkeys (Humphrey, 1971). To a synaesthete, perhaps seeing the color red will evoke the number 5. The point is that if conscious experience is what it feels like to be in a certain state, then "What it feels like" can only mean the specific set of associations that have been established by experience between the stimulus or the situation you now find yourself in, on the one hand, and your memories, on the other. This is what one means by saying that there is something it is like to be you in this state rather than nobody or somebody else: The set of memories evoked by the stimulus (or by actions you perform, etc.), and, crucially, the set of emotional states associated with each of these memories. It is interesting to note that Indian philosophical traditions have placed similar emphasis on the role that emotion plays in shaping conscious experience (Banerjee, 2007).

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

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

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

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

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

In this context, what would it take for a network like Hinton's to be able to access its own representations; and what difference would that make with respect to consciousness?

To answer the first question, the required machinery is the machinery of agenthood; in a nutshell, the ability to do something not just with external states of affairs, but rather with one own's representations of such external states. This crucially requires that the agent be able to access, inspect, and otherwise manipulate its own representations, and this in turn, I surmise, requires mechanisms that make it possible for an agent to redescribe its own representations to itself. The outcome of this continuous "representational redescription" (Karmiloff-Smith, 1992) process is that the agent ends up knowing something about the geography of its own internal states: It has, in effect, *learned* about its own representations. Minimally, this could be achieved rather simply, for instance by having another network take both the input (i.e., the external stimulus as represented proximally) to the first-order network and its internal representations of that stimulus as inputs themselves and do something with them.

One elementary thing the system consisting of the two interconnected networks (the first-order, observed network and the second-order, observing network) would now be able to do is to make decisions, for instance, about the extent to which an external input to the first-order network elicits a familiar pattern of activation over its hidden units or not. This would in turn enable the system to come up with judgments about the performance of the first-order network (Dienes, 2007; Persaud, McLeod, & Cowey, 2007). This is just what we propose below in a preliminary set of simulations.

To address the second question (what difference would representational redescription make in terms of consciousness), if you think this is starting to sound like a higher-order thought theory of consciousness (Rosenthal, 1997), you may be right: Higher-order representations (which we will call metarepresentations in what follows) play a crucial role in consciousness.

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

## 1. The radical plasticity thesis

We 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, others, and, crucially, with itself. We call this claim the "*Radical Plasticity Thesis*", for its core is the notion that learning is what makes us conscious. How so? The short answer, as hinted above, is that consciousness involves not only knowledge about the world, but, crucially, knowledge about our own internal states, or mental representations.

In the following, we describe some preliminary simulation work aimed at capturing these intuitions about the possible role that metarepresentations may play in shaping consciousness.

Please cite this article in press as: Cleeremans, A., et al. Consciousness and metarepresentation: A computational sketch. Neural Networks (2007), doi:10.1016/j.neunet.2007.09.011

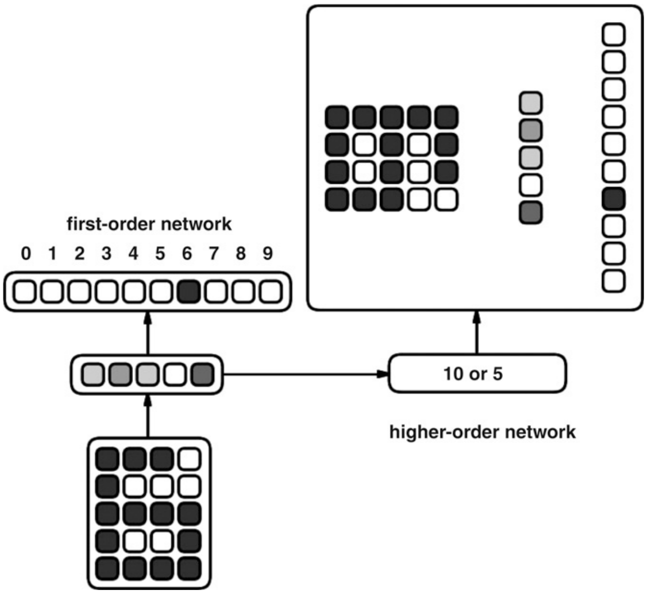

Fig. 1. Architecture of the first network, in which the higher-order network serves as an encoder of the hidden unit patterns of the first-order network.

## 2. Simulations: The digits problem

We illustrate two ways in which metarepresentations can be operationalized and what this might teach us about consciousness. Both simulations involve a first-order network that has to perform a simple task such as digit recognition, and a higher-order network that "observes" the internal states of the first-order network. This second network is thus wholly independent from the causal chain set up by the first-order network.

In the first simulation the higher-order network is simply trained to act as an encoder of the first-order internal states. It learns to reproduce the state of the entire first-order network based on that network's hidden unit patterns.

In the second simulation the higher-order network is given the more complex task of evaluating the first-order network's performance by wagering. In other words, it has to distinguish between "correct" and "wrong" internal states of the first-order network.

### 2.1. Higher-order encoding of first-order internal states

For the first simulations, we constructed a first-order feedforward backpropagation network consisting of 20 input units representing digit shapes, 5 hidden units, and 10 output units representing the 10 digits. Immediately following each presentation the hidden unit activation pattern was copied onto the 5 input units of the higher-order feedforward network, connected to either 10 or 5 hidden units, in turn connected to the 35 output units that corresponded to the number of units in the entire first-order network, as shown in Fig. 1.

One epoch consisted of presentation of all 10 digits. For each of both architectures (higher-order network with 10 or 5 hidden units, identical first-order networks) we trained 5 networks over 1000 epochs with a learning rate of .1 and a momentum of .9, measuring the error proportion (defined, for a particular epoch of training, as the magnitude of RMS output error for that epoch divided by the maximum output error, i.e. output error prior to any training) separately across the output units of the first-order and the higher-order network. Results are shown in Fig. 2 and show comparable learning curves for both architectures.

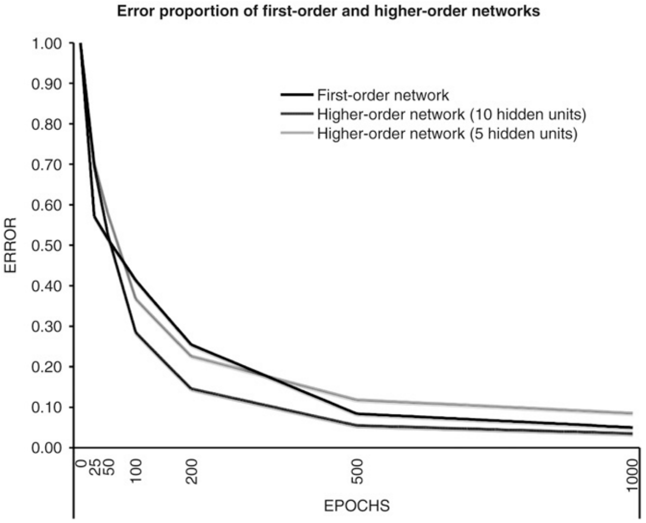

Fig. 2. Error proportion (see text for details) for the first-order network and for both higher-order networks (10 and 5 hidden units).

We can see that initially the first-order network learns at a faster rate than the higher-order network. However, after 50–100 epochs the higher-order network becomes actually better at predicting the entire state of the first-order network based on its hidden unit patterns than the first-order network is at predicting the correct digit from the input pattern. This difference decreases gradually, and for the higher-order network with 5 hidden units we can see that eventually the first-order network again outperforms the higher-order network. This suggests that as soon as some activation stability is achieved in the first-order network's hidden units, these patterns, even though they do not yet permit the first-order network itself to optimize its performance beyond an error proportion of .40, become available to a higher-order network that is able to extract from these hidden units information about the overall state of the first-order network, — information that is in itself not available to that first-order network.

In terms of awareness, this would mean that at some point during the early stages of learning, some aspects of the learned knowledge become available as targets of higher-order representations. In other words, whereas initially unstable first-order knowledge makes it impossible for the higher-order network to consistently learn about them, this changes with training in such a manner that once first-order representations have become sufficiently stable, the higher-order network can then use the structure that they contain so as to improve its own ability to reconstruct the input and the output of the first-order network successfully.

In the next simulation study, we will explore how a higher-order network can make use of this capacity to re-represent

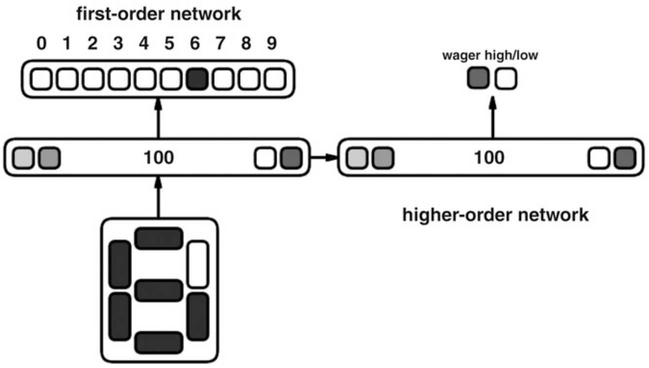

Fig. 3. Architecture of the second network, in which the higher-order networks classifies the hidden unit patterns of the first-order network.

first-order internal states so as to perform a different task, namely, evaluating the performance of the first-order network.

### 2.2. Higher-order classification of first-order internal states: A wagering network

Recently, Persaud et al. (2007) introduced wagering as a measure of awareness, where participants are required to place a high or a low wager on their decision, such as relative to stimulus identification for example. The intuition behind this measure is that people will place a high wager when they have conscious knowledge about the reasons for their decisions, and a low wager when they are uncertain of their decisions. In this, wagering is thus similar to other subjective measures of awareness (Dienes, 2004; Gaillard, Vandenberghe, Destrebecqz, & Cleeremans, 2006). 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 judgment. Despite recent criticism of Persaud et al. 's claims (Seth, 2007), wagering certainly reflects the extent to which an agent is sensitive to its own internal states. This may perhaps be captured by training a higher-order network to use first-order information so as to evaluate the performance of the latter. We therefore aimed at creating a wagering network. For this simulation, the first-order feedforward backpropagation network 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 100 first-order hidden units connected to a different pool of 100 hidden units of a higher-order feedforward network, with 2 output units representing a high and a low wager, as shown in Fig. 3.

A learning rate of .15 and a momentum of .5 were used during training of the first-order network. However, in a first condition of *high awareness*, the second network was trained with a learning rate of .1, and in a second condition of *low awareness*, a learning rate of $10^{-7}$ was applied. The task of the higher-order network consisted of wagering high if it "thought" that the first-order network was providing a correct answer (correct identification of the digit), and to wager low in case the first network gave a wrong answer (misidentification of the digit). Fig. 4 displays the average error curves of 10 networks throughout 200 epochs of training.

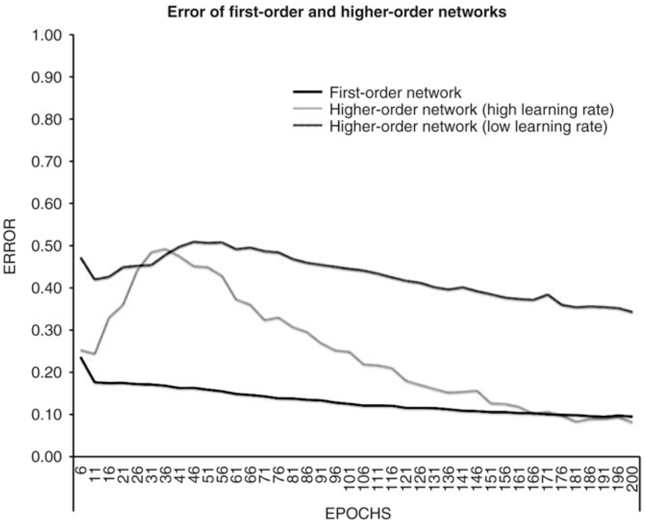

Fig. 4. Error proportion (see text for details) for the first-order network and for both higher-order networks ("high and low consciousness", reflected by using .1 and $10^{-7}$ learning rates).

Despite the gradual learning exhibited by the first-order network, error in wagering increases during the 40 first epochs in both conditions of high and low awareness. Only from the 40th epoch onwards does the higher-order network start to improve the quality of its wagering.

In order to understand the reason for this initial increase, we need to evaluate the networks' performance through an analysis of the recognition rate for the first-order network, and of the wagering strategy for the higher-order network. The first-order network's performance is represented by the percentage of correct identification (the chance level is at .1 since 10 digits are available). Wagering strategy is considered good if the network wagered high in case of correct identification and low in case of misrepresentation. Conversely, the strategy is considered to be poor if a high wager accompanies a incorrect classification, or when a correct identification was only associated to a low wager. As the strategy has the same probability of being good or bad, the chance level is at 50%. The results of this analysis are shown in Fig. 5.

As shown in Fig. 5, the previously identified error extremum at the 40th epoch corresponds in fact to a chance level wagering performance. Further analysis revealed that the higher-order networks mainly used a low-wagering strategy during the first epochs, during which the first-order network is misclassifying most of the digits, whereas is develops a high-wagering strategy at a later stage in learning, when first-order identification becomes progressively more accurate. Thus the error extremum observed in Fig. 5 characterizes the higher-order network's "most doubtful moment" when identification is correct only 50% of the time and no strategy can be applied. 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 or a categorization that will allow it to separate correct identifications from wrong identifications.

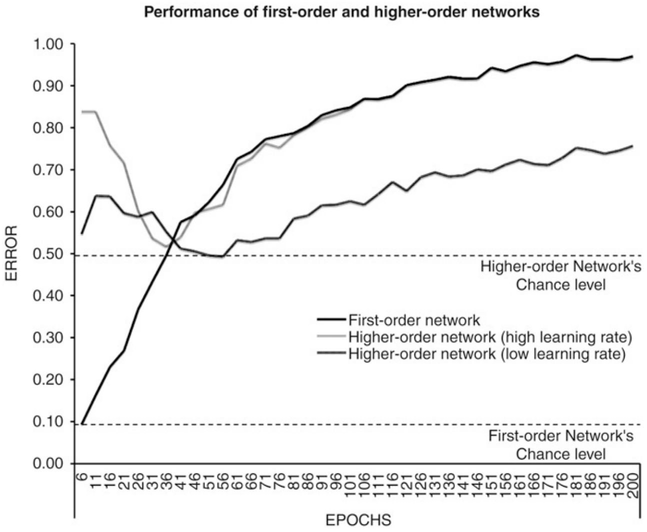

Fig. 5. Performance of the first-order network in terms of correct identifications, and of the higher-order networks in terms of advantageous wagers (high wagers when the first-order network is correct and low wagers when it is wrong).

## 3. Metarepresentation

The simulations sketched above illustrate how a network can be trained to observe the internal states of another network in such a manner that it can use this information to perform tasks that require knowledge of the structure of these internal states — either to reconstruct the corresponding inputs and outputs, or to actually evaluate the extent to which these internal representations will result in successful performance. In both cases, it is interesting to note that while the higher-order, observing network initially performs poorly, it quickly learns enough about the structure of the first-order internal representations to become more accurate in performing its own task. This captures the processes involved in the development of expertise, whereby learning might initially take place in an essentially implicit manner, and be subsequently followed by a period where explicit knowledge becomes available (Bechara, Damasio, Tranel, & Damasio, 1997; Bierman, Destrebecqz, & Cleeremans, 2005; Cleeremans, 2005, 2006; Cleeremans, Destrebecqz, & Boyer, 1998). Automaticity (Shiffrin & Schneider, 1977) would correspond to a third period in skill learning where the acquired metarepresentations become optional or otherwise detached from first-order representations.

What are the conditions under which metarepresentations emerge? Strong, stable, and distinctive representations as they occur in trained neural networks are *explicit* representations, at least in the sense put forward by Koch (2004): They indicate what they stand for in such a manner that their reference can be retrieved directly through processes involving low computational complexity (see also Kirsh (1991, 2003)). Conscious representations, in this sense, are explicit representations that have come to play, through processes of learning, adaptation, and evolution, the functional role of denoting a particular content for a cognitive system.

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, which here takes the form of an "observer" network, 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, our claim here is that such metarepresentations are learned in just the same way as first-order representations, that is, in virtue of continuously operating learning mechanisms. Because metarepresentations are also representations, the same principles that make first-order representations explicit 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. In the simulations we have described, metarepresentations as they occur in the second-order network take the more limited role of indicating relationships between internal representations and the input–output representations.

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

## 4. Conclusion

Thus we end with the following idea, which is the heart of the "Radical Plasticity Thesis": The brain continuously and unconsciously learns not only about the external world, but about its own representations of it. The result of this unconscious learning is conscious experience, in virtue of the fact that each representational state is now accompanied by (unconsciously learnt) metarepresentations that convey the mental attitude with which these first-order representations are held. Thus, from this perspective, there is nothing intrinsic to neural activity, or to information per se, that makes it conscious. Conscious experience involves specific mechanisms through which particular (i.e., stable, strong, and distinctive) unconscious neural states become the target of further processing, which we surmise involves some form of representational redescription in the sense described by Karmiloff-Smith (1992). These ideas are congruent both with higher-order theories in general (Dienes, 2007; Dienes & Perner, 1999; Rosenthal, 1997), but also with those of Lau (2007), who characterizes consciousness as "signal detection on the mind". The simulation work we have sketched here is a first step in implementing these ideas in the form of actual computational principles.

## Acknowledgements

A.C. is a Research Director with the National Fund for Scientific Research (F.R.S.-FNRS, Belgium). B.T. is a postdoctoral fellow at the ULB — Université Libre de Bruxelles. A.P. is a Research Fellow of the Fonds National de la Recherche Scientifique (F.R.S-FNRS, Belgium). This work was supported by an institutional grant from the Université Libre de Bruxelles to A.C., by Concerted Research Action 06/11-342 titled "Culturally modified organisms: What it means to be human in the age of culture", financed by the Ministère de la Communauté Française – Direction Générale l'Enseignement non obligatoire et de la Recherche scientifique (Belgium), by European Commission Grant #043457 "Mindbridge – Measuring Consciousness", and by F.R.F.C / E.S.F. Grant #2.4577.06 "Mechanisms of serial action". Portions of this article were adapted from the following publication: Cleeremans, A. (in press), The radical plasticity thesis, in R. Banerjee & B. Chakrabarti (Eds.), Models of Brain and Mind: Physical, Computational and Psychological Approaches; Progress in Brain Research. Amsterdam: Elsevier.

## References

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: Cambridge University Press.

Banerjee, R. (2007). Buddha and the bridging relations. In R. Banerjee, & B. Chakrabarti (Eds.), *Progress in brain research*, *Models of brain and mind: Physical, computational and psychological approaches*. Amsterdam: Elsevier.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, *275*(5304), 1293–1295.

Bierman, D., Destrebecqz, A., & Cleeremans, A. (2005). Intuitive decision making in complex situations: Somatic markers in an artificial grammar learning task. *Cognitive, Affective & Behavioral Neuroscience*, *5*(3), 297–305.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

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

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

Cleeremans, A. (2006). Conscious and unconscious cognition: A graded, dynamic perspective. In Q. Jing, M. R. Rosenzweig, G. d'Ydewalle, H. Zhang, H. -C. Chen, & C. Zhang (Eds.), *Progress in psychological science around the world*: *Vol. 1. Neural, cognitive, and developmental issues* (pp. 401–418). Hove, UK: Psychology Press.

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

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 United States of America*, *95*(24), 14529–14534.

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. (2004). Assumptions of subjective measures of unconscious mental states: Higher order thoughts and bias. *Journal of Consciousness Studies*, *11*(9), 25–45.

Dienes, Z. (2007). Subjective measures of unconscious knowledge. In R. Banerjee, & B. Chakrabarti (Eds.), *Progress in brain research*, *Models of brain and mind: Physical, computational and psychological approaches*. Amsterdam: Elsevier.

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

Gaillard, V., Vandenberghe, M., Destrebecqz, A., & Cleeremans, A. (2006). Third- and first-person approaches in implicit learning research. *Consciousness and Cognition*, *15*, 709–722.

Hinton. (1986). Learning distributed representations of concepts. Paper presented at the 8th annual conference of the cognitive science society.

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

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

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge: MIT Press.

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

Kirsh, D. (2003). Implicit and explicit representation. In L. Nadel (Ed.), *Encyclopedia of cognitive science*: *Vol. 2* (pp. 478–481). London, UK: Macmillan.

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

Kreiman, G., Fried, I., & Koch, C. (2002). Single-neuron correlates of subjective vision in the human medial temporal lobe. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 8378–8383.

Please cite this article in press as: Cleeremans, A., et al. Consciousness and metarepresentation: A computational sketch. Neural Networks (2007), doi:10.1016/j.neunet.2007.09.011

Lamme, V. A. F. (2003). Why visual attention and awareness are different. *Trends in Cognitive Sciences*, *7*(1), 12–18.

Lau, H. (2007). A higher-order Bayesian decision theory of consciousness. In R. Banerjee, & B. Chakrabarti (Eds.), *Progress in brain research*, *Models of brain and mind: Physical, computational and psychological approaches*. Amsterdam: Elsevier.

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

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

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

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

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.

Seth, A. K. (2007). Post-decision wagering measures metacognitive content, not sensory consciousness. *Consciousness and Cognition*. doi:10.1016/j.concog.2007.05.008.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, *84*, 127–190.

Tononi, G. (2003). Consciousness differentiated and integrated. In A. Cleeremans (Ed.), *The unity of consciousness: Binding, integration, and dissociation* (pp. 253–265). Oxford, UK: Oxford University Press.

Tononi, G. (2007). The information integration theory. In M. Velmans, & S. Schneider (Eds.), *The Blackwell companion to consciousness* (pp. 287–299). Oxford, UK: Blackwell Publishing.

Please cite this article in press as: Cleeremans, A., et al. Consciousness and metarepresentation: A computational sketch. Neural Networks (2007), doi:10.1016/j.neunet.2007.09.011