Cognition 117 (2010) 182–190

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Know thyself: Metacognitive networks and measures of consciousness

Antoine Pasquali [a,c,*,1], Bert Timmermans [b,1], Axel Cleeremans [a,1]

[a] Consciousness, Cognition, and Computation Group, Université Libre de Bruxelles, 1050 Bruxelles, Belgium
[b] Neuroimaging Group, Department of Psychiatry, University Hospital of Cologne, 50937 Köln, Germany
[c] Neurogenics Research Unit, Adam Neurogenics, 20240 Solaro, France

ARTICLE INFO

Article history:
Received 15 April 2009
Revised 5 July 2010
Accepted 11 August 2010

Keywords:
Awareness measures
Metacognitive networks
Wagering

ABSTRACT

Subjective measures of awareness rest on the assumption that conscious knowledge is knowledge that participants know they possess. Post-Decision Wagering (PDW), recently proposed as a new measure of awareness, requires participants to place a high or a low wager on their decisions. Whereas advantageous wagering indicates awareness of the knowledge on which the decisions are based, cases in which participants fail to optimize their wagers suggest performance without awareness. Here, we hypothesize that wagering and other subjective measures of awareness reflect metacognitive capacities subtended by self-developed metarepresentations that inform an agent about its own internal states. To support this idea, we present three simulations in which neural networks learn to wager on their own responses. The simulations illustrate essential properties that are required for such metarepresentations to influence PDW as a measure of awareness. Results demonstrate a good fit to human data. We discuss the implications of this modeling work for our understanding of consciousness and its measures.

1. Introduction

Awareness can be assessed by exploring patterns of association and dissociation between objective (i.e., recognition or discrimination performance) and subjective (i.e., verbal reports and confidence judgments) measures (e.g., Merikle, 1992). Thus, given certain assumptions about sensitivity and exhaustiveness, one can conclude that performance on some task of interest (the objective measure) is guided by unconscious knowledge whenever participants claim to be guessing (the subjective measure) while nevertheless performing better than chance (i.e., the "guessing criterion", e.g., Dienes, Altmann, Kwan, & Goode, 1995). Conversely, whenever we observe a correlation between subjective and objective measures, we can conclude that task performance is at least to some degree subtended by conscious knowledge (i.e., the "zero correlation criterion", e.g., Dienes et al., 1995).

This reasoning, while it remains somewhat controversial (e.g., Holender & Duscherer, 2004; Tunney & Shanks, 2003), subtends all subjective measures of awareness. Its central assumption is that when one is conscious of some piece of information, one is also conscious that one holds this information. Unconscious information can thus be taken to be knowledge about which we have no metaknowledge. There is, however, continuing debate about the extent to which available measures of such metaknowledge are sufficiently sensitive and exhaustive (e.g., Dienes & Seth, 2009; Sandberg, Timmermans, Overgaard, & Cleeremans, 2010).

In this context, Persaud et al. (e.g., Persaud & McLeod, 2007; Persaud, McLeod, & Cowey, 2007) recently reintroduced (i.e., after Ruffman, Rustin, Garnham, & Parkin, 2001) a measure of awareness that aims to be as exhaustive as possible: Post-Decision Wagering (PDW). Through PDW, participants continuously evaluate their performance by

* Corresponding author at: Consciousness, Cognition, and Computation Group, Université Libre de Bruxelles CP 191, Av. F.-D. Roosevelt, 50, 1050 Bruxelles, Belgium. Tel.: +322 650 32 96; fax: +322 650 22 09.
E-mail address: antoine.pasquali@ulb.ac.be (A. Pasquali).
[1] Shared first-authorship.

0010-0277/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2010.08.010

wagering on each decision in tasks such as visual stimulus identification under blindsight (e.g., Stoerig, Zontanou, & Cowey, 2002), string classification in Artificial Grammar Learning (AGL) (e.g., Reber, 1967), or deck selection in the Iowa Gambling task (e.g., Bechara, Damasio, Damasio, & Anderson, 1994). Like other subjective measures, PDW relies on the following logic: Given that participants attempt to maximize their earnings, if participants are aware of the knowledge upon which they make their decisions, then they should wager advantageously, betting high whenever they make a correct decision, and low for errors. Conversely, when they fail to place advantageous wagers, that is, when wagering is independent from above-chance performance, one may conclude that the knowledge that drives performance is unconscious (e.g., Seth, 2008). One purported advantage of PDW over other subjective measures such as confidence judgments is that it is more intuitive and therefore less fraught with the measurement errors possibly associated with the latter (e.g., Koch & Preuschoff, 2007). In other words, PDW is assumed to provide a better, more intuitive and also more easily quantifiable assessment of metaknowledge. This is why Persaud et al. have described PDW as an "objective" measure of awareness (but see Dienes & Seth, 2009, for a critical evaluation).

In this paper, we ask what computational principles are required for the occurrence of metaknowledge as measured by PDW and germane subjective tests of awareness that likewise depend on metacognitive access. To explore this issue, we propose a series of three simulations that capture the results obtained by Persaud et al. in the three experimental paradigms they examined (blindsight, Artificial Grammar Learning, and the Iowa Gambling task). Each simulation is based on the following set of central assumptions.

Our *first core assumption* (see also Cleeremans, 2008; Cleeremans, Timmermans, & Pasquali, 2007) is that evaluating one's own performance, as involved in subjective measures of awareness, requires that the first-order representations that are responsible for performance be accessed in a manner that is independent from their expression in behavior. To see this point, consider any neural network that has learned a particular task,—say, a Simple Recurrent Network (SRN, see Elman, 1990) predicting the next element of a sequence. Over training, such networks learn richly structured internal representations of their domain, as demonstrated for instance by Cleeremans, Servan-Schreiber, and McClelland (1989), who showed that the internal representations learned by an SRN trained to predict the next element of sequences generated by a finite-state automaton (FSA) reflect the abstract structure of the FSA. The network, however, and this is the crucial point we wish to make here, does not know, in any sense, that it possesses this knowledge. Its sensitivity to sequential structure can only be expressed in the context of the prediction task it was trained to perform. To enable such a network to "know that it knows", as would be the case for knowledge held consciously by a human agent, the knowledge that is "in the system" must therefore become knowledge "for the system" (e.g., Clark & Karmiloff-Smith, 1993; Karmiloff-Smith, 1992; Mandler, 2004).

However, as pointed out by Dienes and Perner (1996), representing knowledge into metarepresentations (i.e., into content-explicit representations) is not sufficient. One must also represent oneself as being in possession of that content (attitude-explicit representations). Our *second core assumption* is thus that such attitude-explicit representation requires access to the relevant first-order knowledge in a manner that is independent from the causal chain in which it is embedded, such that not only the content but also the accuracy of the knowledge be represented.

To achieve such a mechanism, we assume that a higher-order network automatically and continuously monitors the performance of a first-order network, in such a way that it is able: (a) to discriminate and classify information contained in the first-order network in an independent manner (i.e., independently of the first-order task), and (b) to provide access to this information (i.e., the information must indicate if the first-order network can be trusted or not) towards any secondary task requiring knowledge about the first-order internal states or performance. Without claiming that this mechanism would present sufficient or even necessary conditions for awareness in general, we intend to show that it can stand as a basic principle of metacognition, and therefore of subjectivity.

2. Simulations

We implemented three networks that simulate the performance of Persaud et al.'s participants in a blindsight situation, in the AGL task, and in the Iowa Gambling task. All simulations, rather than merely fitting Persaud et al.'s data, were designed so as to illustrate the main theoretical assumptions discussed earlier while modeling additional principles based on specific features of each of Persaud et al.'s experiments (all of which are briefly discussed in each simulation's results section).

2.1. Architectures

All three metacognitive networks each consist of two interconnected networks: (a) a three-layers backpropagation feedforward first-order network that performs the main task (i.e., stimulus discrimination, letter string classification, or deck selection, respectively), and (b) a second-order network that continuously evaluates the performance of the first-order network and that consists of a hidden unit layer and of two output nodes representing high or low wagers. In every first-order and second-order network of all simulations, we used a winner-take-all algorithm (automatic selection of the most activated units) in the output (all output activations ranging from 0.0 to 1.0) so as to avoid having to set an arbitrary goodness-of-fit criterion on the output error in order to obtain the networks' responses. Because of different task requirements, the specific network architectures associated with each of the three paradigms differ substantially from each other in terms of: (a) the nature of their higher-order representations, and (b) the implementation of "low" and "high" awareness conditions (a complete description of network architecture, parameters, and patterns is available in Supplementary material).

The fairly similar blindsight and AGL architectures (Fig. 1) have in common that the first-order network is an autoassociator, complemented by a winner-take-all mechanism on the output. The autoassociator has to solve the task problem through a re-representation of the input, that is, it has to create a bimodal distribution of inputs, representing either the presence versus absence of a stimulus (Simulation 1, in which the winner-take-all algorithm selects the most activated output unit), or grammatical versus ungrammatical strings (Simulation 2, in which the winner-take-all algorithm selects the most activated unit for each column, thus representing the most activated letter for each position within a string). This reflects participants' behavioral performance and is what in principle any non-conscious learning mechanism can do: create a simple distribution representing the environment.

The second-order network's hidden units consist of a comparator matrix, representing the match between inputs and outputs of the first-order network's autoassociator, which is effectively the accuracy of the first-order network's internal knowledge. The crucial role of such comparators for the emergence of conscious percepts has been suggested previously (e.g., Frith, Blakemore, & Wolpert, 2000; Gallagher, 2004; Mandler, 2004; Synofzik, Vosgerau, & Newen, 2008), in that they merge internal and external states into unique representations (e.g., Pacherie, 2008; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996; Sperry, 1950; Wolpert & Kawato, 1998). In our networks, each of these comparator units compute the difference between each corresponding pair of first-order input and output units of the autoassociator. Thus they represent the first-order network's error not as a training signal but as a distributed activation pattern, which the second-order network can then access by using a weighted sum of these signed errors to decide on whether to place a high or a low wager. Thus, the second-order network is taught to wager high when the first-order network's response is correct (i.e., the stimulus is present/regular and is accepted as such by the winner-take-all, or the stimulus is absent/irregular and is rejected) and to wager low when the first-order network's response is incorrect (i.e., the stimulus is present/regular but is rejected, or the stimulus is absent/irregular but is accepted). This effectively comes down to setting a decision criterion on the first-order network's error distribution. Crucially, and in contrast with the Iowa Gambling task's simulation: (a) development of the comparison patterns is automatic and unsupervised (that is, not driven by feedback), and (b) the second-order network's access to these patterns for wagering is learned in a pre-training phase, independent of specific first-order patterns or training and testing tasks. These two properties allow for the second-order network to access the relevant first-order knowledge in a manner that is independent from the causal chain in which that knowledge is embedded. Both the unsupervised emergence of the comparison patterns as well as the pre-training of the wagering skill are essential to the second-order network, for they guarantee that the second-order network does not simply learn to match specific first-order inputs and outputs to a high or a low wager. Instead, it learns here to discriminate between cases when the metacognitive network knows that it does or does not know (high wager) and when it does not know if it does or does not know (low wager). Just as people do not have to inspect their behavior in order to decide whether they will place a bet on or be confident about something, they can just do it by judging the accuracy of their internal knowledge, independently of the task to which that judgment pertains.

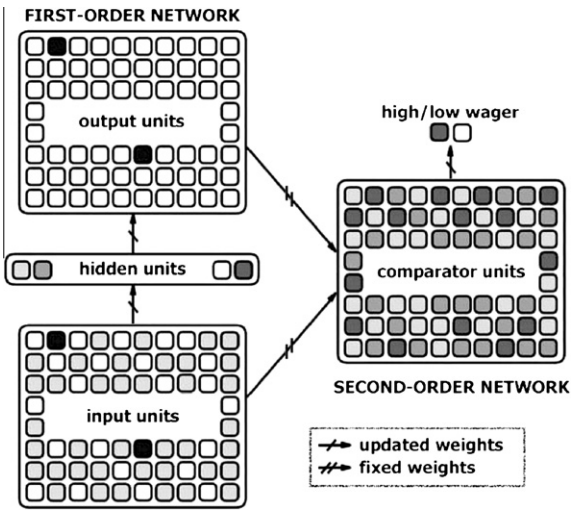

Fig. 1. Network architecture for blindsight and AGL simulations. The network consists of a first-order feedforward backpropagation autoassociator (the winner-take-all mechanism on the output is not represented), of which the input and output units are connected through fixed weight to a second-order comparator, which in turn feeds forward into two wagering units.

The Iowa Gambling task, as modified by Persaud et al., is a fundamentally different paradigm, hence the architecture of the Iowa Gambling task's simulation involves neither an autoassociator nor a comparator. Instead, the first-order network performs a supervised predictive task, and the second-order network's input consists of the activation pattern of the first-order network's hidden units. Unlike the other two simulations, the resulting second-order network's hidden unit representations are therefore neither automatic nor independent from the first-order network, but are instead learned in a manner that specifically depends on the first-order internal knowledge and task (Fig. 2). Indeed, learning occurs this time concurrently in every connection of both networks during the task, and the second-order network's wagering performance is modulated in direct correlation with first-order network's content and accuracy. In this simulation, the first-order network is directly reinforced as a function of decks' outcomes, whereas the second-order network learns to wager high when the outcome of the card selected by the first-order network is a reward, and to wager low when the outcome is a penalty (each card deck having different probabilities of presenting both types of outcomes).

2.2. Simulation 1 – blindsight

2.2.1. Experiment

In their blindsight experiment, Persaud et al. showed that GY (a patient who, under specific circumstances,

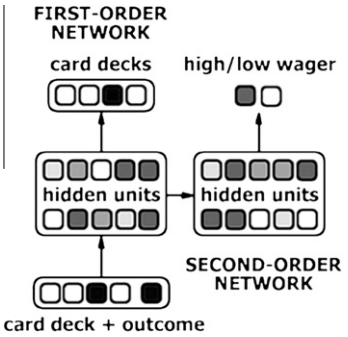

Fig. 2. Network architecture for the Iowa Gambling task's simulation. The network consists of a first-order feedforward backpropagator, of which the hidden units feedforward into a set of second-order hidden units, which in turn feed forward into two wagering units.

makes visual discriminations in the absence of visual awareness), when presented with subthreshold stimuli in his blind field, displayed above chance localization performance but failed to maximize earnings through wagering, suggesting that he was not always aware of the knowledge involved in his decisions for stimulus localization. However, for suprathreshold stimuli (both in normal and blind fields), GY maximized performance as well as earnings.

2.2.2. Simulation

We simulated these results by pre-training 15 networks to discriminate amongst arbitrary stimulus positions and to simultaneously place wagers on their own performance. Distinction between suprathreshold and subthreshold blindsight vision was introduced during a subsequent testing phase, in which the networks classified the patterns they had previously been presented with (suprathreshold), as well as degraded versions of these patterns in which stimulus-to-noise ratio was manipulated by increasing the noise level (subthreshold). As blindsight is commonly associated with lesions in the brain and in particular in V1, we chose to modulate the stimulus-to-noise ratio in the input of a fully connected metacognitive network, rather than simply removing the connections between the first-order network and the second-order network since this would obviously result in a dissociation (an additional way of simulating blindness is reported in Supplementary material).

2.2.3. Results and discussion

As shown in Table 1, the simulations closely capture experimental results. Discrimination performance, as simulated by the first-order network, is well above chance both under subthreshold and suprathreshold conditions (78.5% and 80.0% correct, respectively). However, networks tested in the subthreshold condition fail to wager advantageously on their correct and incorrect discriminations and instead wager at chance level, with only 48.5% of all trials (29 + 19.5%) being followed by an advantageous wager. This is not the case under suprathreshold conditions, where 66% of all trials are accompanied by advantageous wagers (50.5 + 15.5%).

The blindsight simulation illustrates the main principle of higher-order representations that are formed outside of the first-order network's causal chain by means of a comparator with an unsupervised learning mechanism. In addition to the first-order classification boundary, this comparator puts a second boundary on the distributed representations of the first-order error. However, the wagering capacity in this first simulation, while being causally independent from specific first-order patterns, is still connected to the first-order task of pattern recognition and therefore is still dependent in terms of content, because the network learns how to wager at the same time that it learns to recognize patterns. Indeed, when we learn to perceive the world, we learn to trust what we see (supervised pre-training of wagering), after which we are able to trust our own judgment (wagering during pattern testing). Nevertheless, having learned what percepts we can trust, we do not need to re-learn which of our judgments we can trust with every new perceptual input or task. The next simulation addresses this issue.

2.3. Simulation 2 – Artificial Grammar Learning Task

2.3.1. Experiment

In the AGL experiment, Persaud et al. show that, following incidental exposure to strings of letters (i.e., memorize "TSXVPP", "PVPXVT", etc.) produced by an artificial grammar, participants perform above chance on a subsequent and unexpected test asking them to discriminate between novel grammatical and non-grammatical strings while failing to maximize their earnings though wagering (implicit condition). This result is in line with typical implicit learning results (e.g., Dienes et al., 1995), suggesting that people can learn about the structure of the material while remaining unable to verbalize their knowledge. When participants were subsequently made aware of the grammar rules by being told what they were, however, they started to wager advantageously (explicit condition). Discrimination performance also improved but it was maintained, for comparison purposes, at the same level as under incidental conditions by reducing time of exposure to the strings during the test phase.

2.3.2. Simulation

We simulated these results by training two sets of 15 networks to classify artificial grammar strings. The metacognitive networks were similar to those used in the blindsight simulation, with the exception that we implemented the distinction between low and high awareness conditions by manipulating the first-order network's training phase length (short and long training phases corresponding to implicit and explicit conditions, respectively). It is obviously impossible to tell this network what the rules of the grammar are in a symbolic, abstract manner. However, for the purpose of the simulation, we only needed to induce the same improvements as in the experimental situation. As both performance and wagering were enhanced in the explicit condition (indeed, Persaud et al. directly told participants how to improve their performance), we simply let the first-order network learn for a longer period of time to obtain the same effect. We did not simulate the manipulation of string exposure time that would

Table 1
Results of the blindsight simulation.

Localization with	Experiment			Simulation		
Subthreshold stimuli	Correct	Incorrect	Total	Correct	Incorrect	Total
High wager	12	6	18	29	2	31
Low wager	62	20	82	49.5	19.5	69
Total	74	26	100	78.5	21.5	100
Suprathreshold stimuli	Correct	Incorrect	Total	Correct	Incorrect	Total
High wager	72	2	74	50.5	4.5	55
Low wager	18	8	26	29.5	15.5	45
Total	90	10	100	80	20	100

Percentages of localizations and corresponding wagers in low (subthreshold) and high (suprathreshold) consciousness conditions in Persaud et al.'s experiment (reproduced with permission) and in our simulation. Advantageous wagers are underlined.

have maintained performance at a level comparable to the implicit condition.

Crucial to this simulation, we wanted a second-order network that was able to wager independently from specific first-order material, and have it generalize its knowledge to novel material (recall that in the blindsight simulation, all test patterns had been presented previously). To achieve this, both networks were first subjected to pre-training on a set of random patterns, allowing the second-order network to learn how to wager independently of any first-order task and content. Half of these random patterns were accompanied by learning in the first-order network, while the other half were not. In both cases, the second-order network had to wager high when first-order input and output matched, and low when they did not, thus learning to establish a decision criterion based on the first-order network's accuracy. Following pre-training, all first-order network's connections were reset to initial conditions, whereas second-order network's weights were kept as was until the end of the simulation. (see Supplementary material for a detailed account).

2.3.3. Results and discussion

With this new setup, our simulation results again fit behavioral data (Table 2). Networks in the implicit condition performed above chance (71.8% correct), but failed to wager advantageously (56% of all trials). The explicit condition networks were not only better at the discrimination task (98% correct) – as a longer exposure period would predict – but also in placing advantageous wagers above chance (65% of all trials).

In addition to the principle illustrated in the blindsight simulation (higher-order representations formed outside of the first-order causal chain), the AGL simulation demonstrates that a higher-order network can generalize its wagering ability to new materials and contexts. Indeed, people can wager on their performance on any given task, without having to learn each time anew how to do so. In the AGL simulation, the second-order knowledge is generalizable thanks to the specific choice made for the pre-training task, being itself as general as possible: the indiscriminate learning of random patterns. From a developmental perspective, this pre-training simulates the effect of having the second-order network trained on many different tasks in order for the higher-order boundary to be adjusted independently from any first-order specific content. As a consequence, the metarepresentations generated by the new task (the AGL task) fit within the larger area of expertise of the second-order network. Thus the AGL simulation illustrates the principle of independency not only in terms of attitude, but also in terms of content. Finally the last simulation, because of differences in Persaud et al.'s experimental setup, serves to illustrate another kind of metarepresentations that might develop in our brain, and that are causally dependent on the first-order representations.

Table 2
Results of the AGL simulation.

Discrimination	Experiment			Simulation		
Implicit	Correct	Incorrect	Total	Correct	Incorrect	Total
High wager	36	6.5	42.5	36.5	8.5	45
Low wager	44.5	13	57.5	35.5	19.5	55
Total	80.5	19.5	100	72	28	100
Explicit	Correct	Incorrect	Total	Correct	Incorrect	Total
High wager	53.2	7.3	60.5	63.5	0.5	64
Low wager	20.1	19.4	39.5	34.5	1.5	36
Total	73.3	26.7	100	98	2	100

Percentages of discriminations and corresponding wagers by the network in low (implicit condition) and high (explicit condition) consciousness conditions in Persaud et al.'s experiment (reproduced with permission; explicit condition data are extrapolated from Persaud et al. Fig. S2) and in our simulation. Advantageous wagers are underlined. (A detailed breakdown of string classifications is listed in the Supplementary material).

2.4. Simulation 3 – Iowa Gambling task

2.4.1. Experiment

In Persaud et al.'s modified version of the Iowa Gambling task, participants select, on each trial, one of four decks of cards, each with different pay-offs (e.g., some decks offer relatively low rewards but are ultimately advantageous, whereas others offer much larger rewards but also larger penalties, proving disadvantageous over the long run). After deck selection, but before turning over the card (revealing how much was won or lost), participants wager on whether the card will be winning or losing. 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 (i.e., "What would you expect your average winning amount to be by picking 10 cards from deck 1?") (e.g., Maia & McClelland, 2004), wagering follows performance more closely (Fig. 3b). This experiment differs from the others in two basic ways, necessitating a different simulation approach. First, the Iowa Gambling task requires participants to initially explore the material before being able to create any representation about the decks' yields. The resulting meta-representations are thus necessarily dependent on this exploration phase. Second, participants receive feedback on each trial about the quality of their wagering, as the turning of the card immediately reveals whether and how much is won. As a consequence, participants can use this feedback to unconsciously optimize not only their deck selection, but also their wagering. This does not rule out the fact that participants effectively become aware of relative pay-offs along the experiment but this makes wagering in the Iowa Gambling task less suitable as a measure of awareness, since advantageous wagering could in principle emerge in the absence of awareness. Additionally, the high awareness condition consisted in asking participants specific questions regarding strategy every 10th trial after the 20th. This, in contrast to the AGL experiment, affected their wagering but not their task performance. This suggests that in this situation, manipulating awareness only modulates processing in the second-order network, leaving the first-order network's processing intact.

2.4.2. Simulation

Two sets of 15 networks learned to perform the deck selection task while wagering on the gain obtained on each decision. These networks were modified in three ways to reflect the task differences described above. First, the first-order network could not be an autoassociator since the desired states were not available as inputs as in the previous simulations. Instead it had to select one out of four card packs at first and received feedback about the result of that selection (win or loss) only after the selection had been made. Second, because the first-order network wasn't an autoassociator, the second-order network meta-representations could not rely on a comparator, but instead consisted of a hidden layer directly connected to the first-order network's hidden units, and feeding forward into the (wagering) outputs (Fig. 2). Finally, to modulate only wagering performance, we implemented the

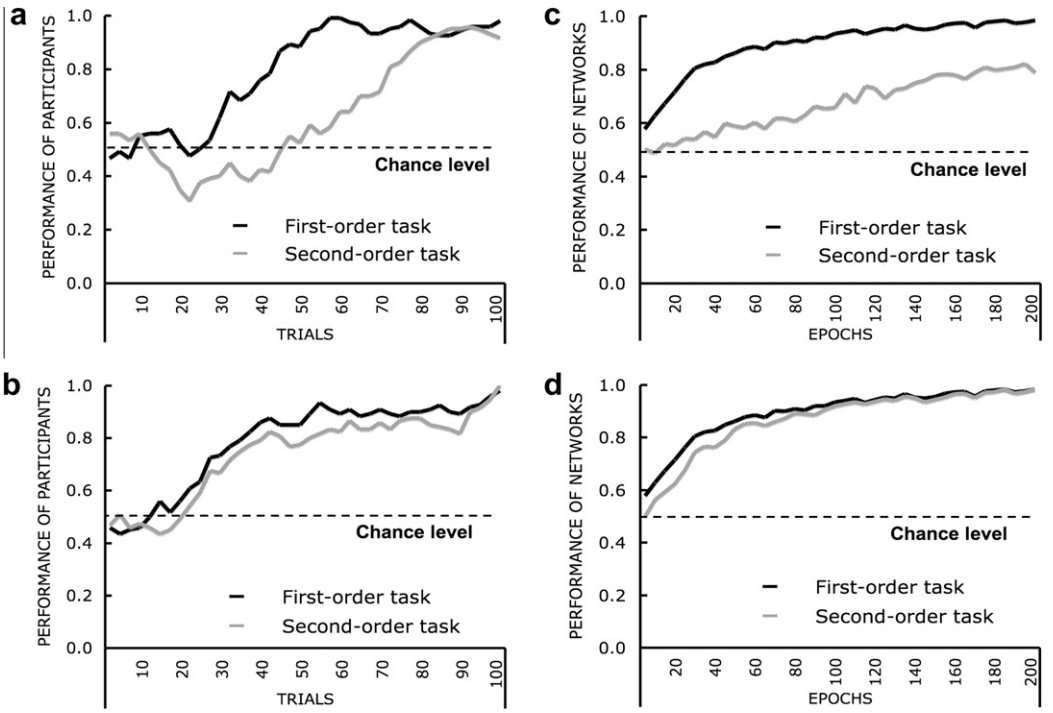

Fig. 3. Results for the Iowa Gambling task's simulation. Network performance is plotted across time (epochs) for: (c) Low Consciousness, and (d) High Consciousness conditions. Persaud et al.'s results are reproduced (with permission) for comparison purpose (a and b for Low and High Consciousness conditions respectively).

distinction between "Low Consciousness" and "High Consciousness" conditions by setting the second-order network's learning rate low or high, respectively, without affecting the first-order network.

2.4.3. Results and discussion

Fig. 3 displays the performance of both networks over time. Just as for the experimental data, wagering performance lags advantageous card deck selection in the "Low Consciousness" condition. By contrast, in the "High Consciousness" conditions, in which we increased the efficacy with which the second-order network could make use of the first-order network's hidden unit representations, wagering closely follows card deck selection.

The simulation of the Iowa Gambling task demonstrates a different way of creating higher-order representations through supervised, rather than unsupervised, learning. The resulting metarepresentations are this time embedded within the causal chain of the first-order network. Therefore, the second-order network lacks independency of content, as was the case in the blindsight simulation, since both the first-order and the second-order knowledge lie within the exact same domain. Crucially, in addition to this, as both networks are causally dependent, metaknowledge is prevented from becoming attitude-explicit (i.e., accuracy is not represented). As a consequence, this knowledge needs not to be conscious, since it remains "in" the system without being available "for" the system. In the general discussion, we elaborate on how the two types of metarepresentations (with or without attitude-explicitness) might be related.

3. Discussion

The simulations demonstrate three ways of modeling the pattern of associations and dissociations between performance and wagering as reported experimentally as distinctions between performance with and without awareness by Persaud et al. (2007). Furthermore, they suggest that a metacognitive network can be trained to evaluate its own performance through the development of metarepresentations that redescribe relevant first-order representations to the network itself. Though we do not claim that metacognitive networks are conscious in any sense, we would like to discuss three fundamental issues on which we believe these simulations shed some light.

3.1. The nature of metacognition

First, the basic assumption that underlies our modeling work is that metacognition, as probed in this context by wagering ability, requires higher-order representations. Here we suggest that such representational redescription emerges when a system is allowed to observe its own internal states. This makes it possible for task-related knowledge not merely to drive performance in the system, but, crucially, to become available as a further object of representation for the system (e.g., Clark & Karmiloff-Smith, 1993; Karmiloff-Smith, 1992; Mandler, 2004). We surmise that this mechanism forms the core of the Higher-Order Thought Theory of Consciousness (e.g., Rosenthal, 1997), which takes it that one is conscious of some content when one is conscious that one knows this content (i.e., when one is conscious of possessing this content). However, we suggest that higher-order representations: (1) do not necessarily provide any conscious access, as their accrual may only render the first-order state content-, but not attitude-explicit, notably when there is no causal independence between first-order and higher order states (the Iowa Gambling task simulation); and (2) can allow the transfer of knowledge from a second-order task to another, notably when they are causally independent from the specific first-order knowledge, (as it is the case in the AGL simulation). Furthermore, even though we did not explore this in the current article, we assume that: (1) metarepresentations allow, as far as the current task requires it, the control of lower-order processes through top-down interactions (hence affecting the conscious states themselves), and that (2) simple second-order representations are by nature unconscious unless further higher-order representations are recursively built upon them.

Despite its apparent similarity to signal detection accounts of metacognition (e.g., Scott & Dienes, 2008), the here presented notion of knowledge becoming available for the system by re-representation in a higher-order network differs from other such signal detection based models in a crucial way. The latter typically make the second-order distinction between confidence and guessing (high and low wagers) on the very signal that is used for first-order classification, 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 (between the first-order and second-order boundary), and confidence (on the far side of the second-order boundaries). In such an account, confidence or high wagers depend essentially on signal strength. However, in our current model (specifically the first two simulations), in which the second-order network's representations lie outside of the first-order causal chain, the second-order classification does *not* depend on the same signal as the first-order task. Instead of wagering high or low based on signal strength, the second-order network re-represents the first-order error, thus basing itself more on a consequence of signal coherence. Therefore, before it can wager, the second-order network, like the first-order network, has to learn how to make a *single*-boundary classification based on this second-order representation (the distributed error representation). Such a classification means that the second-order network has conceptually learned to judge the first-order networks' accuracy, independently of the first-order task.

This difference between our model and the more standard Signal Detection Theory account is substantial, for it impinges on whether one considers that Type I and Type II performance, that is, first-order decisions 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. Our simulations are suggestive that the former may be more fruitful in that they afford additional flexibility and generality.

3.2. Learning to be conscious

Second, our simulations assume that metarepresentations are, just like their corresponding first-order representations, learned based on experience. However, this is accomplished in two fundamentally different ways in the simulations. The first two models use second-order comparators, which form metarepresentations of the difference between on the one hand the current internal state or prediction of a first-order network, and on the other hand the effective external state corresponding to the input (or target) it received. As suggested before, such comparators may play a crucial role in consciousness (e.g., Frith et al., 2000; Gallagher, 2004; Mandler, 2004; Pacherie, 2008; Rizzolatti et al., 1996; Sperry, 1950; Synofzik et al., 2008; Wolpert & Kawato, 1998), and particularly in the sense of enabling agency, as they inform an agent about the adequacy of its own internal states. In these simulations, pattern comparison and the resulting metarepresentations emerge through a non-supervised process and are hence not learned in the classic sense through feedback. However, the way in which such specific metarepresentations produce a high or a low wager is learned in a supervised, feedback-driven way, independently from the pattern-specific comparisons (that occur beforehand). Conversely, in the third simulation, emergence of metarepresentations occurs through supervised reinforcement between two tasks of interest. This second type of metarepresentations lacks the causal independence that would be necessary for a system to know that it possesses internal knowledge.

The fact that our comparators do not learn during the task does not imply that at one point one stops to "learn to be conscious". What it means is that, rather than sudden jumps or shifts in what content becomes conscious, such as a newborn might have, learning to be conscious for adults would involve continuous but infinitesimal adjustments with every novel worldly experience, eventually approaching an asymptote. In these simulations we artificially introduced such an early asymptote by freezing the second-order network's learning, thus providing a snapshot of a system learning to be conscious – for it is obvious that none of the networks here are conscious. What the simulations do suggest is that "learning to be conscious" could evolve from a more supervised learning-based criterion setting, to a more gradual, unsupervised adaptation. The causally dependent metarepresentations exemplified in the third simulation might thus represent a way in which such early criterion setting occurs (see also Cleeremans et al., 2007). Though it lies beyond the scope of this paper, we consider it possible, and in fact most probable, that the use of first-order knowledge in different feedback-driven second-order tasks (as exemplified in the third simulation) may eventually lead to the accrual of metaknowledge that progressively becomes independent of specific first-order tasks, and may thus serve as a comparator that allows the brain to immediately evaluate the accuracy of its own internal signals.

3.3. Objective and subjective awareness

Third, our simulations speak to the distinction between objective and subjective measures of consciousness. The literature on the differences between conscious and unconscious processing is characterized by continuing debates about the proper methodology through which to assess awareness (e.g., Butler & Berry, 2001; Holender, 1986; Merikle & Reingold, 1991; Shanks & StJohn, 1994) but that have tended to ignore the fundamental point that objective and subjective measures concern at first different kinds of knowledge (e.g., Fu, Fu, & Dienes, 2008): Knowledge about the world ("worldly discrimination") (e.g., Lau, 2007) in the case of objective measures, and knowledge about one's own mental states ("mental sate discrimination") in the case of subjective measures. Our simulations provide a mechanism that instantiates this distinction, here not only in terms of content but also in terms of causal independency between different levels of representation: representation of the task (or world) and criterion setting.

In conclusion, our simulations are broadly supportive of the provocative idea that consciousness results from the continuous operation of unconscious learning and plasticity mechanisms that make it possible for a system to redescribe its own activity to itself—a thesis that we have dubbed "the radical plasticity thesis" (Cleeremans, 2008). Thus, the brain not only learns about the world, but also about its own representations of it, so developing, through experience, metarepresentations that inform it about the geography of its own internal states. This in turn makes it possible for an agent to qualify the manner, or the mental attitude, in which first-order knowledge is held: Is this something that I fear, that I hope, that I regret, etc.? Such sensitivity to the qualities of one's own internal states as well as their relationships to other internal states forms the basis of subjectivity and, we claim, is constitutive of what it means for an agent to be conscious.

Acknowledgements

We thank N. Persaud, D. Rosenthal, Z. Dienes, A. Seth, and J.L. McClelland for their helpful comments on earlier versions of this paper. This research was supported by a Grant from the National Fund for Scientific Research (FRS – FNRS, Belgium) to A.P.; B.T. is supported by Marie Curie Action IEF #237502 "Social Brain"; A.C. is a Research Director with the National Fund for Scientific Research (FRS – FNRS, Belgium). This work is supported 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 FRFC/ESF Grant #2.4577.06 "Mechanisms of serial action".

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.cognition.2010.08.010.

References

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50*, 7–15. doi:10.1016/0010-0277(94)90018-3.

Butler, L. T., & Berry, D. C. (2001). Implicit memory: Attention and awareness revisited. *Trends in Cognitive Sciences, 5*, 192–197. doi:10.1016/S1364-6613(00)01636-3.

Clark, A., & Karmiloff-Smith, A. (1993). The cognizer's innards: a psychological and philosophical perspective on the development of thought. *Mind and Language, 8*, 488–519. doi:10.1111/j.1468-0017.1993.tb00299.x.

Cleeremans, A. (2008). The radical plasticity thesis. In R. Banerjee & B.K. Chakrabarti (Eds.), *Progress in brain research, 168*, 19-33. doi:10.1016/S0079-6123(07)68003-0.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation, 1*, 372–381. doi:10.1162/neco.1989.1.3.372.

Cleeremans, A., Timmermans, B., & Pasquali, A. (2007). Consciousness and metarepresentation: A computational sketch. *Neural Networks, 20*, 1032–1039. doi:10.1016/j.neunet.2007.09.011.

Dehaene, S. & Charles, L. (2010). A dual-route theory of evidence accumulation during conscious access. In H. Lau (Chair), *Conscious awareness, perceptual decision making and the bayesian brain. Symposium conducted at the 14th annual meeting of the association for the scientific study of consciousness*, Toronto, Canada (June).

Dienes, Z., Altmann, G., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 1322–1338. doi:10.1037/0278-7393.21.5.1322.

Dienes, Z., & Perner, J. (1996). Implicit knowledge in people and connectionist networks. In *Implicit cognition* (pp. 227–256). Oxford University Press.

Dienes, Z., & Seth, A. (2009). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition*. doi:10.1016/j.concog.2009.09.009.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211. doi:10.1016/0364-0213(90)90002-E.

Frith, C. D., Blakemore, S. J., & Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society, B, 355*, 1771–1788. doi:10.1098/rstb.2000.0734.

Fu, Q., Fu, X., & Dienes, Z. (2008). Implicit sequence learning and conscious awareness. *Consciousness and Cognition, 17*, 185–202. doi:10.1016/j.concog.2007.01.007.

Gallagher, S. (2004). Neurocognitive models of schizophrenia: A neurophenomenological critique. *Psychopathology, 37*, 8–19. doi:10.1159/000077014.

Holender, D. (1986). Semantic activation without conscious activation in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioural and Brain Sciences, 9*, 1–23. doi:10.1017/S0140525X00021269.

Holender, D., & Duscherer, K. (2004). Unconscious perception: The need for a paradigm shift. *Perception & Psychophysics, 66*, 872–881 (PMid:15495911).

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

Koch, C., & Preuschoff, K. (2007). Betting the house on consciousness. *Nature Neuroscience, 10*, 140–141. doi:10.1038/nn0207-140.

Lau, H. (2007). A higher order Bayesian decision theory of consciousness. In R. Banerjee & B.K. Chakrabarti (Eds.), *Progress in brain research, 168*, 35-48. doi: 10.1016/S0079-6123(07)68004-2

Lau, H. (2010). Comparing different signal processing architectures that support conscious reports. In H. Lau (Chair), *Conscious awareness, perceptual decision making and the bayesian brain. Symposium conducted at the 14th annual meeting of the association for the scientific study of consciousness*, Toronto, Canada, June

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 Science USA, 101*, 16075–16080. doi:10.1073/pnas.0406666101.

Mandler, J. M. (2004). *The foundation of mind: Origins of conceptual thought*. New York: Oxford University Press.

Merikle, P. M. (1992). Perception without awareness: Critical issues. *American Psychologist, 47*, 792–795. doi:10.1037/0003-066X.47.6.792.

Merikle, P. M., & Reingold, E. M. (1991). Comparing direct (explicit) and indirect (implicit) measures to study unconscious memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 224–233. doi:10.1037/0278-7393.17.2.224.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition, 107*, 179–217. doi:10.1016/j.cognition. 2007.09.003.

Persaud, N., & McLeod, P. (2007). Wagering demonstrates subconscious processing in a binary exclusion task. *Consciousness and Cognition, 17*, 565–575. doi:10.1016/j.concog.2007.05.003.

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience, 10*, 257–261. doi:10.1038/nn1840.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning & Verbal Behavior, 6*, 855–863. doi:10.1016/S0022-5371(67)80149-X.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*, 131–141. doi:10.1016/0926-6410(95)00038-0.

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.

Ruffman, T., Rustin, C., Garnham, W., & Parkin, A. J. (2001). Source monitoring and false memories in children: Relation to certainty and executive functioning. *Journal of Experimental Child Psychology, 80*, 95–111. doi:10.1006/jecp. 2001.2632.

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition*. doi:10.1016/j.concog.2009.12.013.

Scott, R. B., & Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 34*, 1264–1288. doi:10.1037/a0012943.

Seth, A. K. (2008). Post-decision wagering measures metacognitive content, not sensory consciousness. *Consciousness and Cognition, 17*, 981–983. doi:10.1016/j.concog.2007.05.008.

Shanks, D. R., & StJohn, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17*, 367–395. doi:10.1017/S0140525X00035032.

Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative Physiology and Psychology, 43*, 482–489. doi:10.1037/h0055479.

Stoerig, P., Zontanou, A., & Cowey, A. (2002). Aware or unaware: Assessment of cortical blindness in four men and a monkey. *Cerebral Cortex, 12*, 565–574. doi:10.1093/cercor/12.6.565.

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition, 17*, 219–239. doi:10.1016/j.concog.2007.03.010.

Tunney, R. J., & Shanks, D. R. (2003). Subjective measures of awareness and implicit cognition. *Memory & Cognition, 31*, 1060–1071. PMid:14704021.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks, 11*, 1317–1329. doi:10.1016/S0893-6080(98)00066-5.