FINAL DRAFT of Timmermans, B., Schilbach, L., Pasquali, A., & Cleeremans, A. (2012) Higher-Order Thoughts in Action: Consciousness as an unconscious redescription process. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1412-1423. doi:10.1098/rstb.2011.0421.

Higher-Order Thoughts in Action:

Consciousness as an unconscious re-description process

Bert Timmermans[a], Leonhard Schilbach[b], Antoine Pasquali[c,d], & Axel Cleeremans[c]

[a] Neuroimaging Group, Psychiatry & Psychotherapy Clinic, University Hospital of Cologne, Kerpener Str. 62, 50937 Cologne, Germany; corresponding author, bert.timmermans@uk-koeln.de

[b] Max Planck Institute for Neurological Research, Gleueler Str. 50, 50931 Cologne, Germany;

[c] Consciousness, Cognition and Computation Group, Université Libre de Bruxelles CP 191BR, Av. F.-D. Roosevelt 50, 1050 Brussels, Belgium;

[d] Neurogenics Research Unit, Adam Neurogenics, 133 Marine de Solaro, 20240 Solaro, France

Abstract

Metacognition is usually construed as a conscious, intentional process whereby people reflect upon their own mental activity. Here, we instead suggest that metacognition is but an instance of a larger class of representational re-description processes that we assume occur unconsciously and automatically. From this perspective, the brain continuously and unconsciously learns to anticipate the consequences of action or activity on itself, on the world, and on other people through three predictive loops: An inner loop, a perception–action loop, and a self–other (social cognition) loop, which together form a tangled hierarchy. We ask what kinds of mechanisms may subtend this form of enactive metacognition. We extend previous neural network simulations and compare the model with Signal Detection Theory, highlighting that while the latter approach assumes that both type I (objective) and type II (subjective, metacognition-based) decisions tap into the same signal at different hierarchical levels, our approach is closer to dual-route models in that it assumes that the re-descriptions made possible by the emergence of meta-representations occur independently and outside of the first-order causal chain. We close by reviewing relevant neurological evidence for the idea that awareness, self-awareness and social cognition involve the same mechanisms.

Keywords: Consciousness, metacognition, blindsight, artificial grammar learning, neural networks, social cognition

1. Introduction

There is undoubtedly a relationship between awareness and metacognition, for our common understanding of conscious knowledge is simply that it is knowledge that we know we possess. Congruently, it is precisely in those cases where our behaviour is guided by knowledge we do *not* know we possess that we speak of *unconscious* knowledge. Colloquially, thus, metacognition, or "cognition about cognition", appears to be fundamental to our understanding of consciousness. However, metacognition is usually construed as a controlled, intentional process whereby people intentionally and effortfully reflect upon their own mental activity. Here, we would instead like to suggest that metacognition is but an instance of a larger class of *representational re-description* processes [1] that we assume occur unconsciously, automatically, and continuously. From this perspective, the brain is continuously and unconsciously learning to anticipate the consequences of action or activity on itself, on the world, and on other people. In so doing, we shall argue, it learns to re-present its own activity to itself, so developing systems of meta-representations that characterise the manner in which first-order representations are held. Such systems of meta-representations both enable conscious experience (for it is in virtue of such meta-representations that the agent "knows that it knows") and define its subjective character (for each agent's meta-representations will be idiosyncratic, shaped by its experience with the world and with others).

To support these ideas, we begin by discussing the relationships between consciousness and metacognition. Next, we ask what kinds of mechanisms are necessary to subtend it. We argue that Signal Detection Theory (SDT), as applied to the study of consciousness, has a descriptive character that we should like to see replaced by a mechanistic account. We propose such an account in the next section, based on the neural network models we initially introduced in two previous papers [2,3]. Next, we analyse the performance of such models through Signal Detection Analysis, explore their implications for our understanding of consciousness, and overview relevant neurological evidence. We close by suggesting that consciousness is something that the brain learns to do rather than a static property of certain neural representations and not others. This we call the "Radical Plasticity Thesis".

2. Metacognition

Metacognition covers a lot of ground. It has been variously construed as the ability to reflect upon one's own mental activity ("cognition about cognition"), as awareness of possessing task-relevant knowledge (so-called judgment knowledge [4]), or as the introspective mechanism that lies at the core of perceptual awareness (i.e., sensory metacognition). A number of recent papers have addressed both the neurobiological underpinnings of metacognition [5–7], as well as its functions and mechanisms [8,9].

The complex relationship between consciousness, self-awareness, and metacognition is the object of an ongoing debate (e.g., [10,11] – see also [8] for an overview). In a nutshell, the argument hinges on whether metacognition is taken to be a precondition or a consequence of consciousness. Contemporary theories of consciousness, in this respect, roughly fall into one of two categories: those that see capacity for metacognition as a consequence of content becoming conscious and therefore available to higher-order processes and introspection (so-called "fame-in-the-brain" approaches), and those that assume that some form of metacognition is a necessary prerequisite for consciousness.

"Fame in the brain" theories, introduced by Dennett [12,13], typically assume that consciousness occurs whenever particular conditions are fulfilled, such as stability and strength or complexity of a knowledge representation, which can result from processes such as re-entrant processing and/or from synchrony of neural processing. Essentially, it is assumed that the brain is a large dynamical system in which stable, attractor states come in and out of existence as a result of continuously operating global constraint satisfaction processes. The main functional consequence of such states is that the information they convey then becomes available to the global workspace [14–16] for further information processing, such as cognitive control, or conscious access. However, one problem with "fame-in-the-brain" proposals is that there is no particular property of the information contained in conscious representations, apart from strength, stability or complexity, that sets it qualitatively apart from information contained in unconscious representations. All information remains first-order information in the system, and some of that information somehow gives rise to conscious awareness of it.

As an alternative point of view, approaches that take higher-order- or meta-representations as a prerequisite for consciousness hold that in order for content to become conscious, a system needs to be able to represent its internal states *to* itself. In other words, for a system to be conscious of its internal states, said internal states have to become available to inspection, in addition to serving

their first-order functions. As Karmiloff-Smith [1] put it: knowledge in the system has to become knowledge for the system. First-order systems — systems that merely transform, however appropriately, inputs into outputs — can never know *that* they know: They simply lack the appropriate machinery [17]. This points to a fundamental difference between sensitivity and awareness. Sensitivity merely entails the ability to respond in specific ways to certain states of affairs. Sensitivity does not require consciousness in any sense. A thermostat can appropriately be characterised as being sensitive to temperature, just as the carnivorous plant *Dionaea muscipula* (Venus flytrap) may appropriately be described as being sensitive to movement on the surface of its leaves. But our intuitions tell us that such sensitive systems (thermostats, photodiodes, transistors, cameras, carnivorous plants) are not conscious. They do not have "elementary experiences", they simply have no experiences whatsoever. Sensitivity can involve highly sophisticated knowledge, and even learned knowledge, but such knowledge is always first-order knowledge, it is always knowledge that is necessarily embedded in the very same causal chain through which processing occurs.

Awareness, on the other hand, always seems to minimally entail the ability of knowing *that* one knows. This ability, after all, forms the basis for the verbal reports we take to be the most direct indication of awareness. And when we observe the absence of such ability to report on the knowledge involved in our decisions, we conclude that the decision was based on unconscious knowledge. Thus, it is when an agent exhibits *knowledge* of the fact that he is sensitive to some state of affairs that we take this agent to be a conscious agent. This *second-order* knowledge, we argue, critically depends on *learned* systems of metarepresentations, and forms the basis for conscious experience of the first-order knowledge that is the target of such metarepresentations. Despite remaining heavily debated, this higher-order approach to consciousness has received substantial support recently [10,18–22] (see also [8] for a recent overview) and is currently enjoying renewed interest.

Irrespective of whether one sees metacognition as a consequence or as a prerequisite to awareness there remains the questions of what mechanisms subtend it. In this respect, Lau [23] has defended the idea that metacognition involves the brain performing signal detection on its own representations. For instance, in a typical visual detection or discrimination task aimed at investigating task performance and awareness, participants have an "objective" discrimination performance, and a "subjective" awareness rating. SDT approaches to awareness [9,24–28] model this relationship by assuming that for each of these judgments, the participant's (and the brain's) task comes down to representing the outside world in terms of stimulus and noise, and looking for

decision criteria to set both apart in objective (type I) and subjective (type II) terms. In general terms, this comes down to calculating two sensitivities and criteria. Type I sensitivity d'_1 is, as usual, based on the proportion of hits with respect to the proportion of false alarms in the context of the actual task, and criterion c_1 represents the bias with which the participant tends to be conservative versus risk-taking (in detection tasks; or selects one response option over the other in discrimination tasks). Type II sensitivity d'_2, however, which is the degree to which one can tell apart one's correct from one's false responses, is thus the number of "awareness hits" with respect to "awareness false alarms." This, if awareness is measured by rating one's confidence in one's response, d'_2 reflects the proportion of high confidence ratings for my correct responses with respect to the proportion of high confidence ratings for wrong responses, whereas c_2 reflects my bias in terms of how prone I am to rate my confidence as high or low. The relationship between type I and type II SDT analysis has been described in depth elsewhere [29].

Within this general framework however, important differences exist between how "fame-in-the-brain" or higher-order approaches characterise this relationship. Recent modelling work [27] has laid out the different classes of possible models that follow from the above distinction within a SDT framework. The study distinguishes three types of models: *first-order models*, which assume that one stream of information accounts for both behavioural output and awareness of this output, *dual-channel models*, which assume that information that informs behaviour is essentially processed along a different channel than that which informs awareness of this information, and *hierarchical models*, which assume that information is first processed on a first-order level (which determines behaviour), and that a second-order level is necessary to make the information available to awareness. The modelling results [27] show that hierarchical SDT models outperform first-order or dual-channel models.

SDT, however, offers essentially a descriptive account of the relationships between type I and type II performance. Here, building on earlier work, we would like to propose a computational account [2,3] of these relationships. This proposal is motivated by different reasons.

First, as mentioned before, both "fame in the brain" and higher-order approaches as operationalised in SDT somehow assume that metacognition, whether a consequence or a prerequisite, is necessarily tied to consciousness. Here we argue that metacognition may be an instance of a larger class of learning-related representational re-description processes [1] that, we assume, occur unconsciously and automatically.

Second, we believe that, whereas SDT might provide a conceptual description of what occurs in any given visual detection or discrimination task (as mentioned above: the brain performing signal detection on itself), it offers no explanation as to how such signal detection might come about and therefore remains largely descriptive: it is not because people behave as if performing a signal detection task that this is how the brain produces this behaviour. This is not an argument about biological plausibility (which has also been criticised for neural network models), but about explanatory power. In our opinion SDT models lack an account of how the brain develops criteria, how it develops a representation of the world, and how it develops awareness. In our view, it is crucial to incorporate an organism's interaction with the world in order to understand how metacognition develops.

Third, conceptually, type II SDT in the context of awareness is somewhat ambiguous. In a type I task, there is, objectively, a stimulus present or not, and we can say there is one, or not – there is no a priori relationship (d'_1) between the two. Thus, my ratings can correspond to or diverge from the actual probability of a stimulus being present in the experiment. A type II task (e.g., confidence ratings) are completely different. There is a probability of correct decisions (which is a match between the world an the type I decision), but I do not simply provide subjective ratings that correspond or diverge from this probability. This is because a guess is just that, a guess. Confidence in a response A (instead of B) indeed means that I thought it was A, but when I claim to guess, I do not say "A is wrong," and that it should be B – rather, it means that for all I care it could be either of them. Overall, there are usually no (or very few) trials in which I know I was wrong, I am just not sure whether I was right. Indeed, if I consistently say "guess" *only* for trials where I make an error, I am in fact fully aware (see zero correlation criterion [30]). So in principle, irrespective of the relative proportion of guesses on correct versus incorrect trials (the "misses" versus the "correct rejections"), those "guess" trials should contribute in equal proportions, or not at all, to how I represent my decisions to myself, since when I guess, I do not state that my type I decision was wrong. Thus, at least in our opinion, type II tasks cannot be seen simply as a higher-level equivalent of type I tasks. There are many ways in which one can define the relationship between type I and type II decision axes, but those described by Maniscaldo and Lau [27] do not include a *mechanism* that accounts for the accrual over time on both decision axes and how their relationship comes to be established.

Fourth, on a more general note, in our view SDT, irrespective of whether it is implemented as a first-order, dual-channel, or hierarchical model, assumes (1) that a noisy but rich signal enters the sensory channels, and (2) that the brain represents one or two sensitivities (d') and sets at least two

criteria (c) that allow for the selection of the adequate type I and type II outputs. Apart from the fact that these criteria have to be arbitrarily chosen and hence that there is no explanation on how they come about, this approach is reminiscent of traditional filter models and of spectatorial accounts of cognition in general, whereby the senses receive massive (though noisy) amounts of information, and where the passive observer's brain is merely tasked to extract the signal. In this respect, one of the important variables manipulated in Maniscaldo and Lau's [27] hierarchical models is a decay factor, which determines how much of the first-order information remains for the second-order classification. This suggests, first, that somehow at one point there is an enormous amount of information (rich phenomenal consciousness) that dissipates over time, leaving only limited access to whatever remains [31,32], and second, that consciousness is essentially a passive endeavour. Indeed, using a decay, one has to subscribe to the fact that consciousness "slips through our fingers" – whereas in fact there have been recent findings suggesting that consciousness takes time [33], and that over this time, many misconstruals can happen [34,35]. In fact, it has been argued that conscious content is but an "illusion" created by the brain based on piecemeal sensory input in combination with priors (partial awareness hypothesis [36]; see also [37]), a notion which, to some extent, is also in line with an enactive view on consciousness [38,39], whereby the agent, embedded in an environment, is not a spectator but plays an active role in constructing his awareness of that environment and of himself (see below for an elaboration of this idea). Thus, even if one accepts that SDT criteria can be influenced by priors, there is no account of how this might happen. Taken together with the second point, SDT accounts are very useful at a descriptive level, but lack a developmental perspective, both in terms of how they come about through interaction of an organism with the world, and in terms of how conscious content is generated based on priors acquired through such interactions. The simulation work we carried out in Pasquali *et al*. [2] is an attempt to offer an alternative, computationally oriented, account. We revisit this work in the next section.

3. A hybrid neural network approach

We recently proposed a neural network approach to metacognition [2,3]. The core idea of our approach, which bears some resemblance to the actor-critic models introduced by Sutton and Barto [40], is that two independent networks (a "first-order" network and a "second-order" network) are connected to each other in such a way that the entire first-order network is input to the second-order network (**figure 1**). This means that all the units of the first-order network are used as input for a

second network, which can then in principle learn to discriminate the different ways in which the first-order network's internal representations match the outside world.

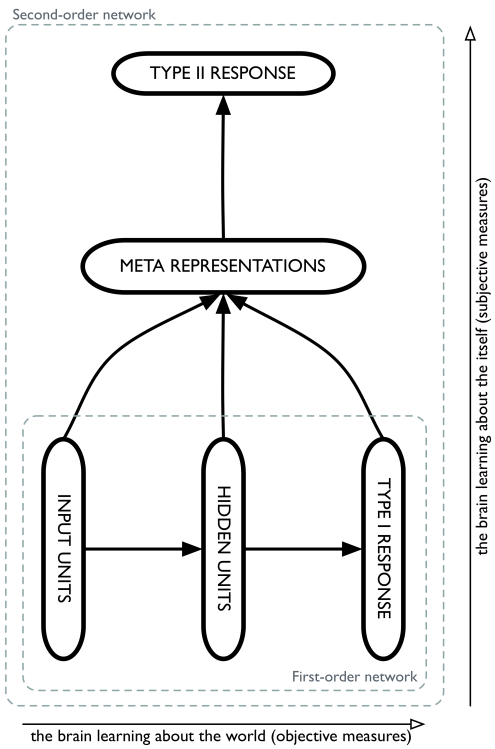

Figure 1. General architecture of a metacognitive network. A first-order network, consisting for instance of a simple three-layers backpropagation network, is trained to perform a simple classification task and thus contains knowledge that links inputs to outputs in such a way that the network can produce type I responses. This entire first-order network then constitutes the input to a second-order network, the task of which consists of redescribing the activity of the first-order network in some way. Here, the task that this second-order network is trained to perform is to issue type II responses, that is, judgments about the extent to which the first-order network has performed its task correctly. One can think of the first-order network as instantiating cases where the brain learns about the world, and of the second-order network as instantiating cases where the brain learns about itself.

Both networks are, for instance, simple feedforward back-propagation networks. The first-order network is trained to perform a simple discrimination task, that is, to produce type I responses, whereas the second is trained to judge of the accuracy of the first-order network's responses, that is, to perform type II judgments. In its more general form, as depicted in **figure 1**, such an architecture would also be sufficient for the second-order network to also perform other kinds of judgments, such as distinguishing between an hallucination and a veridical perception, developing knowledge

about the overall geography of the internal representations held by the first-order network, or forming propositional attitudes.

The fundamental difference between this type of model (a "metacognitive network") and SDT models is that the former learns and develops both first- and second-order representations over time. Pasquali *et al*. [2] instantiated the general architecture depicted in **figure 1** in different ways. One instantiation was a strictly hierarchical model (**figure 2*a***) whereas the other is best described as implementing a hybrid between dual-route models and hierarchical models (**figure 2*b***).

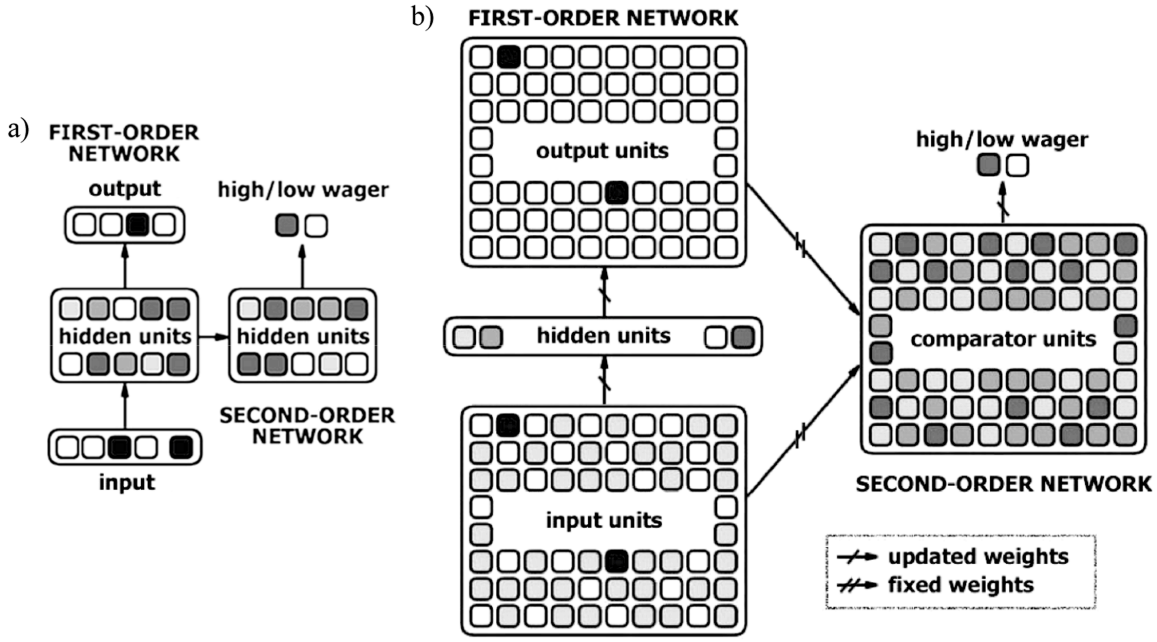

Figure 2. (a) Network architecture for the Iowa Gambling Task simulation (see [2], simulation 3). The network consists of a first-order feedforward backpropagator, of which the hidden units feedforward into a set of second-order hidden units, which in turn feed forward into two wagering units. **(b)**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

The hierarchical instantiation, which we will here dub "hidden unit-readers" (**figure 2*a*)** (see [3], and [2], simulation 3), directly reads out the first-order network's internal representations from its hidden units (containing the relationships between input and output patterns [41]. The model is hierarchical because the sensory input needs to be fully processed by the first-order network before it becomes available to the second-order network. The information contained in the second-order network is directly dependent on the information contained in the first-order network in that the hidden unit patterns predict both the first-order and the second-order responses.

Re-representing knowledge through meta-representations (i.e., "content-explicit representations") is not sufficient, however: one must also represent oneself as being in possession of that content ("attitude-explicit representations" [42]). Such attitude-explicit representations require 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 is represented. Indeed, it has been suggested that metacognition hinges upon encoding the precision of a representation, because this would allow organisms not only to evaluate what they know, but to engage in prospective error monitoring and optimisation of decision making, for instance by smoothing the accrual of evidence for the "right" decision over time [43].

We also explored the characteristics of a second instantiation (**figure 2*b***; 'comparator units', [2]: simulations 1 and 2), which indirectly reads out the first-order network's internal representations by comparing first-order input with first-order output (the latter of which is, in fact, the computational consequence of the hidden unit patterns). In these networks, the second-order network lies outside of the first-order causal chain, because the information used by the first-order network to execute its task is not the information used by the second-order network to place a high or a low wager. Thus, they are in principle dual-channel models. Still, since both networks "plug into" the same basic knowledge (first-order performance; albeit in a different way, see below), this type of model is effectively a hybrid between hierarchical and dual-route models.

Our hybrid models thus depend on two core assumptions: First, evaluating one's own performance requires that the first-order representations that are responsible for performance be accessed in a manner that is independent from their expression in behaviour. Second, one must possess attitude-explicit representations that require 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 is represented. The first of these assumptions refers to the hierarchical component of the models, whereas the second refers to their dual-channel aspect. Obviously, the notion of independence of the first-order causal chain is also present in dual-channel SDT models. One of the consequences of using non-dual channel SDT to model type I and type II decisions is that, when there is no type I sensitivity, then there is no type II sensitivity: when there is no signal to discriminate between the presence or absence of a stimulus, or between two stimuli, there should in principle be no signal to base one's subjective rating on – something which, in the context of sensory metacognition, is at least plausible. However, Scott *et al*. [44] recently demonstrated why a model of metacognition should exhibit such independence. Specifically, they showed, in an artificial grammar learning (AGL) task, that participants could perform better than

chance in expressing judgments about their own performance (type II decisions) in spite of the fact that their performance (type I discrimination) was actually at chance! Such findings have two implications. First, strictly first-order and hierarchical models cannot account for such dissociations, which is suggestive that only dual-channel models have enough generality. Second, such findings support the idea that the information contained in the first-order network can be used in different, perhaps orthogonal decision criteria. Our hybrid-hierarchical comparator models do precisely that, in that they use the prediction error of the first-order network in a different way for first- and second-order decisions. In particular, while the first-order network takes its decisions based on the performance error (the standard SSE), the second-order network's decisions are based on a more detailed *pattern* representation of the first-order error. Thus, the second-order network learns to redescribe the error committed by the first-order network explicitly, as a pattern of activation rather than as a scalar signal. This is what enables it to leverage information that may not be captured by the first-order error. In principle, this might reflect the fact that, even if a first-order decision is predominantly subject to bias without any discriminative sensitivity, there is still enough information in the first-order performance signal in order to detect when one is wrong and when right in a discrimination task. In other words, the second-order network has a finer grained access to the first-order error, precisely because it can "look at" the error by representing it as a (potentially manipulable) pattern of activation, rather than just use it to guide output, as the first-order network does. In light of Scott *et al*.'s [44] data, this would mean that, even though the overall first-order error with respect to string grammaticality cannot be used to distinguish between strings in a type I task, the *way in which* those strings elicit errors is detectable by the second-order system, and hence reflected in above-chance type II judgments.

4. An SDT analysis of the hybrid metacognitive model

Our simulations were able to successfully account for the pattern of associations and dissociations between performance and confidence (or wagering) observed in the Iowa Gambling task, in an Artificial Grammar Learning task, and in blindsight. Here, we sought to analyse the hybrid model's performance in terms of SDT. Thus, we performed SDT analyses on the performance of the network in the Artificial Grammar Learning Task and in blindsight. ([2]; electronic supplementary material).

In the Artificial Grammar Learning Task simulation, the first-order network was trained, as in Persaud *et al*. [45], to discriminate grammatical from non-grammatical strings of letters, while the second-order network was trained to produce wagers on the first-order network's decisions. We showed [2] how the model was able to capture the patterns of associations and dissociations

between classification performance and wagering in the two conditions (implicit and explicit) tested by Persaud *et al*. [45].

Here, to analyse the model's performance using SDT, we replicated our original simulations, inserting a test block – instances of new grammatical strings and of non-grammatical strings – after every block of the learning phase and not only after the third (implicit condition) and the twelfth (explicit condition) block as it was the case in the original study. This small modification of the simulation setup allowed us to capture the networks' performance at every step during the learning phase (**figure 3*a***). As expected, type I sensitivity d'_1 steadily increases from 0 to a maximum value through learning, reflecting a progressively larger proportion of hits – correct discriminations of the new grammatical strings – than of false alarms – incorrect discriminations of the ungrammatical strings. In addition, networks tend to lose their initial conservative bias (type I c) as their knowledge develops. At the end of the learning phase, the neural networks end up with perfect knowledge of the grammar, as suggested by a high type I sensitivity and a null type I criterion. Type II sensitivity and criterion follow roughly the same pattern, although d'_2 does not increase as much as d'_1 and although c_2 here appears to already start from a neutral value (but higher initial criterion values were sometimes obtained in other simulations). As a reminder, the second-order network had already been trained in a pre-training phase and no more updates of its internal weights occurred afterwards, that is, during the actual learning phase. Thus, the second-order network behaves as a simple observer of the first-order network's knowledge and yet, its type II performance improves just as well through the learning phase, as reflected by a greater sensitivity and a neutral bias at the end of the task. Finally, by comparing type I and type II measures on the figures, one may notice that objective performance seems to have shaped subjective performance in this simulation, just as one would have predicted from a purely hierarchical architecture.

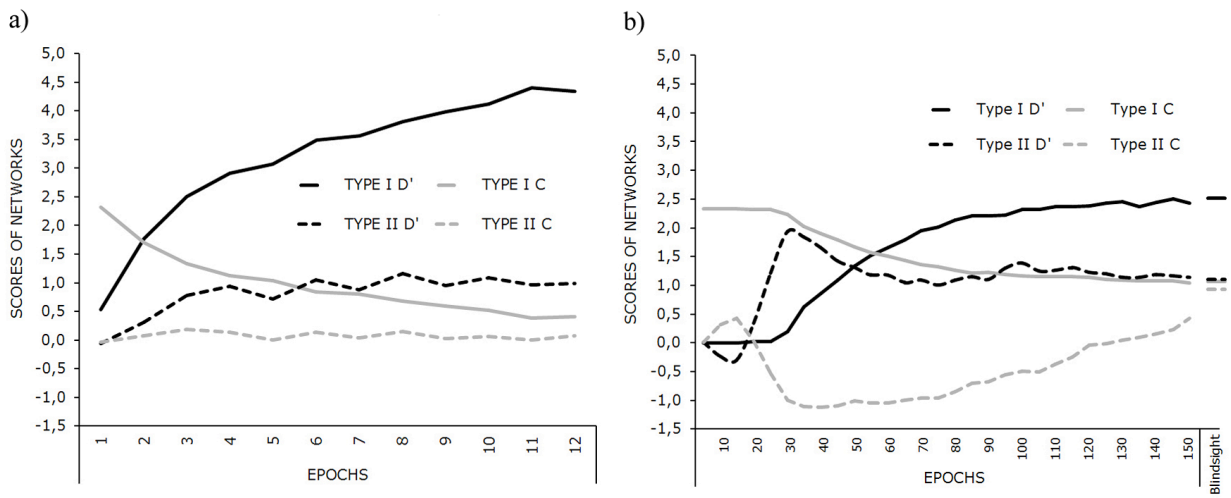

Figure 3. SDT analyses of the model's performance. A (type I and type II scores) chart reflects both objective and subjective measures in the AGL simulation ; B (type I and type II scores) chart reflects both objective and subjective measures in the blindsight simulation. (Data points for blindsight test type I d' and type II c overlap)

Our second implementation of the hybrid model was dedicated to capturing blindsight. In their Blindsight experiment, Persaud *et al*. [45] showed that blindsight subject GY (i.e., a patient who, under specific circumstances, makes visual discriminations in the absence of visual awareness), when presented with sub-threshold stimuli in his blind field, displayed above chance localisation performance but failed to maximise his earnings through wagering, suggesting that he was not always aware of the knowledge involved in his decisions for stimulus localisation. However, for supra-threshold stimuli (both in normal and blind fields), GY maximised performance as well as earnings. We successfully simulated these results [2] by pre-training the networks to discriminate amongst arbitrary positions of a stimulus and to simultaneously place wagers on their own performance. The distinction between supra-threshold and sub-threshold blindsight vision was introduced during a subsequent testing phase, in which the networks classified the patterns they had previously been presented with (supra-threshold), as well as degraded versions of these patterns in which stimulus-to-noise ratio was manipulated by increasing the noise level (sub-threshold). Here, we look at how the model's performance develops over time, and at how the model accounts for blindsight in light of Persaud's data.

To track the model's performance over time, we used the same procedure as for the AGL simulation, inserting test blocks after each block of the pre-training phase. We thus captured the networks' objective and subjective performance through the pre-training phase – results at the 150th block reflecting one's normal performance in a standard subliminal detection task –, as well as in a post-test blindsight condition for which the level of background noise in input was raised (see **figure 3*b***). Only after a short time of adaptation—the required time for the networks to learn to see anything, which may end around block 30 in the pre-training phase—type I performance seems to evolve perfectly normally. With training, d'_1 starts to increase, as the networks progressively become able to discriminate between noise and signals. However c_1 never reaches the null value, indicating the maintenance of a conservative policy. This, of course, is due to the fact that a few of the stimuli are displayed below the noise threshold and hence cannot be discriminated properly by the networks. Keeping a conservative bias thus prevents the networks from exhibiting too high a rate of false alarms. By contrast, type II scores seem rather peculiar. By the time the networks "learn to see", type II d' has reached its maximum value, and type II c is at its lowest, that is,

second-order networks have acquired a very high sensitivity but also a very liberal bias. One might think that they are somehow fully "open-minded", which pays off since subjective performance overrides the lack of objective knowledge in this case. Following this phase, type II sensitivity returns to a more moderate value while the criterion's slope tends towards a conservative value, as if bounded again by type I knowledge. Finally, type II scores in the post-test blindsight situation confirm our earlier findings [2], that is, a preserved sensitivity but a highly conservative bias. Whereas our overall results match the general findings by Persaud *et al*. [45], this criterion-setting account of blindsight diverges from Persaud *et al*.'s data, which suggest a decreased sensitivity, and not a criterion setting problem was underlying the failure to optimise wagering. However, Overgaard *et al*. [46,47], showed that this decreased-sensitivity account is linked to use of dichotomous measures such as the high vs. low wagers used by Persaud *et al*., whereas use of more graded measures reveals that in fact sensitivity *is* preserved but that patients use a very conservative criterion, which is what our current analysis suggests as well, and what others propose in this issue [11].

Our analyses thus highlight the hybrid character of the model. Indeed, in the AGL simulation, type II performance directly depends upon type I performance, whereas in the blindsight simulation, the second order network is able to build relevant meta-knowledge despite the first-order network's poor performance.

In closing, we should stress that the models we have presented have substantial limitations. Two such limitations are worth highlighting. The first is that the models fail to be dynamical. Responses are computed in a single time step, whereas we envision the relevant type I and type II processes as unfolding over time. The second is that the models fail to be recurrent: The meta-representations developed in the second-order network cannot influence the representations developed in the first-order network. Going beyond these two limitations is important for the following reason: When responses take time to be computed by a first-order network that contains multiple levels (e.g., 6 or 7 layers of hidden units), the second-order network may actually, were it able to influence the states of the first-order network, compute or at least bias the appropriate type I response even before the first-order network has completed its own processing. In other words, the second-order network would then be able to *predict* future states of the output layer of the first-order network. This would capture a central idea in our framework, namely that the brain continuously learns to predict the consequences of activity in one region on activity on other regions (what we call the "inner loop", see below). Augmenting our models with the necessary computational mechanisms will require using different, fully recurrent, dynamical learning algorithms.

5. Learning to be conscious: Metacognition as Radical Plasticity

What are the implications of this approach to metacognition as a dynamic representational re-description process? First, this approach suggests that metacognition (and hence, consciousness) takes time, at different time scales, that is, over a single trial, over learning, and over development. Second, this approach suggests that metacognition, far from being mere filtering as perhaps suggested by SDT, is an active, trained construction process. Recent work supports the idea that one can train people to gain conscious access to their own representations. For instance, participants can be trained to improve their performance in subliminal perception tasks [48], aversive learning can teach people to make novel olfactory distinctions [49], and imposing a deadline on simultaneous type I and type II ratings interfered with the degree to which participants were able to identify their correct responses [33] (interestingly, type I performance was also affected, but only on those trials for which people had claimed to be sure, suggesting that disruption of this metacognitive signal affects lower-level processing). It has been suggested [43] that gradual learning of (type II) precision estimates over a certain amount of time is particularly useful "in situations where the causes of perceptual evidence may change unpredictably over time, and as such may provide a better account of the sort of fluid, ongoing sensorimotor integration that characterises everyday activities such as riding a bicycle." Indeed, the creation of a conscious experience of the world may protect us and our brain from piecemeal and unpredictable sensory input.

Second, we would instead like to suggest that metacognition is but an instance of a larger class of representational re-description processes that, as stated before, occur unconsciously and automatically. From this perspective, the brain is continuously and unconsciously learning to anticipate the consequences of action or activity on itself, on the world, and on other people (see below for elaborations on the latter two). There is considerable evidence for such hierarchical predictive mechanisms in the human brain [50], through which the brain continuously attempts to minimise "surprise" or conflict by anticipating its own future activity based on learned priors. Through these predictive mechanisms, the brain develops systems of meta-representations that characterise and qualify the target first-order representations. Such learned re-descriptions, enriched by the emotional value associated with them, form the basis of conscious experience. Learning and plasticity are, thus, central to metacognition and consciousness, to the extent that experiences only occur in experiencers that have learned to know that they possess certain first-order states and that have learned to care more about certain states than about others. Cleeremans [19,51] has termed this view the "Radical Plasticity Thesis." While this paper is concerned primarily with meta-representation as a prerequisite for consciousness, this "caring about" aspect is equally crucial to

our model of consciousness, in that the knowledge that resides in those meta-representations (i.e., the knowledge about the first-order representations) has to have relevance for the organism. It has to matter to an organism whether the first-order state is A or B. Such relevance may be related to prospective error monitoring [43], or may be related to motivational and emotional components.

The idea that predictive re-description processes take place unconsciously can in fact be argued to form the core of the Higher-Order Thought (HOT) Theory of consciousness [21], according to which a representation is a conscious representation when one is conscious *of* it. In other words, by HOT, it is in virtue of the occurrence of (unconscious) higher-order thoughts "that we are now conscious of some content," that the content becomes phenomenally conscious. This, we surmise, requires the ability for the agent to re-describe its own states to itself as suggested above. We further suggest that a system's ability to re-describe its own knowledge to itself minimally requires (1) the existence of recurrent structures that enable the system to access its own states, and (2) the existence of predictive models (meta-representations) that make it possible for the system to characterise and anticipate the occurrence of first-order states. Importantly however, here, and in contrast to HOT, such meta-representational models (1) may be local and hence occur anywhere in the brain, (2) can be sub-personal, (3) are subject, just like first-order representations, to learning and plasticity mechanisms and, hence, can themselves become automatic.

Note that the proposed metacognitive architecture instantiates the minimal requirements necessary to enable a cognitive system to distinguish between veridical perceptions and hallucinations (something a pure first-order system would be unable to do) and, more generally, to develop the metacognitive knowledge necessary to represent the manner in which its own first-order knowledge is held, that is, propositional attitudes (is this a belief? a hope? a regret?).

6. Beyond consciousness: Three predictive loops

As discussed above, the core idea of our proposal is that the brain is continuously and unconsciously learning to anticipate the consequences of action or activity on itself, on the world, and on other people. Thus we have three closely interwoven loops that link the brain with itself, with the world, and with other agents, all driven by the same prediction-based mechanisms (**figure 4**). A first, internal or "inner loop", involves the brain re-describing its own representations to itself as a result of its continuous and unconscious attempts of predicting how activity in one region influences activity in other regions. In other words: The brain does not know in and of itself that there is a causal link between, say, activity in supplementary motor area and activity in primary

motor cortex, or between any other cerebral regions that are so causally linked. The knowledge contained in such feedforward links is thus implicit to the extent that there is no mechanism to access it directly. Our proposal, largely based on Friston's own analysis [52], is that the brain learns to render this implicit knowledge explicit by re-describing it through unconscious prediction-driven mechanisms. This is essentially the mechanism that our simulations attempt to capture.

The second loop is the familiar "perception-action loop". It results from the agent as a whole continuously predicting the consequences of its actions on the world.

The third loop is the "self-other loop", and links the agent with other agents, again using the exact same set of prediction-based mechanisms as involved in the other two loops. The existence of this third loop is constitutive of conscious experience, we argue, for it is in virtue of the fact that as an agent I am constantly attempting to model other minds that I am able to develop an understanding of myself. The processing carried out by the inner loop is thus causally dependent on the existence of both the perception-action loop and the self-other loop, with the entire system forming a "tangled hierarchy" (e.g., Hofstadter's concept of "a strange loop" [53]) of predictive internal models.

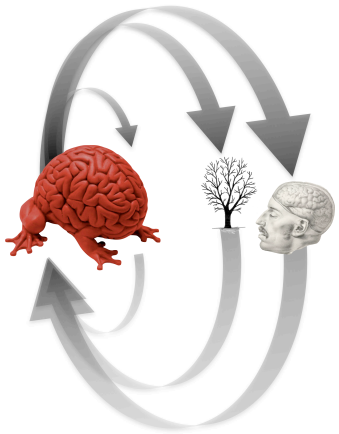

Figure 4. Three tangled loops (see text for details)

This third predictive loop thus extends beyond the agent into the social world. Consistently with the recent proposal by Carruthers [10], we surmise that understanding ourselves depends on ability to anticipate the consequences of our actions on other agents. Roughly, successfully anticipating how other agents will react to the actions we direct towards them requires that we have built internal models of how such agents will react to our actions. We assumed that such model building is enabled by automatic prediction of the other's actions in ongoing dynamic interaction [37, 54].

Recently Schilbach *et al*. [55,56] have suggested that, ontogenetically, becoming an expert in social cognition may crucially depend on social interaction while later competencies of more detached,

reflective social cognition (mirroring, mentalising) could be a result of reactivating the neural networks forged during social interactions (neural "re-use" [57]) and representationally re-describing these interaction-based capacities [1,19]. Crucial to this third loop, rather than seeing such a re-description as an internally generated, qualitatively different representation of discrete knowledge about the world, the "social" re-description is an ongoing learning process driven by increasingly complex interactive contexts, for instance when moving from dyadic to triadic interaction, which creates the possibility and need to communicate with respect to an external, third object or person [58]. In this light, e.g. language might not only be shaped by social interaction, but also the other way around, with the gradual development of language providing a scaffolding that allows implicit social know-how to develop in explicit social knowledge. Social context as a driving force for learning has, indeed, been recognised in language learning [59], child development [60], and social cognition [61]. Recently, it has also been suggested that mirror neurons might be the result of reinforcement learning [62–64]. Thus, the third loop conceptualises metacognition as resulting from predictive learning mechanisms that allow for agents to simultaneously learn about the environment as well as about their own internal representations. The ongoing re-descriptions that this entails, make for a potential explanation of how implicit precursors to mentalising (such as gaze following) later develop into explicit Theory of Mind and our capacity to consciously reason about others and ourselves [65].

Finally, the idea that all three loops may be subtended by the same mechanisms is supported by recent findings that metacognition, social interactions, and the processing of self-relevance all involve the recruitment of a common set of brain areas. Using an activation-likelihood estimation (ALE) approach, Schilbach *et al*. [66] recently investigated the statistical convergence of results from functional neuroimaging studies that had respectively targeted social cognition, emotional processing and unconstrained cognition, based on the assumption that a "common denominator" could exist in cognitive terms, consisting in a reliance on introspective processes, in particular prospective meta-cognition. By exploring the commonalities of the results from these three individual meta-analyses by means of a conjunction analysis, the authors were, indeed, able to provide empirical evidence for a shared neural network localised in dorso-medial prefrontal cortex and in the precuneus. These two regions are known to be critical hubs in the neurofunctional architecture of the human brain [67–73] and have been shown to be closely related to introspective ability [6]. Crucially, comparing the results of our conjunction analysis to the recent findings by Fleming at al. [6] demonstrates anatomical overlap both in PFC and the precuneus (**figure 5**).

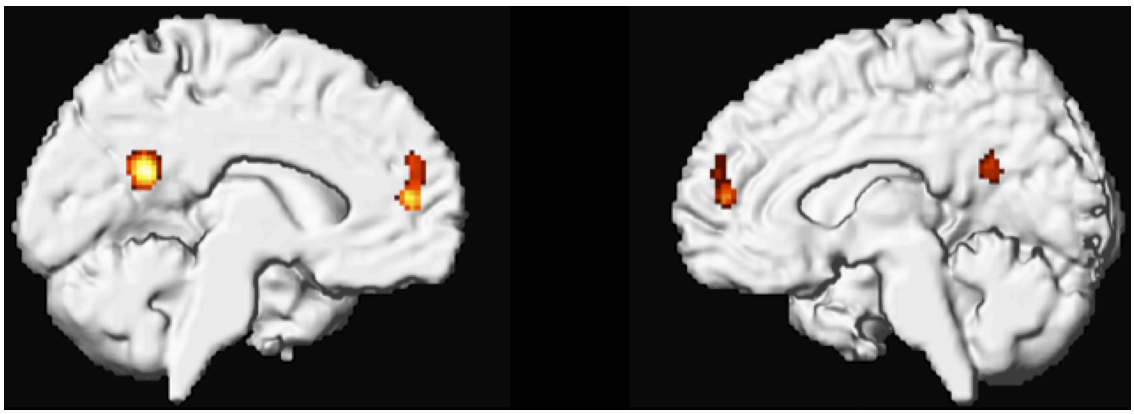

Figure 5. Comparison of studies of Fleming *et al*. [6] and Schilbach *et al*. [66]. Neuroanatomical overlap between areas related to individual differences in metacoginive abilities as reported in [6], and significant results of a triple conjunction analysis of ALE meta-analyses targeting functional neuroimaging studies of social cognition, emotional processing and unconstrained cognition [66]. Statistical convergence of functional neuroimaging results in dorso-medial prefrontal cortex and the precuneus are displayed on the surface view of the MNI single subject template. Taken from Schilbach *et al*. [66].

Interestingly, the two brain regions that appear to be involved both in social cognition and introspective or meta-cognitive processes are part of what has become known as the "default mode of brain function" [67]. We have recently argued that this convergence might be taken to suggest that the physiological baseline of the human brain, i.e. the default mode network, is related to a psychological baseline of social cognition [55]. Here, we extend this argument by suggesting that social interactions might enable introspective processes and conscious experience while relying on changes in the activity of the default mode network. Congruently, Carhart-Harris and Friston [74] have recently argued that the default mode network might realise the Freudian secondary process, i.e. the "mode of cognition of the ego", or in other words, normal waking consciousness. Strikingly, this analysis is rooted in a Bayesian perspective on the brain, which assumes that the brain uses internal hierarchical models to predict its sensory inputs and suggests that neural activity tries to minimise the ensuing prediction-error or (Helmholtzian) free energy [52,74]. Consistent with the proposal of key regions of the default mode network subserving introspective processes and social cognition and our claim that these abilities take time to develop, it has been found that connectivity within the Default Mode Network (DMN) develops through ontogeny [75,76]. Importantly, such developments hinge upon interactions with the environment and might be necessary to establish a balance between internally oriented cognition and engagement with the external world. Apart from the empirical evidence for an anatomical overlap of the brain regions relevant for introspection and social interaction, Carhart-Harris & Friston's account [74] can also be taken to suggest that all of the three loops, which we assume are relevant for metacognition, rely on similar neural

mechanisms, namely internal models that are used to predict network changes based either on sensory input or on endogenously generated activation.

7. Conclusion

Overall, our perspective is thus akin to the sensorimotor or enactive perspective [77] and to the general conceptual framework provided by forward modelling (e.g., [54]) in the sense that awareness is linked with knowledge of the consequences of our actions. Crucially, however, we extend the argument inwards (the inner loop) and further outwards (the self-other loop), and specifically towards social cognition (see also [78]). Our representations of ourselves are shaped by our history of interactions with other agents. Learning about the consequences of the actions that we direct towards other agents uniquely requires more sophisticated models of such other agents than when interacting with objects, for agents, unlike objects can react to actions directed towards them in many different ways as a function of their own internal state. A further important point here is that caretakers act as external selves during development, interpreting what happens to developing children for them, and so providing meta-representations where they lack. In this light, theory of mind can thus be understood as rooted in the very same mechanisms of predictive re-descriptions as involved when interacting with the world or with one self (see also [37]).

Thus we end with the following idea, which we call the "Radical Plasticity Thesis": The brain continuously and unconsciously learns not only about the external world and about other agents, but also about its own representations of both. The result of this unconscious learning is conscious experience, in virtue of the fact that each representational state is now accompanied by (unconscious learnt) meta-representations that convey the mental attitude with which the first-order representations are held. 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 re-description in the sense described by Karmiloff-Smith [1].

Acknowledgments

Bert Timmermans is supported by a European Commission Marie Curie Fellowship FP7-PEOPLE-IEF 237502 “Social Brain”. Leonhard Schilbach is supported by the Koeln Fortune Program of the Medical Faculty, University of Cologne and the Volkswagen Foundation. Axel Cleeremans is a Research Director with the National Fund for Scientific Research (F.R.S.-FNRS, Belgium).

References

1 Karmiloff-Smith, A. 1992 *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.

2 Pasquali, A., Timmermans, B. & Cleeremans, A. 2010 Know thyself: Metacognitive networks and measures of consciousness. *Cognition* **117**, 182-190. (doi 10.1016/j.cognition.2010.08.010)

3 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)

4 Scott, R. B. & Dienes, Z. 2008 The conscious, the unconscious, and familiarity. *J. Exp. Psychol. Learn.* **34**, 1264—1288. (doi 10.1037/a0012943)

5 Fleming, S. M. & Dolan, R. J., The neural basis of accurate metacognition, *Phil. Trans. R. Soc. B.*, this issue.

6 Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J. & Rees, G. 2010 Relating introspective accuracy to individual differences in brain structure. *Science* **329**, 1541–1543. (doi 10.1126/science.1191883)

7 Rounis, E., Maniscalco, B., Rothwell, J., Passingham, R. & Lau, H. 2010 Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neurosci.* **1**, 165–175. (doi 10.1080/17588921003632529)

8 Lau, H. & Rosenthal, D. 2011 Empirical support for higher-order theories of conscious awareness. *Trends Cogn. Sci.* **15**, 365–373. (doi 10.1016/j.tics.2011.05.009)

9 Pleskac, T. J. & Busemeyer, J. R. 2010 Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychol. Rev.* **117**, 864–901. (doi 10.1037/a0019737)

10 Carruthers, P. 2009 How we know our own minds: The relationship between mindreading and metacognition. *Behav. Brain Sci.* **32**, 121—138. (doi 10.1017/S0140525X09000545)

11 Ko, Y. & Lau, H., A detection theoretic account of blindsight suggests a link between conscious perception and metacognition, *Phil. Trans. R. Soc. B.*, this issue.

12 Dennett, D. C. 1991 *Consciousness explained*. Boston, MA: Little, Brown and Co.

13 Dennett, D. C. 2001 Are we explaining consciousness yet? *Cognition* **79**, 221–237. (doi 10.1016/S0010-0277(00)00130-X)

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

15 Dehaene, S., Kerszberg, M. & Changeux, J.-P. 1998 A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl. Acad. Sci. USA* **95**, 14529–14534. (doi 10.1073/pnas.95.24.14529)

16 Dehaene, S. & Naccache, L. 2001 Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition* **79**, 1–37. (doi 10.1016/S0010-0277(00)00123-2)

17 Clark, A. & Karmiloff-Smith, A. 1993 The Cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind Lang.* **8**, 487-519.

18 Carruthers, P. 2000 *Phenomenal consciousness: A naturalistic theory*. Cambridge, UK: Cambridge University Press.

19 Cleeremans, A. 2011 The Radical Plasticity Thesis: How the brain learns to be conscious. *Front. Psychology* **2**, 86. (doi 10.3389/fpsyg.2011.00086)

20 Kriegel, U. 2009 *Subjective consciousness: A self-representational theory*. Oxford, UK: Oxford University Press.

21 Rosenthal, D. M. 2005 *Consciousness and Mind*. Oxford, UK: Clarendon Press

22 Rosenthal, D. M., Higher-order awareness, misrepresentation, and function, *Phil. Trans. R. Soc. B.*, this issue.

23 Lau, H. 2008 A higher-order Bayesian decision theory of consciousness. In *Models of brain and mind. Physical, computational and psychological approaches* (eds R. Banerjee & B. K. Chakrabarti), pp. 35—48, *Prog. Brain Res.*, Elsevier. (doi 10.1016/S0079-6123(07)68004-2)

24 Clifford, C. W. G., Arabzadeh, E. & Harris, J. A. 2008 A good bet to measure awareness. *Trends Cogn. Sci.* **1**. (doi 10.1016/j.tics.2008.02.011)

25 Snodgrass, M. T. U., Bernat, E. & Shevrin, H. 2004 Unconscious perception: A model-based approach to method and evidence. *Percept. Psychophys.* **66**, 846–867. (doi 10.3758/BF03194978)

26 Maniscalco, B. & Lau, H. 2012 A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn.* **21**, 422–430. (doi 10.1016/j.concog.2011.09.021)

27 Maniscalco, B. & Lau, H. Under review. The signal processing architecture underlying subjective reports of sensory awareness.

28 Scott, R. B. & Dienes, Z. 2010 Fluency does not express implicit knowledge of artificial grammars. *Cognition* **114**, 372—388. (doi 10.1016/j.cognition.2009.10.010)

29 Galvin, S. J., Podd, J. V., Drga, V. & Whitmore, J. 2003 Type II tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychon. B. Rev.* **10**, 843–876.

30 Dienes, Z., Altmann, G. T. M., Kwan, L. & Goode, A. 1995 Unconscious knowledge of artificial grammars is applied strategically. *J. Exp. Psychol. Learn.* **21**, 1322—1338. (doi 10.1037/0278-7393.21.5.1322)

31 Block, N. 2007 Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behav. Brain Sci.* **30**, 481–99. (doi 10.1017/S0140525X07002786)

32 Sperling, G. 1960 The information available in brief visual presentation. *Psychol. Monogr.* **74**, 1–29. (doi 10.1037/h0093759)

33 Mealor, A. & Dienes, Z. 2012 No-loss gambling shows the speed of the unconscious. *Conscious. Cogn.*, **22**, 228–237. (doi 10.1016/j.concog.2011.12.001)

34 de Gardelle, V., Sackur, J. & Kouider, S. 2009 Perceptual illusions in brief visual presentations. *Conscious. Cogn.* **18**, 569–577. (doi 10.1016/j.concog.2009.03.002)

35 Kouider, S. & Dupoux, E. 2004 Partial awareness creates the "'illusion'" of subliminal semantic priming. *Psychol. Sci.* **15**, 75–81. (doi 10.1111/j.0963-7214.2004.01502001.x)

36 Kouider, S., de Gardelle, V., Sackur, J. & Dupoux, E. 2010 How rich is consciousness? The partial awareness hypothesis. *Trends Cogn. Sci.* **14**, 301–307. (doi 10.1016/j.tics.2010.04.006)

37 Frith, C. D. 2007 *Making up the mind: How the brain creates our mental world*. Oxford, UK: Blackwell.

38 Noë, A. 2004 *Action in perception*. Cambridge, MA: MIT Press.

39 Noë, A. 2009 *Out of our heads: Why you are not your brain and other lessons from the biology of consciousness*. London, UK: MacMillan.

40 Sutton, R.S. & Barto, A.G. 1998 *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

41 Hinton, G. E. 1986 Learning distributed representations of concepts. In *Proc. 8th Annu. Conf. Cogn. Sci. Soc., Amherst, MA, August 1986*, pp. 1–12. Hillsdale, NJ: Lawrence Erlbaum.

42 Dienes, Z. & Perner, J. 1996 Implicit knowledge in people and connectionist networks. In *Implicit cognition* (ed G. Underwood), pp. 227-256. Oxford, UK: Oxford University Press.

43 Yeung, N. & Summerfield, C., Metacognition in human decision making: Confidence and error monitoring, *Phil. Trans. R. Soc. B.*, this issue.

44 Scott, R. B., Dienes, Z. & Seth, A. K. 2011 Higher-order awareness without first-order accuracy: implications for models of consciousness. Paper presented at the 15th Annual Meeting of the Assoc. for the Scientific Study of Consciousness, Kyoto, Japan, June 2011. (See http://www.theassc.org/files/assc/Program_201106010_update.pdf for the abstract.)

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

46 Overgaard, M., Fehl, K., Mouridsen, K., Bergholt, B. & Cleeremans, A. 2008 Seeing without seeing? Degraded conscious vision in a blindsight patient. *PLoS ONE* **3**, e3028. (doi 10.1371/journal.pone.0003028)

47 Overgaard, M. & Sandberg, K., Kinds of access: Different methods for report reveal different kinds of metacognitive access, *Phil. Trans. R. Soc. B.*, this issue.

48 Schwiedrzik, C. M., Singer, W. & Melloni, L. 2008 Sensitivity and perceptual awareness increase with practice in metacontrast masking. *J. Vis.* **9**, 18. (doi 10.1167/9.10.18)

49 Li, W., Howard, J. D., Parrish, T. B. & Gottfried, J. A. 2008 Aversive learning enhances perceptual and cortical discrimination of indiscriminable odor cues. *Science* **319**, 1842—1845. (doi 10.1126/science.1152837)

50 Friston, K. 2008 Hierarchical models in the brain. *PLoS Comput. Biol.* **4**:e1000211. (doi 10.1371/journal.pcbi.1000211)

51 Cleeremans, A. 2008 Consciousness: The radical plasticity thesis. In Models of brain and mind: Physical, computational and psychological approaches (eds R. Banerjee & B. K. Chakrabarti), *Prog. Brain Res.* **168**, 19-33. (doi 10.1016/S0079-6123(07)68003-0)

52 Friston, K. 2006 A free energy principle for the brain. *J. Physiology-Paris* **100**, 70—87. (doi 10.1016/j.jphysparis.2006.10.001)

53 Hofstadter, D.R. 2007 *I am a strange loop*. New York, NY: Basic Books

54 Wolpert, D. M., Doya, K. & Kawato, M. 2003 A unifying computational framework for motor control and social interaction. *Phil. Trans. R. Soc. B* **358**, 593–602. (doi 10.1098/rstb.2002.1238)

55 Schilbach, L., Eickhoff, S. B., Rotarska-Jagiela, A., Fink, G. R. & Vogeley, K. 2008 Minds at rest? Social cognition as the default mode of cognizing and its putative relation to the "default system" of the brain. *Conscious. Cogn.* **17**, 457—467.

56 Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T. & Vogeley, K. In press. Toward a second-person neuroscience. *Behav. Brain Sci.*, target article.

57 Anderson, M. L. 2010 Neural reuse: A fundamental organizational principle of the brain. *Behav. Brain Sci.* **33**, 245–266. (doi 10.1017/S0140525X10000853)

58 Carpendale, J. E. M. & Lewis, C. 2004 Constructing an understanding of mind: The development of children's social understanding within social interaction. *Behav. Brain Sci.* **27**, 79—150. (doi 10.1017/S0140525X04000032)

59 Kuhl, P. K. 2007 Is speech learning "gated" by the social brain? *Dev. Sci.* **10**, 110–120. (doi 10.1111/j.1467-7687.2007.00572.x)

60 Reddy, V. 2008 *How infants know minds*. Cambridge, MA: Harvard University Press.

61 Becchio, C., Sartori, L. & Castiello, U. 2010 Toward you: The social side of actions. *Curr. Dir. Psychol. Sci.* **19**, 183—188. (doi 10.1177/0963721410370131)

62 Catmur, C., Walsh, V. & Heyes, C. 2009 Associative sequence learning: The role of experience in the development of imitation and the mirror system. *Phil. Trans. R. Soc. B* **364**, 2369–2380. (doi 10.1098/rstb.2009.0048)

63 Heyes, C. 2010 Where do mirror neurons come from? *Neurosci. Biobehav. R.* **34**, 575–583. (doi 10.1016/j.neubiorev.2009.11.007)

64 Triesch, J., Jasso, H. & Deak, G. O. 2007 Emergence of mirror neurons in a model of gaze following. *Adapt. Behav.* **15**, 149–165. (doi 10.1177/1059712307078654)

65 Frith, C. D. & Frith, U. 2012 Mechanisms of social cognition. *Annu. Rev. Psychol.* **63**, 287-313. (doi 10.1146/annurev-psych-120710-100449)

66 Schilbach, L., Bzdok, D., Timmermans, B., Fox, P.T., Laird, A.R., Vogeley, K., Eickhoff, S.B. 2012 Introspective minds: Using ALE meta-analyses to study commonalities in the neural correlates of emotional processing, social & unconstrained cognition. *PLoS ONE* **7**, e30920. (doi 10.1371/journal.pone.0030920)

67 Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A. & Shulman, G. L. 2001 A default mode of brain function. *Proc. Natl. Acad. Sci. USA* **98**, 676—682. (doi 10.1073/pnas.98.2.676)

68 Cavanna, A. E. & Trimble, M. R. 2006 The precuneus: A review of its functional anatomy and behavioural correlates. *Brain* **129**, 564—583. (doi 10.1093/brain/awl004)

69 Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J. & Sporns, O. 2008 Mapping the structural core of human cerebral cortex. *PLoS Biol.* **6**, e159. (doi 10.1371/journal.pbio.0060159)

70 Buckner, R. L., Andrews-Hanna, J. R. & Schacter, D. L. 2008 The brain's default network: Anatomy, function, and relevance to disease. *Ann. N. Y. Acad. Sci.* **1124**, 1—38. (doi 10.1196/annals.1440.011)

71 Fransson, P. & Marrelec, G. 2008 The precuneus/posterior cingulate cortex plays a pivotal role in the default mode network: Evidence from a partial correlation network analysis. *NeuroImage* **42**, 1178—1184. (doi 10.1016/j.neuroimage.2008.05.059)

72 Margulies, D. S., Vincent, J. L., Kelly, C., Lohmann, G., Uddin, L. Q., Biswal, B. B., Villringer, A., Castellanos, F. X., Milham, M. P. & Petrides, M. 2009 Precuneus shares intrinsic functional architecture in humans and monkeys. *Proc. Natl. Acad. Sci. USA* **106** 20069—20074. (doi 10.1073/pnas.0905314106)

73 Glahn, D. C., Winkler, A. M., Kochunov, P., Almasy, L., Duggirala, R., Carless, M. A., Curran, J. C., Olvera, R. L., Laird, A. R., Smith, S. M., Beckmann, C. F., Fox, P. T. & Blangero, J. 2010 Genetic control over the resting brain. *Proc. Natl. Acad. Sci. USA* **107**, 1223—1228. (doi 10.1073/pnas.0909969107)

74 Carhart-Harris, R. L. & Friston, K. J. 2010 The default-mode, ego-functions and free-energy: A neurobiological account of Freudian ideas. *Brain* **133**, 1265—1283. (doi 10.1093/brain/awq010)

75 Fair, D. A., Cohen, A. L., Dosenbach, N. U. F., Church, J. A., Miezin, F. M., Barch, D. M., Raichle, M. E., Petersen, S. E. & Schlaggar B. L. 2008 The maturing architecture of the brain's default network. *Proc. Natl. Acad. Sci. USA* **105**, 4028–32. (doi 10.1073/pnas.0800376105)

76 Kelly, A. M., Di Martino, A., Uddin, L. Q., Shehzad, Z., Gee, D. G., Reiss, P. T., Margulies, D. S., Castellanos, F. X. & Milham, M. P. 2009 Development of anterior cingulate functional connectivity from late childhood to early adulthood. *Cereb. Cortex* **19**, 640–657. (doi 10.1093/cercor/bhn117)

77 O'Regan, J. K. & Noë, A. 2001 A sensorimotor account of vision and visual consciousness. *Behav. Brain Sci.* **24**, 883—917. (doi 10.1017/S0140525X01000115)

78 Graziano, M. S. A. & Kastner, S. 2011 Human consciousness and its relationship to social neuroscience: A novel hypothesis. *Cognitive Neurosci.* **2**, 98—113. (doi 10.1080/17588928.2011.565121)

Supplementary Material

Simulation 1 – Blindsight

General architecture. The architecture of the networks is depicted in Fig. 2b (main article). The first-order network was a backpropagation autoassociator, consisting of a 100-unit input layer, itself connected to a layer of 60 hidden units, which were finally connected to a 100-unit output layer. The second-order network was a feedforward backpropagation network, the input of which consisted of a 100-unit comparison matrix, representing the match between the first-order input and output layers. Each of these units calculated the difference between each pair of corresponding input and output units in the first-order network. For instance, if first-order input unit #6(i) had an activation of 0.6 and the first-order output unit #6(o) had an activation of 0.7, the activation of second-order comparison unit #6(c) would be -0.1. It is important to note that these units were thus representing the error of prediction of each first-order output unit. In other words, instead of using this error to drive learning by means of backpropagation (in the first-order network only), the error is represented as an *activation pattern*, which can be used to drive any secondary task, in this case wagering (in output of the second-order network). The 100 comparison units were connected to two output units representing a high and a low wager. Initial connection weights were between -1.0 and 1.0 for the first-order network and between 0.0 and 0.1 for the second-order output (wager). The comparator weights of the second-order network were set to 1.0 for the connections from the first-order input layer, and to -1.0 for the connections from the first-order output layer.

Wagering specifics. Importantly, the comparator weights between the first-order network and the second-order network (1.0 and -1.0), were not subject to learning. Only the first-order network's weights were modified by learning, as were the second-order network's weights between the comparator units and the wagering units. This allowed for the second-order network to learn to wager at the same time the first-order network learned the discrimination task, while wagering nevertheless remained independent from the particular activation patterns of the first-order network. Instead, the second-order network learned to base its wagers on the degree to which a specific presented pattern (first-order input) corresponded to the internal signal it elicited (first-order output).

Patterns. Network pre-training set consisted of 200 patterns, half of which represented mere noise (unit activations chosen random between 0.0 and 0.02), and half of which represented a possible stimulus (for each pattern, 99 out of 100 units had an activation between 0.0 and 0.02, and one unit had an activation between 0.0 and 1.0. We did not always use an activation of 1.0, as we wanted to create situations in which the network would be unable or have great difficulty distinguishing and localizing the stimulus (subthreshold condition), and hence would learn to wager low in such cases. As the first-order network is contrastive, for "stimulus present" trials the first-order target patterns consisted of 99 units set to 0.0, and one unit set to 1.0, whereas the target patterns for the "stimulus absent" trials consisted of 100 units set to 0.0. A discrimination threshold was applied on the first-order network, so that at least one unit had to be above 0.5 for the pattern to be distinguished from mere noise. The target of the second-order network consisted either of the "high wager" unit set to 1.0 and the "low wager" unit set to 0.0 if (a) the first-order network localized a stimulus (threshold set such that the output unit activation > 0.5) and the stimulus was presented (input contained the stimulus), or (b) the first-order network did not recognize any stimulus and there was none presented; either the second-order network's target was "high wager" = 0.0 and "low wager" = 1.0, if (c) the first-order network recognized a stimulus that was not presented (hallucination), or (d) the first-order network failed to recognize a stimulus that was presented (blindness).

Pre-training. Each network was pre-trained on the 200 patterns for 150 epochs. first-order network's learning rate was 0.9, while second-order network's learning rate (between comparator and wagering units) was set to 0.1. All units' momentum was 0.0, while temperature was 1.0. This pre-training allows the second-order network to learn the degree to which it can trust what the first-order network recognizes. It corresponds to a healthy brain learning to make distinctions between what it does and doesn't see.

Testing. We tested the network in three different conditions. Each of these represented a different way of manipulating the signal-to-noise ratio, and hence a different degree and nature of blindness. First, under "Suprathreshold stimulus" condition, the networks were presented with the same set of 200 patterns as in pre-training. Second, networks were tested under two different blindness conditions. In the "Subthreshold stimulus" condition (representing the Blindsight condition), blindness was simulated by adding noise (+ 0.0012) to every input of the first-order network, except the one representing the stimulus. In the "Low Vision" condition, blindness was simulated by

reducing the activation of the stimuli (instead of varying from 0.0 to 1.0, they varied between 0.0 and 0.3).

Additional Results. As shown in Table 1, simulating blindness by reducing signal strength leads to very different results than by adding noise, and does not result in blindsight. For "Low Vision", we observe a situation unrelated to blindsight but rather reflecting blindness in general, in which the networks completely fail to show any discriminatory ability (50.3% correct) but are still able to wager well above chance (advantageous wagers in 69.6% of the trials). This effect is a consequence of the first-order network's discrimination threshold, which is difficult to attain in this "Low Vision" condition, but which has however no consequence on the spreading of input and output activations to the second-order network. In other words, the first-order network will hardly detect anything, causing the second-order network to lose confidence in the first-order network for the "stimulus present" patterns. Therefore the second-order network wagers low every time it feels a stimulus should have been detected. Thus, although the first-order network does not detect any stimulus, wagers are warranted and wagering performance remains advantageous. Simply said, the metacognitive network "knows" that it is blind.

Robustness of the results. The blindsight simulations are the only ones that depend on the specific choices made for the different parameters (learning rates, epochs, noise), as we wished to reproduce situations of blindsight and of blindness without resulting to more extreme measures such as, for instance, cutting the connections. We included a different possibility of simulating blindness, to illustrate the impact of such choices.

Low Vision	**Correct**	**Incorrect**	**Total**
High Wager	<u>31,83</u>	11,93	43,77
Low Wager	18,43	<u>37,80</u>	56,23
Total	50,27	49,73	100,00

Table 1. Additional results of the Blindsight simulation. Here blindness is simulated by manipulating the signal-to-noise ratio through decreased stimulus activation. Optimal wagers are underlined.

Simulation 2 – Artificial Grammar Learning Task

General architecture. The architecture was largely similar to that used in the first simulation, and is depicted in Fig. 2b. The first-order network was a backpropagation autoassociator, consisting of a 48-unit input layer (representing a string of minimum 3 items and maximum 8, each being one of 5 possible letters, constructed according to the selected artificial grammar), connected to a layer of 40 hidden units, who were connected to the 48-unit output layer. The input of the second-order network consisted of a 48-unit comparison matrix, representing, as in Simulation 1, the difference between the first-order input and output activations. These units were connected to two output units representing a high and a low wager. Initial connection weights were between -1.0 and 1.0 for the first-order network and between 0.0 and 0.1 for the second-order output. The comparator weights of the second-order network were set to 1.0 for the connections from the first-order input layer and to -1.0 for the connections from the first-order output layer.

Wagering specifics. As in Simulation 1, the comparator weights were not subject to learning, for the exact same reasons. The basis of the wagering mechanism was identical.

Patterns. We used 80 random patterns for pre-training. For actual training and testing, we constructed 75 patterns according to a specific Grammar A, and 30 patterns according to a Grammar B, each pattern representing a string from three to eight letters. All input/target patterns consisted of activations of either 1.0 or 0.0. A specific winner-take-all mechanism was added to the output layer in order to select the successive letters of the string that the first-order network recognized. The target of the second-order network was determined in a way identical to Simulation 1.

Pre-training and training. The two network sets were first subjected to *pre-training* on 80 *random* patterns for 60

epochs, allowing the second-order network to learn how to wager independently of *any* first-order task. In order to achieve this, half of the 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. 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*. During the actual *training* phase, only the first-order networks were trained again on 45 patterns of Grammar A. Of 30 networks, 15 were assigned to a "High Consciousness" condition and were trained for 12 epochs, while 15 networks in the "Low Consciousness" condition were trained for only 3 epochs. During their respective periods of training, first-order network's and second-order network's learning rates were set to 0.4, units' momentum was 0.5, and temperature was 1.0. The use of different learning phases for the first-order network and the second-order network in this simulation provides an illustration of first-order and second-order independency in the brain.

Testing. We tested all 30 networks on the same set of 60 patterns, consisting of 30 novel Grammar A patterns, and 30 Grammar B patterns.

Robustness of the results. The simulations do not depend on the specific choices made for the different parameters (learning rates, momentums, epochs), but should be manipulated one by one in order to maintain the generalization effect on the second-order knowledge.