

CHAPTER 4

A higher order Bayesian decision theory of consciousness

Hakwan C. Lau^{1,2,*}

¹Wellcome Trust Functional Imaging Laboratory, University College London, 12 Queen Square, London WC1N 3BG, UK
²Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford, OX1 3UD, UK

Abstract: It is usually taken as given that consciousness involves superior or more elaborate forms of information processing. Contemporary models equate consciousness with global processing, system complexity, or depth or stability of computation. This is in stark contrast with the powerful philosophical intuition that being conscious is more than just having the ability to compute. I argue that it is also incompatible with current empirical findings. I present a model that is free from the strong assumption that consciousness predicts superior performance. The model is based on Bayesian decision theory, of which signal detection theory is a special case. It reflects the fact that the capacity for perceptual decisions is fundamentally limited by the presence and amount of noise in the system. To optimize performance, one therefore needs to set decision criteria that are based on the behaviour, i.e. the probability distributions, of the internal signals. One important realization is that the knowledge of how our internal signals behave statistically has to be learned over time. Essentially, we are doing statistics on our own brain. This 'higher-order' learning, however, may err, and this impairs our ability to set and maintain optimal criteria for perceptual decisions, which I argue is central to perception consciousness. I outline three possibilities of how conscious perception might be affected by failures of 'higher-order' representation. These all imply that one can have a dissociation between consciousness and performance. This model readily explains blindsight and hallucinations in formal terms, and is beginning to receive direct empirical support. I end by discussing some philosophical implications of the model.

Keywords: consciousness; Bayesian; signal detection; fMRI

Introduction

This article describes a theoretical framework for characterizing perceptual consciousness. People receive information from the outside world through their sense organs, and produce actions in reaction to the external stimuli. However, the

brain seems to perform more than just the mechanical transformation of sensory inputs into motor outputs. Often, the person is also said to be subjectively and consciously aware of the objects of perception. I call this phenomenon 'perceptual consciousness', or sometimes 'consciousness' for short. Here I describe a model that formally characterizes the conditions under which this occurs.

Many theories have already been proposed on this topic, but the framework described here differs

*Corresponding author. Tel.: +44 (0)1865 271 444;
Fax: +44 (0)1865 310447; E-mail: hakwan@gmail.com

from them in an important aspect, which is that I do not treat consciousness as the same thing as superior information processing. In other words, I consider that, when the subject is consciously aware of the stimuli, the basic effectiveness of information processing is not necessarily higher. This may differ from common interpretations of most contemporary models of consciousness, which equate consciousness with global processing (Baars, 1988; Dehaene et al., 2003), system complexity (Tononi, 2004), or depth or stability of computation, etc. (Cleeremans, 2005). Commitment to these models often leads to the prediction that consciousness, as compared to the lack of it, will lead to some absolute advantage in terms of information processing. Specifically, within the context of perception, the prediction would be that when one consciously perceives something, rather than sensing it unconsciously, one is always better at identifying it or discriminating it from something else. Intuitively, this may seem plausible enough. In fact, this is often a tacit assumption in experimental studies of perceptual consciousness, even in studies that are conducted by scientists who are not committed to any specific model. This is reflected by the fact that many researchers (Rees et al., 2002) take forced-choice identification or discrimination performance as an index of consciousness: if the performance is high (hits, or high average accuracy), we consider the stimuli consciously perceived, and if the performance is low (misses, or near-chance average accuracy), we consider the stimuli not consciously perceived.

However, despite its ubiquity, the assumption that conscious perception is associated with high performance is unsupported by current empirical data (Lau, in press). In the author's opinion, the general question of whether consciousness plays any special function is still open to empirical investigation. Many have assumed that consciousness might be necessary in executive control or in the generation of spontaneous voluntary action, but recent studies revealed several surprising contradictions to these assumptions (Wegner and Wheatley, 1999; Wegner, 2003; Wegner et al., 2003; Dijksterhuis et al., 2006; Lau et al., 2006, 2007). Whereas none of these show that consciousness has no special function at all, at least they

remind us that we should be cautious in accepting theoretical speculations. Consciousness may be functionally less powerful than being assumed previously.

Specifically, within the context of visual perception, we have good reason to think that consciousness is not necessary for good performance in forced-choice detection or discrimination tasks (Lau, in press). This is due to the well-documented phenomenon of blindsight (Weiskrantz, 1986, 1999). After lesions to the primary visual area, blindsight patients report a lack of visual consciousness in the affected region of their visual field. However, when forced to make a decision as to whether something was presented in the region, or to discriminate between two stimuli such as gratings with different orientations, the patients performed well above chance, even though they said they were merely guessing. In some circumstances, they could even guess correctly ~80–90% of the time. This challenges the view that consciousness is the same as high basic effectiveness of information processing.

The foregoing considerations motivate the formulation of a new model that can allow for the dissociation between perceptual consciousness and the basic effectiveness of information processing. We will first discuss a standard method to characterize the basic effectiveness of information processing, which is signal detection theory, and then we consider how we can further extend it so that it could also characterize consciousness.

Signal detection theory

In studies of perception, the subject's performance is often characterized by using signal detection theory (Green and Swet, 1966; Macmillan and Creelman, 1991). Let us take the example in which the subject is presented with a visual stimulus in half of the trials, and a blank screen in the other half. The subject is required to say whether the stimulus is present or absent in each trial. According to the theory of signal detection, the subject's behaviour could be characterized by the detection sensitivity (d') of the subject and the criterion for detection (c). The former is a measure

of perceptual capacity and the latter reflects the decision strategy used. If we assume that there is an internal decision signal (e.g. firing rate in the visual cortex) with which the subject determines whether a stimulus is presented or not, one could construct the probability distribution function for the decision signal given that the stimulus is present, and for the decision signal given that the stimulus is absent (Fig. 1). The fact that the signal strengths for both conditions are reflected by probability distributions means that there is variability or fluctuation in the signal. One usually assumes that these distributions are Gaussians and of equal variance, and the mean for the “stimulus present” distribution is higher than the mean for the “stimulus absent” distribution.

On each trial/presentation, the subject has an internal signal of a particular strength, and has to decide based on this whether the stimulus is present. According to the theory, the subject sets a criterion (c) and responds “yes” (“there is a stimulus”) if the internal signal exceeds c , or responds “no” if the internal signal is lower than c . When c is low, we can say that the observer is adopting a liberal strategy (saying “yes” frequently). When c is high, we can say that the observer is adopting a conservative strategy

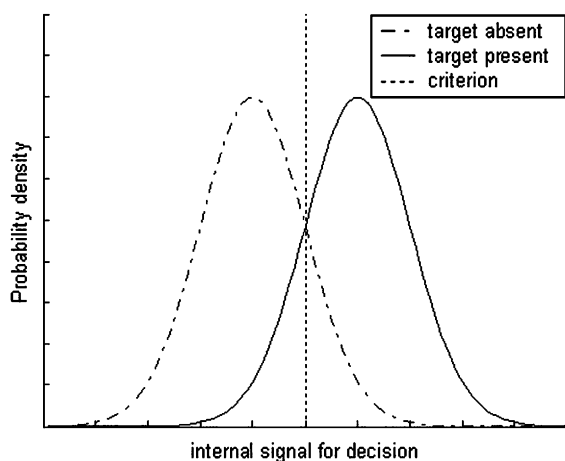


Fig. 1. Standard model of signal detection. Assuming that the target is present in 50% of trials, to perform optimal detection (i.e. to minimize errors), one would set the criterion at a point that best separates the two Gaussian distributions, that is right between their means.

(saying “no” frequently). So long as the two distributions overlap, the subject makes errors, because at the signal strength where the distributions overlap, sometimes the signal is present and sometimes it is not. The subject can only make an informed guess, but in the long run there will still be errors. These errors could take the form of false positives or false negatives, which means the subjects either say “yes” when the signal is actually absent, or say “no” when the signal is actually present. The degree of overlap of the distributions characterizes the perceptual sensitivity, and is measured by d' , which is the distance between the two distributions in terms of their variance. The smaller the degree of overlap, the higher the sensitivity.

Note that this model also applies to the discrimination between two stimuli, because we can think of the blank as a stimulus that differs from the target visual stimulus. So instead of distinguishing between the stimulus and blank, the subject distinguishes between stimulus A and stimulus B. Formally it is the same.

Criterion setting and maintenance reflect consciousness

Many studies take d' as a measure of perceptual consciousness. If $d'=0$ the experimenter claims that the subject does not consciously perceive the stimulus. Other studies compare ‘hits’ and ‘misses’. ‘Hits’ are just trials where the target is present, and the internal signal strength is higher than the criteria and therefore the subject responds “yes”. Misses are trials where the target is present, but the signal strength is lower than the criteria, and thus the subject responds “no”. Therefore, comparing ‘hits’ against ‘misses’ is essentially comparing trials with high internal signal strength and trials with low internal signal strength. In terms of task performance or accuracy, ‘hits’ are by definition 100% correct, ‘misses’ are 0% correct.

I argue that neither d' nor internal signal strength is a good measure of consciousness.

It is useful to remind ourselves that signal detection theory was developed partly in order to characterize the behaviour of simple electronics.

A functional photodiode has a $d' > 0$, but it is hard to argue that it is conscious of light. Similarly, blindsight patients show high d' in detecting stimulus in regions of the impaired visual field, and yet, they do not report perceptual consciousness. Sensitivity measure d' characterizes the basic effectiveness of information processing, which we have argued is not necessarily the same as consciousness.

Similarly, the variability of internal signal strength is a basic feature of any noisy detection system. The fact that SDT is useful in characterizing performance means that the internal signal fluctuates, which is why we need to represent the signal strength in terms of probability distributions. The fact that the internal signal fluctuates means that there will naturally be 'hits' and 'misses', if the appropriate criterion is being set. Comparing trials with high internal signal strength against trials with low internal signal strength is just comparing different degrees of effectiveness of information processing (100% performance vs. 0% performance). Similar to d' , this captures the objective aspect of perceptual processes, but do not reflect the subjective nature of consciousness (Lau, in press).

I argue that the criterion for perceptual decisions is more relevant to study of consciousness, because of studies of both blindsight and normal observers.

One account of why blindsight patients deny conscious perception of the stimuli is that they adopt an extreme criterion (c) for detection. In other words, despite their high d' , they use a very conservative strategy, and respond "no" all the time in a detection situation (Campion and Latto, 1985). This explains their apparent lack of awareness and is also compatible with SDT, because c and d' are independent, in the sense that subjects can set whatever criterion they see fit, regardless of their d' . Alternatively, Azzopardi and Cowey (1997, 1998) have reported that blindsight patients fail to maintain a stable criterion in a detection situation (i.e. a Yes-No task), but has no such problem when they perform '2-alternative forced-choice' tasks, which involve distinguishing the spatial or temporal arrangement of two stimuli. This means that when they perform in a detection

task, the criterion they use for decision changes across trials. This leads to an inflation of the measured sensitivity for the detection, but not for the '2-alternative forced-choice' task. The authors argue that this is why the behaviour of blindsight patients is so unusual. Taken together these suggest that failing to set and maintain the criterion properly might be an important factor that contributes to blindsight.

Another reason why we consider decision criteria to be important for consciousness is to due to the results of a recent study (Lau and Passingham, 2006). In that study, I have shown that given the same d' , discrimination accuracy, reaction time, and similar stimuli, the same normal subjects can report different levels of perceptual consciousness in two different conditions. After a forced-choice response to discriminate between a square or a diamond, the subjects were asked to also say whether they actually saw the target or they had just guessed. This procedure is based on the "commentary key" paradigm used to test blindsight patients (Weiskrantz, 1999). One can consider the additional Seen/Guessed judgements within the framework of SDT, in that they require the adoption of additional criteria. In other words, instead of setting one criterion to classify the signal strength as high or low, and thus respond "yes it is a square" and "no it is not a square" respectively, one could set three criteria to classify the signal into four ranges, and respond "yes I see it is a square", "yes I guess it is a square", "no I guess it is not a square", and "no I see that it is not a square". Considered this way, the change of proportions of trials reported as 'Seen' rather than 'Guessed' reflects a change in the criteria between seeing and guessing, which could occur even if d' is kept constant across the different conditions.

These considerations suggest that setting and maintaining the criteria appropriately might be an important aspect to perceptual consciousness. Or at the very least, to the extent to which consciousness could be characterized by SDT, the setting and maintenance of criteria seem more relevant than d' and signal strength. However, an account is still needed as to why in some cases subjects set the criterion in an unusual way, and how that affects consciousness.

Optimal criterion setting and its failure

Bayesian decision theory (Kersten et al., 2004) offers a general perspective to the optimal setting of a criterion setting, and may therefore help us to understand its failure. The probability distributions in Fig. 1 represent the probability that the internal signal would be of a certain strength, given that the target stimulus is present or not. When making an optimal decision, one would like to know the opposite, i.e. the probability that the target stimulus is present, given a certain internal signal strength. This could be easily worked out by using Bayes theorem, which mathematically relates any pair of reverse conditional probabilities $P(A|B)$ (probability that A given B) and $P(B|A)$ (probability that B given A) by taking into account the prior probabilities $P(A)$ and $P(B)$. The important prior information here is how frequently the stimulus is presented in general. Quite often, in psychophysical experiments we present the stimulus on 50% of the trials, and tell the subjects so. In this case, the prior information does not bias the optimal criterion; one could determine the criterion by looking at Fig. 1 alone, if the objective is just to maximize accuracy.

After working out the probability that the stimulus is present, given different internal signal strengths, we just set a criterion such that a signal beyond that strength implies that it is likely that the stimulus is present ($P > 0.5$), and a signal below that strength implies it is likely to be absent. We should note that Bayesian decision theory allows a more generalized view to optimality, in that it could also take into account the payoffs and punishments for different types of correct responses and errors, so that one maximizes the expected payoffs instead of accuracy. Here, for simplicity we assume that the objective is just to maximize accuracy, which is equivalent to stipulating that any correct response is worth the same as the avoidance of any incorrect response.

Under these assumptions, i.e. unbiased prior and maximizing accuracy, one would set the criterion for decision at a level that best divides the two probability distributions as shown in Fig. 1, right between the two distributions. This is the optimal criterion, in the sense that it produces a

minimum amount of errors. Why do blindsight patients not fix the criterion stably at this optimal point? In the case of the study where d' was matched in two conditions, why would the same subjects set the additional Seen/Guessed criteria differently in two conditions, given that the same d' in the conditions implies similar probability distributions for the stimuli (assuming that they are both Gaussians and of equal variance)?

The solution I offer is as follows: given that the optimal criteria are determined by the probability distributions for the stimuli, one's knowledge of these distributions is important. However, the distributions describe the probabilistic behaviour of the internal signal, which has to be learned over time. The learning of one's own internal signal produces representations concerning the internal signal, which itself is a representation of the external stimuli. In this sense, we are creating representations of representations, and thus they are described as "higher-order" representations (Rosenthal, 2000, 2002). I propose that perceptual consciousness depends on our Bayesian decisions, i.e. criterion setting, based on these higher order representations.

Let us take a simple imaginary example by assuming that we try to detect a dim light by using the firing rate in the primary visual cortex as the only source of evidence. If there is no light, the neurons may fire on average at 10 Hz, with a standard deviation of 5 Hz. If there is light, the neurons may fire on average at 15 Hz, again with a standard deviation of 5 Hz. So given this information, a reasonable subject who has the goal to maximize accuracy would set a criterion at 12.5 Hz, and then say there is light when the firing rating exceeds that criterion and say that there is no light if it does not. It follows, then, that what criterion is set depends on what the subjects *think* their own average firing rates are when there is a light, and when there is no light. To make an estimate of the average level of a fluctuating signal, one has to statistically sample the data over time. So essentially, the subject has to learn the probabilistic behaviour for their own neurons. If this learning fails, or is incomplete, such that the subject makes a biased or incorrect estimate of the firing rates, the criterion they set would be suboptimal.

How exactly this learning occurs is beyond the scope of this paper. Previous work on criterion setting in psychology (Treisman, 1984) has suggested methods of learning that do not involve explicitly modelling the probability distributions, but that could be treated as a special heuristic that satisfies the same goal of learning the distributions and basing the criterion on the result of that learning. Without making any strong assumptions about the form of the distributions, a general solution to this learning problem could take the form of a standard Bayesian learning procedure. In the simple case where there are only two stimulus conditions (e.g. dim light or no light) and the subjects are told whether they make correct decisions after responding, the learning should be a fairly straightforward problem. However, in real life when stimuli are multidimensional and feedback is not always available, this could take more complicated forms of unsupervised learning. The critical point here is that the result of this learning affects the setting and maintaining of criteria in the detection or discrimination of stimuli, which might be important for the normal functioning of perceptual consciousness. We now turn to how different cases of failure of perfect learning can produce behaviour that characterizes disturbance of perceptual consciousness.

Misrepresentations

One obvious way of failing to learn the probability distributions is to grossly misrepresent the mean or the variance of the signal (Fig. 2). So in the example given above, the subject could overestimate the firing rate given that the dim light is present, so that it is thought that the average rate is 25 Hz instead of 15 Hz. The subject might then set the criterion, reasonably, at the mid-point (i.e. 17.5 Hz) between the expected averages for the signal present and signal absent conditions, which are 10 and 25 Hz. This high criterion of 17.5 Hz would lead to a high portion of false-negatives, because actually the average firing rate is only 15 Hz when the dim light is on; subject will be missing most of it. The subject would behave as if the dim light is not perceived most of the time,

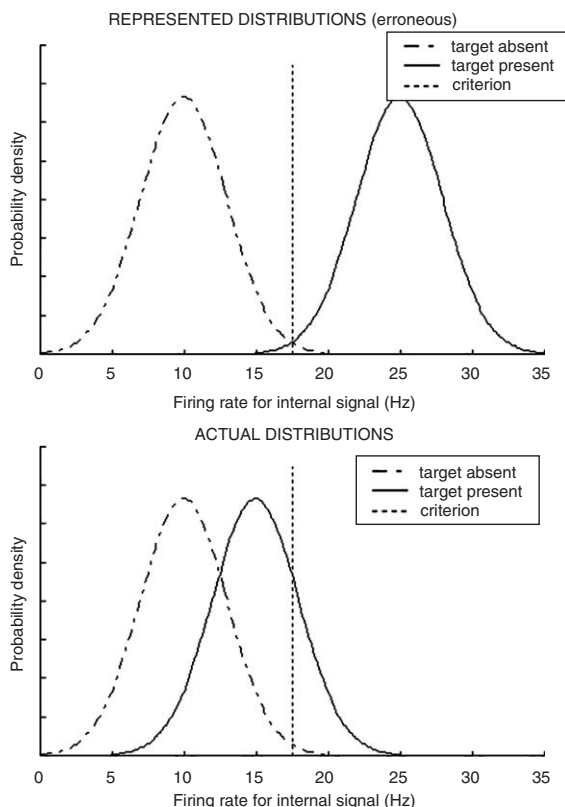


Fig. 2. Misrepresentation. Because one could only set the criterion based on the learned distributions (upper graph), instead of the actual distributions (lower graph), if the two are different, i.e. the learned distributions are incorrect, one sets the criterion suboptimally. Here only one form of misrepresentation is depicted. However, one could imagine the other case, such as representing both the target absent and target present distributions as higher than reality, or both lower than reality, etc.

because the subject responds “no” even when the stimulus is present and the firing rate is at its most likely frequency. The actual d' , however, remains the same, because it measures the distance between the actual distributions, but not the learned distributions.

This would be similar to the behaviour exhibited by blindsight patients, where a negative response is usually given in detection, although a fairly high d' is implied by other indirect forced-choice measures. The primary visual cortex is likely to be one important major source of the internal signal. After a lesion to the cortex, the actual internal

signal decreases dramatically, with the remaining weak signal possibly routing through subcortical pathways. If the subject fails to learn that the signal has dropped, and uses the old criterion for detection, this is equivalent to overestimating the signal as described above. In fact, it is likely that the primary visual cortex is also a source of the baseline noise, i.e. the signal when no stimulus is present. A lesion to the primary visual cortex is thus likely to shift both probability distributions (signal present and absent) to the negative direction. Failing to learn this shift can result in a dramatic positive deviation from the optimal criteria. Understood this way, blindsight could be partly due to a failure to learn the reduction in signal strength after lesion to the primary visual cortex; after the lesion, because of the inappropriate criterion, blindsight patient mis-classifies stimuli as noise.

Hallucinations could be treated as the opposite within this framework. Hallucination could be treated as the production of false-positives, in the sense that noise is being mis-classified as stimuli. By this definition, one hallucinates while dreaming; in dreams we consciously perceive stimuli that are not really there. According to the present framework, this could be due to the underestimation of the signal strength. When brain activity is monitored by electroencephalogram (EEG), sleep can be divided into different stages by the EEG pattern. Dreams are more likely to be reported during a stage of sleep that is characterized by rapid eye movement (REM), and brain activity of relatively high frequency and intensity. Let us assume that the overall signal during REM-sleep is higher. If the brain maintains the same criterion for detection over alternations of REM and non-REM sleep, it would be predicted that false-positives are a lot more likely during REM-sleep, because of the higher signal intensity. Perhaps during sleep, when the brain is not actively learning, it only makes a general estimation of the probability distributions of signal for both REM and non-REM sleep combined, and that is why we fail to set a appropriately high criterion during REM sleep. With this inappropriately low criterion for detection, one mis-classifies noise as stimuli.

Ambiguity

Another way to fail to fully learn the probability distributions is to learn it with high ambiguity (Fig. 3). Ambiguity is formally defined as uncertainty about probabilities (Ellsberg, 1961). The graphs in Figs. 1 and 2 represent the probability that the signal intensity is of a certain strength, given that the signal is absent or present. The fact that we use a line to represent the bell shape curves means that at each signal intensity level, there is a definite number that represents the probability. In reality, this is possible: the probability that when there is a stimulus, the probability that the signal strength is between x and $x+1$ could be exactly 0.01245, for example. However, for any subject who is learning this distribution, one could only learn this with a certain degree of uncertainty. The probability might be expected to be at 0.01245

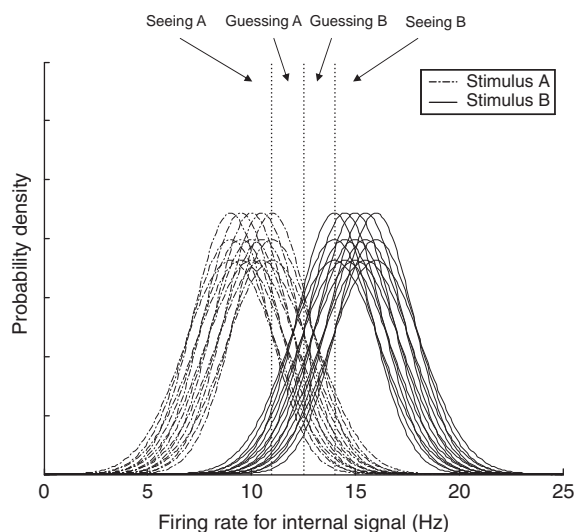


Fig. 3. Ambiguity. Rather than representing the distributions with absolute certainty, one might represent them as a possible range of values (ambiguity). Here one represents the probability distributions with standard deviations 1.8–2.2 Hz, and means of 9–11 Hz and 14–16 Hz respectively for Stimulus A and B. These ranges could be considered as similar to confidence intervals or error bars. When subjects were asked whether they were “guessing”, they could be setting three criteria as depicted in the diagram. It is possible that when the level of ambiguity increases, the criteria for guessing might change accordingly so that more trials are classified as “guessing”.

(i.e. it is the most likely), but the 95% confidence interval may cover between 0.01241 and 0.001248. In other words, when we try to learn the information on Fig. 1, which depicts reality, in our brain we try to create the same graphs that best represent the same information, but possibly with error bars on it, to reflect how certain we are about our estimation of reality (Fig. 3). In this sense, representing ambiguity is not a failure, but rather a useful way to capture the degree to which we are certain. However, having very high ambiguity means that we are not really certain about our estimation, which is not favourable for us.

Recall that under the assumptions we have taken, the optimal way to set the criterion is to set it such that it best classifies the two distributions for stimulus present and stimulus absent, which is the mid-point between the two means. For signal strength above the criterion, we expect that it is more likely that the stimulus is present than not. However, even if we estimate that the probability that the stimulus is present is bigger than 0.5, it may not be significantly bigger than 0.5. Maybe the error bar or confidence interval is so large that it covers the point of 0.5. In other words, we estimate the stimulus to be there, rather than not there, but we are not so sure about this estimation.

This formally characterizes guessing. Obviously, having guessed A instead of B we think that A might be marginally more likely. But it is counted as guessing because we are not ‘significantly’ certain about our decision. Significance here could be defined as in statistics, i.e. when we are significantly certain about a decision, the 95% confidence interval (or whatever other confidence level) for the likelihood that we are correct does not cover 0.5 (chance). The amount of guessing depends on the level of ambiguity.

This offers a possible explanation as to why sometimes given the same d' , discrimination accuracy, reaction time, and similar stimuli, the same subjects can report different levels of perceptual consciousness in two different conditions (Lau et al., 2006). This is because although the underlying distributions are the same, the subjects may learn the distributions for the two conditions with different level of ambiguity (uncertain regarding the distributions). Subjects

set the criteria for Seen/Guessed at the signal strength at which they become significantly certain about the discrimination, and this depends on the level of ambiguity. In other words, if the error bars are large and they overlap greatly, for the distributions, they set the criteria accordingly so that many trials are classified as “Guesses”. The underlying actual distributions, however, could be exactly the same, and thus the actual performance levels are matched.

Dynamic fluctuation

Finally, we briefly consider another type of failure to learn the distributions properly, which is to fail to make stable estimates of the distributions. In a way, this is a form of misrepresentation, but here the emphasis is on how this misrepresentation fluctuates dynamically. Unless the subjects have learned the distributions perfectly and precisely, one would expect that in every trial, the subjects would acquire new information and thereby change the estimation accordingly. In this sense, dynamic fluctuation is a sign that we are learning the distributions. If the learning is effective, one would expect that the estimation of the distributions to fluctuate less and less, and eventually converge to the true forms. However, if the learning itself is not optimal, this fluctuation may continue.

As mentioned above, it has been reported that blindsight patients fail to maintain a stable criterion in a detection situation (i.e. a Yes–No task), but have no such problem when they perform ‘2-alternative forced-choice’ tasks (Azzopardi and Cowey, 1997, 1998). This might seem hard to explain, because if they fail to maintain a stable criterion in the Yes–No task because they fail to remember where to put the criterion, they should also fail to maintain a stable criterion in the other task. However, if the failure is due to the fluctuation in the learned representation for the stimulus-absent distribution only, but not for the other distributions, this could explain the jittering of criterion for the task of detection but not discrimination tasks.

Fluctuation is interesting to consider because unlike the other two forms of failure, it actually affects d' as measured by conventional methods. This is because if one's learned distributions fluctuate, one's criterion also fluctuates; presumably one sets the criterion according to the most up-to-date estimation. When the criterion fluctuates, effectively we have a reduced d' . This has been suggested to be the explanation for why in detection tasks, blindsight patients have a d' that is lower than expected, when one estimates it from the d' for the 2-alternative forced-choice' tasks (Azzopardi and Cowey, 1997, 1998).

Empirical support

It is important to note that the aforementioned three types of failure are not incompatible with each other. One could exhibit all three problems, or just one, or two of them. Therefore it is not really an issue at the moment to determine which is *the* failure that causes disturbance of perceptual consciousness. In reality all three of them may play some role. It is a matter for future research to empirically investigate the relative importance of each type of failure of representation in different contexts. At this stage, we consider empirical evidence that support the general notion that perceptual consciousness depends on the representation of the probability distributions that describe the behaviour of the internal signal.

One basic prediction of this framework is the dissociation between consciousness and detection/discrimination performance. This is because except in the case where dynamic fluctuation is significant, the basic detection and discrimination performance is determined by the actual probability distributions. However, consciousness depends on *the representation* of the probability distributions, by which we set and maintain the criteria. This representation, as we discussed above, may err. Therefore the framework predicts that given the same d' , the same subjects performing the same detection/discrimination task may report different levels of consciousness under different conditions, if the distributions are represented differently.

In a recent study that has been mentioned above (Lau et al., 2006), we showed exactly that. In that task, subjects were required to discriminate between a square and a diamond figure (Fig. 4). We also asked subjects, after the discrimination in each trial, to state whether they consciously saw the identity of the target or that they just guessed what it was. Therefore, we have both an objective forced-choice measure of performance, as well as a subjective measure of perceptual consciousness. The target was metacontrast masked at different stimulus onset asynchrony (SOA, i.e. the temporal distance between the target and the mask), so as to produce different conditions with various levels of difficulty. We capitalized on the fact that the masking function (performance against SOA) is U-shaped for metacontrast masking, which means that there will always be two SOA points at which the performance levels will be matched (Fig. 5). We found that at the two SOA points, the subjective levels of perceptual consciousness differed, in that in the shorter SOA condition subjects claimed to be guessing more frequently. Within the present theoretical framework, this could be understood in terms of different levels of ambiguity for the two SOA conditions.

If this difference in levels of consciousness is due to a difference in the higher order representations, that is the representation of how the internal signal behaves statistically, it should be associated with a difference in neural activity in the relevant brain area. The prefrontal cortex is likely to play an important role in forming and maintaining these higher order representations, because it has also been implicated in studies of uncertainty and learning (Daw et al., 2005; Huettel et al., 2006; Yoshida and Ishii, 2006). Also, it receives anatomical projections from areas of all sensory modalities, and has been considered as situated at the top of the information processing hierarchy of the brain (Goldman-Rakic, 1995; Fuster, 1997; Passingham et al., 2005). The internal signal, on the other hand, is likely to be represented in the occipital and temporal cortices, where neurons code specific visual information. Therefore, our models predicts that if we compare the trials for the two SOA points, where the subjective level of perceptual consciousness differs but forced-choice

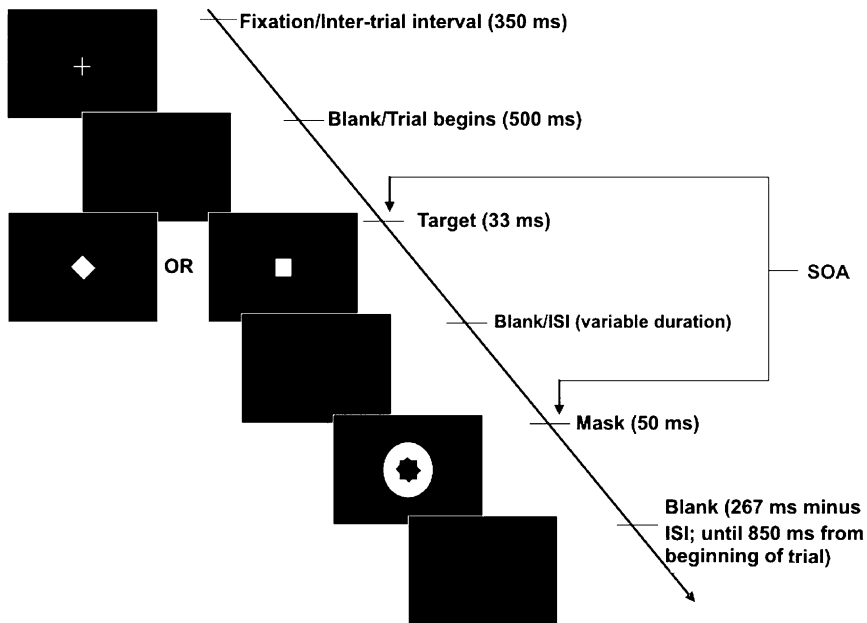


Fig. 4. Visual discrimination task with metacontrast masking. After the presentation of the target and the mask, the participants were first asked to decide whether a diamond or a square was presented. Then, they had to indicate whether they actually saw the target, or that they simply guessed the answer. Shown in the brackets are the durations of each stimulus.

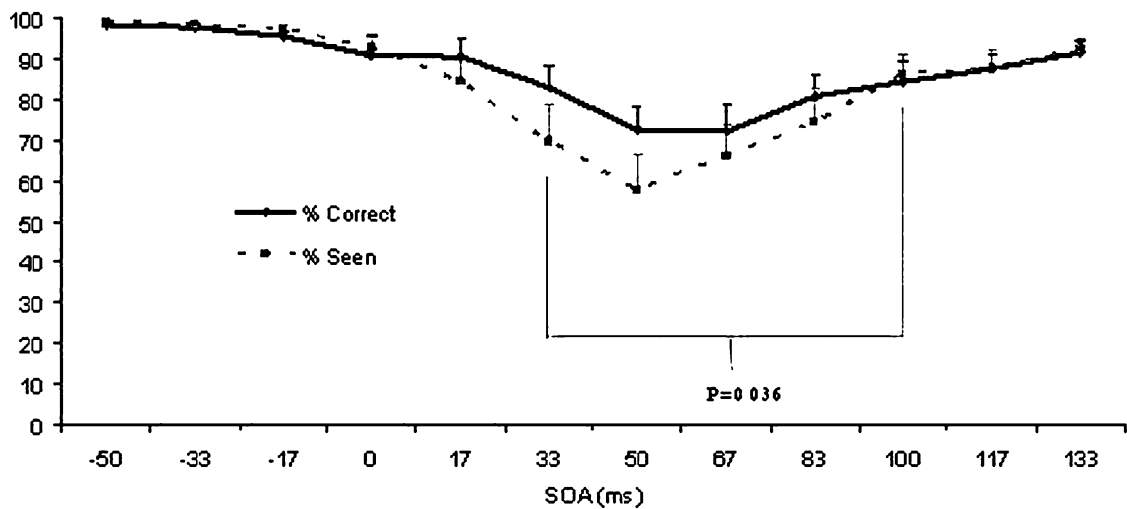


Fig. 5. Consciousness and performance. These results were obtained using the procedure described in Fig. 1, except that it also included trials where the mask was presented before the target (paracontrast masking). Note that at the SOAs where the performance levels (% correct) were the same (e.g. 33 and 100 ms), the awareness levels (% seen) differed significantly.

performance does not, there should be a difference in activity in the prefrontal cortex but not the in the early visual areas. In fact, this is what we observed (Lau et al., 2006). This is a counter-intuitive finding because most theories of visual consciousness suggest that the critical neural correlate should be in visual areas (Zeki and Bartels, 1999; Lamme and Roelfsema, 2000; Lamme, 2003). Even though some researchers have proposed that a ‘frontal-parietal network’ that might be important for consciousness, they typically suggest that this is only important in addition to the visual areas (Rees et al., 2002). However, in our study, the dorsolateral prefrontal (DLPFC) cortex is the only area where we could find a significant difference in activity. This fits with the central idea of the model that perceptual consciousness depends on higher order representations (in the prefrontal cortex), and it can change in the absence of a difference in the internal visual signal (in occipital and temporal visual areas).

Interestingly, the DLPFC has also been implicated in a study of blindsight. Sahraie et al. (1997) have reported results from a study on a blindsight subject (known as GY). The subject has a lesion to the primary visual cortex that affects roughly half of his visual field, stimuli presented to which yield no phenomenal visual awareness. The authors presented to this “blind field” slowly moving ($3^\circ/\text{s}$) stimuli of which the subject was unaware, and found that the subject could nonetheless discriminate the direction of the horizontal movement at slightly above 80% correct. This visual stimulation was associated with a lack of significant activation in the DLPFC. However, when the speed of the movement of the stimuli was increased to $20^\circ/\text{s}$, the subject reported a sense of awareness even though the visual presentation was to the “blind field” (a phenomenon known as type II blindsight), and the performance of discrimination was above 90% correct. This visual stimulation was also associated with a significant activation of DLPFC. The performance levels for discrimination task in these conditions were different, but could be considered roughly matched, because they were both well above chance. Incidentally, when visual stimuli were presented to the unimpaired field of the blindsight subject, there was also significant

activation in the DLPFC. These results, as in the study discussed above (Lau et al., 2006), support the idea that activity in the DLPFC vary in relation to the level of awareness, even when performance level is not an important contributing factor.

Finally, the DLPFC has also been implicated in studies concerning sleep and dreams (Maquet et al., 1996; Muzur et al., 2002). As explained earlier, within the present framework dreams and hallucinations could be considered as the opposite of blindsight. If the internal signal intensity increases, but the higher order representations and thus also the criterion remain the same, we are likely to produce false-positives for conscious perception, and this formally characterizes dreams and hallucinations. We have argued that in REM sleep the internal signal is likely to be higher than in non-REM sleep. And we know that dreams are likely to be reported during REM sleep, and unlikely to be reported during non-REM sleep. Therefore, one way that the aspects of dreams regarding perceptual consciousness could be explained is that activity in the DLPFC should be similar for both REM and non-REM sleep. This possibility reflects that the higher order representations and criterion remain the same between REM and non-REM sleep. If this is true, the higher internal signal intensity during REM sleep would produce false-positives. This pattern of activity is in fact found in the DLPFC when neural activity during REM and non-REM sleep was compared using positron emission topography (PET) (Maquet et al., 1996; Muzur et al., 2002). Compared to wakefulness, during non-REM sleep many areas are deactivated. During REM sleep, most areas are reactivated to normal wakefulness level. However, the DLPFC remains deactivated in REM sleep, in a level similar to that during non-REM sleep. In fact, the DLPFC is the only area in the prefrontal cortex that remains deactivated (Muzur et al., 2002).

Taken together, these results are compatible with the notion that the subjective aspects of perceptual consciousness depend on the higher order representations of how the internal signal behave, which are likely to be associated with the prefrontal cortex. Future studies should further

examine their interactions with areas that are likely to represent the internal signal, such as occipital and temporal areas as in the case of vision.

Finally, the model predicts that manipulating the activity in the prefrontal cortex, and thus the higher order representations, should change the level of perceptual consciousness but not forced-choice performance in detection/discrimination tasks. This offers a solution to the controversy as to whether lesion to the prefrontal cortex impairs consciousness. Critics (Pollen, 1995, 1999) have argued that there has been a lack of reported cases of such prefrontal damaged patients. However, most previous studies of visual consciousness have focused on forced-choice performances, which we predict should not show any difference if only the higher order representations are manipulated. Instead, subjectively reported perceptual consciousness should change. We are currently using transcranial magnetic stimulation (TMS) to test these hypotheses.

Philosophical issues and final remarks

We started off by arguing that signal detection theory is too simple to characterize perceptual consciousness. One could argue that the model we presented here, that perceptual consciousness depends on the setting and maintaining of criterion based on representations of the statistical behaviour of internal signals, is not substantially more complicated. Does this really solve the “hard problem” of explaining the subjective nature of consciousness in terms of physical facts (Chalmers, 1996)? If d' is a bad measure of consciousness because a photodiode could have a d' value as high as a conscious human being, does the same criticism not apply to our model? One could imagine building a device that learns dynamically its only internal signal for detection/discrimination, and maintains optimal criterion according to Bayesian decision theory. Does this make the device conscious?

I do not intend to claim that the present framework readily solves all these problems, or at least I am not going to argue so within the space of this paper. The foregoing thought experiment is

an interesting one, but one should not overlook the complexity of a moving robot that could dynamically learn the probability distributions for its internal signal for performing detection/discrimination in a changing environment. However, here I only recommend a minimal interpretation of the model: it formally characterizes perceptual consciousness, in the sense that it describes the conditions under which consciousness is intact or disturbed in a human subject. It is not supposed to explain all features of consciousness. The main point I wish to make is that perceptual consciousness depends on some form of criterion setting, though it does not mean that all forms of criterion setting is directly relevant to consciousness. For instance, there might be a specific criterion for conscious perception, and a different one for responding or communicating the information to others. The important hypothesis is that the principles for the maintaining and the setting of the criterion for conscious perception should be compatible with the framework described here.

I also argue that because the present model takes the form of a higher order representational theory as discussed frequently in philosophy (Rosenthal, 2000, 2002), it shares similar philosophical explanatory power. To the degree to which higher order representational theories in general can solve the philosophical problems associated with ‘explaining’ consciousness, I speculate that the present model probably does at least equally well. In fact, I am going to argue that it is likely to be more attractive than many other versions of higher order representational theories.

One problem of higher order representational theories is the problem of mismatch between higher and lower levels of representations. Normally, in the current philosophical literature, both the higher and lower level of representation is taken to represent information regarding the external world. If the first level representation represents ‘red’ but the higher order representation represents ‘green’, it is unclear what the conscious experience should be (Neander, 1998). However, in the present model, the higher order representation represents a scale by which the first-order representation (the internal signal) could be interpreted. The internal signal carries no fixed meaning unless

one is to have some access to the higher order representations; a firing rate of 5 Hz in the early visual cortex could mean that a signal is very likely to be present, or very unlikely to be so, depending on the higher order representations. Similarly, the higher order representations do not make sense outside of the context of the internal signal. This way, a mismatch between the levels in the above sense is simply not possible: their content cannot directly contradict, because they are never meant to duplicate each other.

Finally, the present model should also shed light on the function of perceptual consciousness. I have argued earlier than one motivation of the model is that the function of consciousness is unclear. However, if perceptual consciousness depends on the correct knowledge of the variability of the internal signal, when one is conscious one should also be able to perform functions that depend also on this knowledge. An obvious example might be optimal betting, based on one's own performance (Persaud et al., 2007).

As a closing remark, I further suggest that some form of social interaction may also depend on this knowledge which underlies perceptual consciousness. Several observers could make optimal joint decisions by combining their information regarding the same external stimulus. This optimal joint decision could be predicted by a Bayesian approach, in which the team of observers could be considered as a "Bayesian committee". These approaches typically assume that each of the observers know their own variance of their internal signal. In other words, if the present model is correct, that perceptual consciousness depends on the correct knowledge of the variability of one's own internal signal, subjects who are perceptually conscious should be suitable candidates for joining a Bayesian committee which gives optimal team responses. One could imagine being required to team up with another observer who has a certain detection sensitivity. In order to maximize joint performance, when opinions differ one would discuss with the partner and compromise based on the relative levels of confidence in the decision in a particular trial. It is not difficult to see that one would rather team up with normal observers rather than with blindsight patients who have the

same sensitivity. If the team partners claim that their confidence is low and they are guessing all the time even when their responses are correct, negotiating and compromising for an optimal joint response becomes impossibly difficult.

Acknowledgments

I thank Chris Frith, Peter Dayan, Peter Latham, Allan Hobson, Karl Friston, and Tim Behrens for discussions that contributed to the development of the presented ideas. I also thank Chris Frith and Dick Passingham for comments on an earlier version of this manuscript. This work is supported by the Wellcome Trust.

References

- Azzopardi, P. and Cowey, A. (1997) Is blindsight like normal, near-threshold vision? *Proc. Natl. Acad. Sci. U.S.A.*, 94: 14190–14194.
- Azzopardi, P. and Cowey, A. (1998) Blindsight and visual awareness. *Conscious. Cogn.*, 7: 292–311.
- Baars, B.J. (1988) *A Cognitive Theory of Consciousness*. Cambridge [Cambridgeshire]. Cambridge University Press, New York.
- Campion, J. and Latta, R. (1985) Apperceptive agnosia due to carbon monoxide poisoning. An interpretation based on critical band masking from disseminated lesions. *Behav. Brain Res.*, 15: 227–240.
- Chalmers, D. (1996) *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, Oxford.
- Cleeremans, A. (2005) Computational correlates of consciousness. *Prog. Brain Res.*, 150: 81–98.
- Daw, N.D., Niv, Y. and Dayan, P. (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.*, 8: 1704–1711.
- Dehaene, S., Sergent, C. and Changeux, J.P. (2003) A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proc. Natl. Acad. Sci. U.S.A.*, 100: 8520–8525.
- Dijksterhuis, A., Bos, M.W., Nordgren, L.F. and van Baaren, R.B. (2006) On making the right choice: the deliberation-without-attention effect. *Science*, 311: 1005–1007.
- Ellsberg, D. (1961) Risk, ambiguity, and the savage axioms. *Q. J. Econ.*, 75: 643–669.
- Fuster, J.M. (1997) *The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe* (3rd ed.). Lippincott-Raven, Philadelphia, PA.
- Goldman-Rakic, P.S. (1995) Architecture of the prefrontal cortex and the central executive. *Ann. N.Y. Acad. Sci.*, 769: 71–83.

- Green, D. and Swet, S. (1966) *Signal Detection Theory and Psychophysics*. Wiley, New York.
- Huettel, S.A., Stowe, C.J., Gordon, E.M., Warner, B.T. and Platt, M.L. (2006) Neural signatures of economic preferences for risk and ambiguity. *Neuron*, 49: 765–775.
- Kersten, D., Mamassian, P. and Yuille, A. (2004) Object perception as Bayesian inference. *Annu. Rev. Psychol.*, 55: 271–304.
- Lamme, V.A. (2003) Why visual attention and awareness are different. *Trends Cogn. Sci.*, 7: 12–18.
- Lamme, V.A. and Roelfsema, P.R. (2000) The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.*, 23: 571–579.
- Lau, H.C. (in press). Are we studying consciousness yet? In: Weiskrantz L., Davies M. and Parker A. (Eds.), *Chichele Lectures 2006: Frontiers of Consciousness*. Oxford University Press, Oxford.
- Lau, H.C. and Passingham, R.E. (2006) Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proc. Natl. Acad. Sci. U.S.A.*, 103: 18763–18768.
- Lau, H.C., Rogers, R.D. and Passingham, R.E. (2006) On measuring the perceived onsets of spontaneous actions. *J. Neurosci.*, 26: 7265–7271.
- Lau, H.C., Rogers, R.D. and Passingham, R.E. (2007) Manipulating the experienced onset of intention after action execution. *J. Cogn. Neurosci.*, 19(1): 81–90.
- Macmillan, N. and Creelman, C. (1991) *Detection Theory: A User's Guide*. Cambridge University Press, Cambridge, England.
- Maquet, P., Peters, J., Aerts, J., Delfiore, G., Degueldre, C., Luxen, A. et al. (1996) Functional neuroanatomy of human rapid-eye-movement sleep and dreaming. *Nature*, 383: 163–166.
- Muzur, A., Pace-Schott, E.F. and Hobson, J.A. (2002) The prefrontal cortex in sleep. *Trends Cogn. Sci.*, 6: 475–481.
- Neander, K. (1998) The division of phenomenal labor: a problem for representational theories of consciousness. *Philos. Perspect.*, 12: 411–434.
- Passingham, R.E., Rowe, J.B. and Sakai, K. (2005) Prefrontal cortex and attention to action. In: Humphreys G.W. (Ed.), *Attention in Action*. Taylor & Francis Group, Inc., London, UK.
- Persaud, N., McLeod, P. and Cowey, A. (2007) Post-decision wagering objectively measures awareness. *Nat. Neurosci.*, 10(2): 257–261.
- Pollen, D.A. (1995) Cortical areas in visual awareness. *Nature*, 377: 293–295.
- Pollen, D.A. (1999) On the neural correlates of visual perception. *Cereb. Cortex*, 9: 4–19.
- Rees, G., Kreiman, G. and Koch, C. (2002) Neural correlates of consciousness in humans. *Nat. Rev. Neurosci.*, 3: 261–270.
- Rosenthal, D.M. (2000) Consciousness, content, and metacognitive judgments. *Conscious. Cogn.*, 9: 203–214.
- Rosenthal, D.M. (2002) How many kinds of consciousness? *Conscious. Cogn.*, 11: 653–665.
- Sahraie, A., Weiskrantz, L., Barbur, J.L., Simmons, A., Williams, S.C. and Brammer, M.J. (1997) Pattern of neuronal activity associated with conscious and unconscious processing of visual signals. *Proc. Natl. Acad. Sci. U.S.A.*, 94: 9406–9411.
- Tononi, G. (2004) An information integration theory of consciousness. *BMC Neurosci.*, 5: p. 42.
- Treisman, M. (1984) A theory of criterion setting: an alternative to the attention band and response ratio hypotheses in magnitude estimation and cross-modality matching. *J. Exp. Psychol. Gen.*, 113: 443–463.
- Wegner, D.M. (2003) The mind's best trick: how we experience conscious will. *Trends Cogn. Sci.*, 7: 65–69.
- Wegner, D.M., Fuller, V.A. and Sparrow, B. (2003) Clever hands: uncontrolled intelligence in facilitated communication. *J. Pers. Soc. Psychol.*, 85: 5–19.
- Wegner, D.M. and Wheatley, T. (1999) Apparent mental causation. Sources of the experience of will. *Am. Psychol.*, 54: 480–492.
- Weiskrantz, L. (1986) *Blindsight. A case study and implications*. Oxford University Press, Oxford, UK.
- Weiskrantz, L. (1999) *Consciousness Lost and Found*. Oxford University Press, Oxford.
- Yoshida, W. and Ishii, S. (2006) Resolution of uncertainty in prefrontal cortex. *Neuron*, 50: 781–789.
- Zeki, S. and Bartels, A. (1999) Toward a theory of visual consciousness. *Conscious. Cogn.*, 8: 225–259.