PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

*Phil. Trans. R. Soc. B* (2012) **367**, 1401–1411
doi:10.1098/rstb.2011.0380

*Research*

# A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition

**Yoshiaki Ko**[1,*] **and Hakwan Lau**[1,2]

[1]*Department of Psychology, Columbia University, New York, NY 10027, USA*
[2]*Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, The Netherlands*

Blindsight refers to the rare ability of V1-damaged patients to perform visual tasks such as forced-choice discrimination, even though these patients claim not to consciously see the relevant stimuli. This striking phenomenon can be described in the formal terms of signal detection theory. (i) Blindsight patients use an unusually conservative criterion to detect targets. (ii) In discrimination tasks, their confidence ratings are low and (iii) such confidence ratings poorly predict task accuracy on a trial-by-trial basis. (iv) Their detection capacity ($d'$) is lower than expected based on their performance in forced-choice tasks. We propose a unifying explanation that accounts for these features: that blindsight is due to a failure to represent and update the statistical information regarding the internal visual neural response, i.e. a failure in metacognition. We provide computational simulation data to demonstrate that this model can qualitatively account for the detection theoretic features of blindsight. Because such metacognitive mechanisms are likely to depend on the prefrontal cortex, this suggests that although blindsight is typically due to damage to the primary visual cortex, distal influence to the prefrontal cortex by such damage may be critical. Recent brain imaging evidence supports this view.

**Keywords:** blindsight; signal detection theory; consciousness; metacognition

## 1. INTRODUCTION

What accounts for the difference between perceptual processes of which we are consciously aware, and perceptual processes of which we are completely unaware? Blindsight is a classic neurological phenomenon that can shed light on this question [1–3]. Typically damaged in the primary visual cortex, these patients claim to have no conscious visual experience when static visual stimuli are presented to their subjectively 'blind' field. And yet, when required to guess the identity of such stimuli (such as discriminating between whether a horizontal or vertical grating pattern was presented), they can perform well above chance level.

This peculiar dissociation between subjective experience and the ability to perform visual tasks in blindsight is critical for our understanding of the nature of conscious perception. This is because, in most currently popular experimental paradigms for comparing conscious versus unconscious perception, task performance is typically a confounding factor. For instance, in masking experiments [4], subjects usually perform visual tasks (e.g. discrimination) at a much lower level—often at chance level—in the 'unconscious' condition, where subjects do not consciously see the visual targets, whereas performance is typically higher in the 'conscious' condition. When we compare the two conditions, we do not know whether we are comparing different levels of conscious perception *per se*, or just different levels of processing capacity [5,6]. On the other hand, in blindsight, there is above-chance task performance even when conscious visual experience is lacking. This means we can match task performance between blindsight and normal vision conditions when comparing the two [7], sometimes even within the same subject (in cases where brain damage affects only part of the subject's visual field). This is one critical feature of blindsight, which makes explaining it a special challenge to any theory of conscious awareness.

Many studies of blindsight focus on identifying the neural correlates that reflect the phenomenon itself [8–15], and what structural damage and changes account for the abolition of conscious visual experience [16,17]. These results have contributed to views that the primary visual cortex may be necessary for conscious visual experience [18], and that subcortical processing may be essentially unconscious [19]. Currently, however, these correlates do not provide a mechanistic explanation of blindsight.

Here we adopt a different approach, and describe a potential explanation of blindsight based on signal detection theory (SDT) [20] and similar pioneering work by previous authors [21,22]. We offer a systematic treatment of the multiple aspects of psychophysical

* Author for correspondence (yk2450@columbia.edu).

One contribution of 13 to a Theme Issue 'Metacognition: computation, neurobiology and function'.

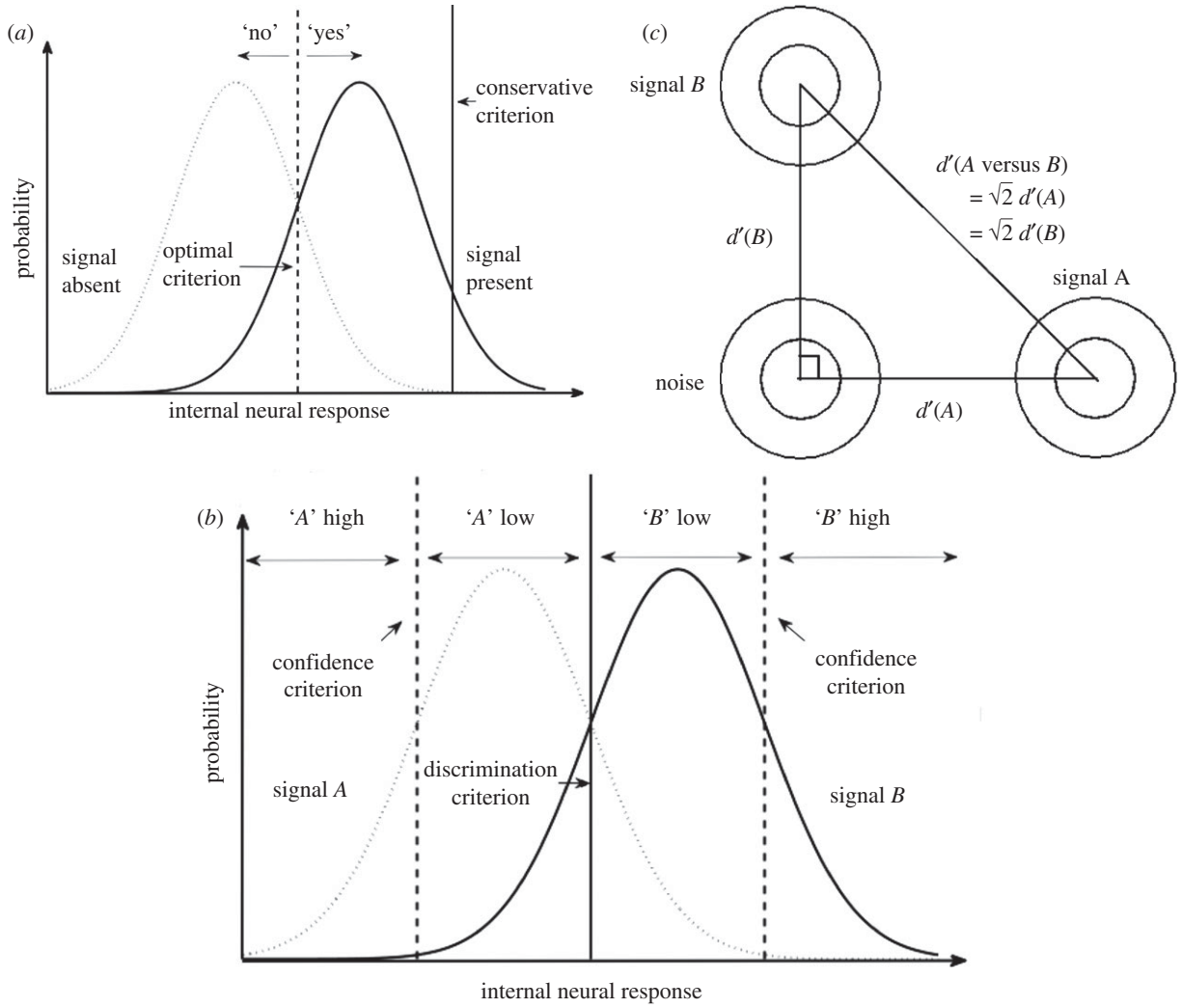


Figure 1. (*a*) Basic signal detection theory (SDT) diagram for a detection task. Subjects make decisions based on a set criterion, answering 'yes' when the internal neural response exceeds the criterion and 'no' otherwise. Using the optimal criterion (dashed line) minimizes errors, whereas using a conservative criterion (solid line) causes the subject to answer 'yes' very rarely. (*b*) SDT diagram for discrimination task with confidence ratings. In addition to a decision criterion (solid line), responses are also broken down into 'low' and 'high' confidence based on whether the internal response exceeds the confidence criteria (dotted lines). (*c*) A schematic of the relationship between detection and discrimination (2AFC). Here, the probability distributions of the internal neural responses are viewed 'from above', represented by co-centric circles. The legs of the right triangle represent the decision dimension where detections are performed, whereas the hypotenuse represents the dimension for discrimination between *A* and *B*. Assuming that signals *A* and *B* are independent and uncorrelated, and that their detection sensitivities are identical, discrimination sensitivity should be equal to $\sqrt{2}$ times the detection sensitivity.

performance in blindsight, and provide a computational framework under which these multiple aspects can be explained by a single mechanism. We then relate the putative computational mechanism to recent results in neuroimaging to discuss their implications.

## (a) *What needs to be explained?*

Although we are interested in the nature of the conscious experience in blindsight, this is hard to define clearly and is therefore less suitable to be the target of explanation for a formal computational theory. Fortunately, blindsight is a well-studied phenomenon. This means we can focus on certain patterns of behaviour that presumably reflect the disturbed nature of conscious awareness. These psychophysical findings can be cast within the context of SDT [20].

According to SDT, subjects make decisions about whether or not a certain stimulus is presented based on noisy internal neural response distributions and the setting of a criterion to differentiate them (figure 1*a*). One key concept of the theory is that the criterion used to make the perceptual decision is mathematically independent from the capacity to detect the target in the presence of noise (i.e. signal-to-noise ratio, $d'$), in the sense that the two are not necessarily correlated.

Because of the mathematical independence between criterion and $d'$, one way to characterize a critical aspect of blindsight behaviour is to say that despite having an above-zero $d'$ (i.e. capacity to process visual signal and to perform visual tasks), blindsight patients may use an unusually conservative criterion for detection, which results in them saying 'no' nearly all the time to the question of 'do you see something?' or 'is there something presented?' This is sometimes taken as a trivializing interpretation of blindsight [23], presumably because conventionally $d'$ is meant to be the important measure in

psychophysical experiments, whereas criterion is often seen as an unwanted subjective bias that could potentially contaminate measures of task performance, such as accuracy rate. Whereas it is true that measures such as accuracy (percentage correct) can be influenced by the criterion, this does not mean that the criterion is an uninteresting annoyance. If blindsight patients set their criterion for detection to be unusually high, then we need an explanation as to why this is the case.

Another important aspect of blindsight is that—in discrimination tasks—even when blindsight patients perform at above chance level, they claim that they are merely 'guessing'. We can characterize this in terms of psychophysical measures, such as confidence ratings, subjective visibility ratings or post-decision wagering responses [22], which are often collected in such experiments. In one study, we tested the well-known patient GY, whose damage to the primary visual cortex was largely restricted to the left hemisphere (affecting the right visual field) [15]. Although we matched for discrimination performance between the sighted visual field and the 'blind' field, GY still gave lower subjective visibility ratings for the blind field compared with the normal-sighted field. These subjective ratings can be generated by having appropriate confidence criteria within the framework of SDT (figure 1*b*). Therefore, once again, what needs to be explained is why blindsight patients use such unusually conservative criteria for high confidence.

The third essential psychophysical feature concerns type II capacity [24], i.e. trial-by-trial correspondence between confidence ratings (or other subjective ratings such as visibility) and accuracy. In normal subjects, usually there is a fairly good correspondence, such that when they rate high confidence, they are more likely to be correct than when they rate low confidence. However, in patient GY, it was shown that the degree of this correspondence was reduced in the blind field when compared with the normal-sighted field, even when the discrimination task-performance level was matched between the two fields [15,25].

Finally, blindsight patients' $d'$ is lower than what one would expect from their capacity to perform two-alternative forced-choice (2AFC) tasks. Although, in recent years, the term 2AFC has often been used to refer to discrimination tasks in general, in the psychophysics tradition, its usage is restricted to tasks in which the same two stimuli are presented in every trial and subjects are required to identify the spatial or temporal arrangement of such stimuli [26]. For example, a spatial 2AFC task may require the subjects to distinguish between these two alternatives: a grating pattern on the left and a blank on the right versus a blank on the left and a grating pattern on the right. Defined as such, it is known that for normal subjects, $d'$ for 2AFC tasks has a principled mathematical relationship with $d'$ for a corresponding detection task (such as detection of the grating pattern versus the blank, corresponding to the 2AFC example task mentioned above). Specifically, one would expect $d'$ for 2AFC tasks to be roughly equal to $\sqrt{2}$ times $d'$ for the corresponding detection task (figure 1*c* gives the explanation). However, it was found that in blindsight patient GY, his $d'$ for detection was lower than one would expect based on his performance in the corresponding 2AFC task [27]. This result has also been replicated in monkeys with lesions to the primary visual cortex [9].

The list of psychophysical features described above by no means forms an exhaustive description of blindsight, which is associated with many other characteristic features [1]. However, we focus on these four features here because we consider them to be relatively central to conscious awareness. They can be summarized in less technical terms as follows. (i) Blindsight patients often say 'no' to the question 'do you detect something being presented?' (ii) When they successfully discriminate things, they seem to be just subjectively guessing, as reflected by the low confidence ratings they give. (iii) Even when we divide these confidence ratings into groups of relatively high and relatively low, such ratings are not very meaningful in that they do not discriminate correct from incorrect trials as accurately as they do in normal subjects. They seem to be placing these ratings relatively randomly. Finally, (iv) their ability to detect things seems to be particularly poor, relative to their ability to discriminate things by forced-choice. We see from this that blindsight is not a phenomenon restricted to solely type I or type II tasks, but instead leaves psychophysical signatures of both types. With the aforementioned descriptions, even without knowing the exact phenomenology of blindsight, one can see that visual awareness is clearly disturbed in these patients; these seem to be the essential psychophysical properties that characterize a disturbance in conscious awareness. Explaining these four psychophysical properties certainly may not be equivalent to explaining consciousness, but it seems to be a good start. Below, we argue that these properties can be parsimoniously explained by a single metacognitive mechanism.

### (b) *A unifying account based on criterion setting*

As discussed earlier, two of the main psychophysical features of blindsight are related to criterion setting (figure 1*a*,*b*): in blindsight, subjects set their criteria for detection and confidence ratings in discrimination to be overly conservative. Here, we give an account of how the two may be related. We also give an intuitive account of how the same mechanisms may also lead to the other two features, namely, reduced type II capacity and lowered detection $d'$.

Given the distributions of internal responses for each stimulus condition (e.g. target present versus target absent), and making certain assumptions regarding the probability of each condition (that they happen equally frequently, for example), one can determine an optimal criterion for maximizing accuracy in detection, i.e. one that maximizes payoff under an even payoff structure (figure 1*a*). Therefore, one possibility could be that subjects voluntarily choose not to use such an optimal criterion, or else they fail to compute this optimal criterion despite the information available. However, a more interesting possibility is that the subject is misinformed regarding the distributions of the internal responses. Note that the representations of the internal response distributions describe how one's internal neural response behaves statistically over time. Such awareness does not come naturally; one has to

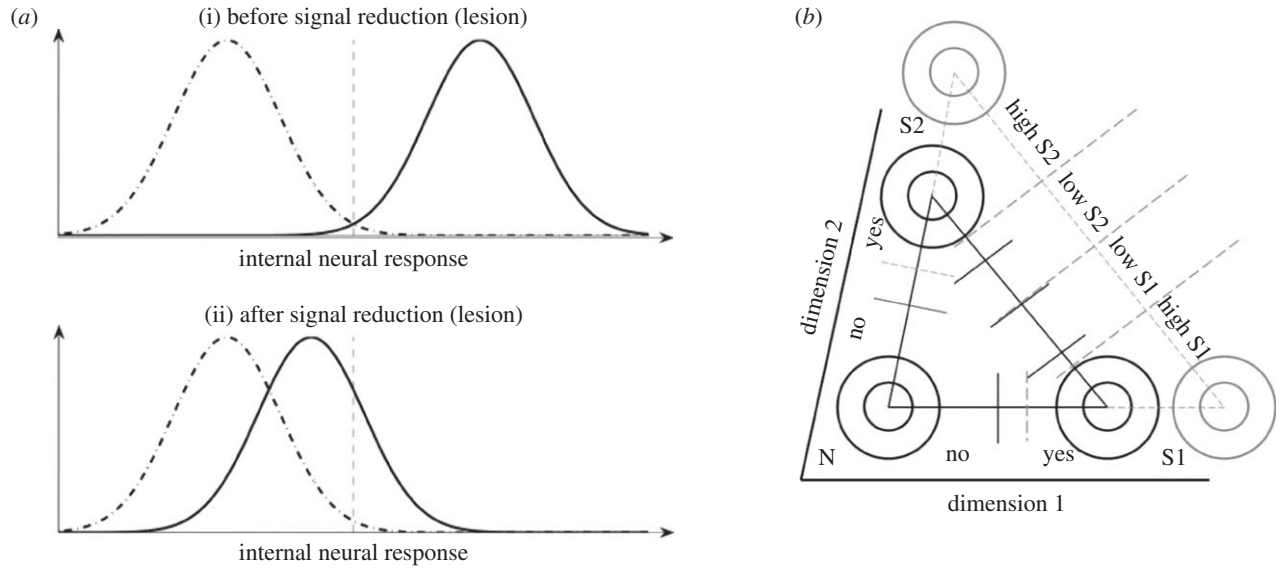


Figure 2. (*a*) A decrease in distribution mean necessitates criterion updating. Dashed dotted and solid Gaussian distributions represent internal neural response distributions for target absent and present conditions, respectively, whereas the dashed line represents the criterion used. An originally optimal criterion (i) becomes conservative when mean response strength for the 'signal present' distribution is reduced (ii). Such may be the case after a lesion to the visual cortex. (*b*) Another schematic depicting the relationship between detection and discrimination (here not necessarily 2AFC; hence the triangle is not right-angled). Overestimating the strength of the signal present distributions causes conservative detection criteria to be used (grey segments on N-S1 and N-S2 axes). In discrimination, this also causes confidence criteria to be placed further away from the decision criterion, as the distributions are represented as further away from each other (grey segments on S1–S2 axis, see also §2). This leads to more trials being classified as low confidence.

learn about it [28]. If one's representations of the internal distributions are not perfect, even if one tries to set an optimal criterion accordingly, then the criterion may turn out to be suboptimal.

Lau argued that this may be the case in blindsight: after a lesion to the primary visual cortex, the overall internal neural response reduces drastically [29,30]. To achieve optimality, one must lower the criterion to match the new distributions (figure 2*a*). But if one fails to learn the new distributions and stubbornly continues to apply the old criterion, then the criterion used becomes too conservative.

If blindsight patients are indeed setting their detection criterion conservatively because they fail to represent the distributions correctly, then this would naturally explain their conservative criteria for high confidence. This is because detection and discrimination tasks are closely related (figure 1*c*); when the internal response for detection drops (after a lesion), the internal response for a related discrimination task drops accordingly. As in detection, if one fails to update changes in the distributions and sets confidence-rating criteria based on the old distributions, the criteria used will be too conservative. In this case, this leads to subjects rating low confidence too frequently (figure 2*b*). Therefore, a single account of suboptimal criterion setting owing to failure to update representations of internal response distributions can account for why blindsight patients are conservative in detection and under-confident in discrimination at the same time.

Let us now turn to the next feature: reduced type II performance in discrimination, i.e. that confidence ratings become less diagnostic of whether a discrimination response was accurate. In general, in normal subjects, high confidence trials are more likely to be correct. One can see this in figure 1*b*: trials that pass the criterion for high confidence are more likely to be correct trials; the overlap between the two distributions in these regions is low. On the other hand, the trials that fail to pass the criterion for high confidence (i.e. trials in the middle region) are less likely to be correct, because this is the region where there is much overlap between the two distributions, thus making it difficult to distinguish between the two stimulus possibilities. Therefore, given the two distributions and decision criteria, if a subject consistently rates high confidence to trials in the outer regions (far left and far right) and low confidence to trials in the middle, the subject's confidence ratings would be diagnostic of whether the discrimination was accurate. On the other hand, if the subject occasionally rates high confidence for trials that fall between the confidence criteria, i.e. in the middle region, and sometimes rates low confidence for trials in the outer regions, then the correlation between confidence and accuracy would decrease. But why would this happen? One possibility is that the confidence criteria are not stationary, but instead jittering around from left to right from trial to trial, as suggested by Azzopardi & Cowey [21]. Effectively, this would be occasionally taking trials from the outer regions to be low confidence (when the criteria happen to be far apart owing to jitter), and occasionally taking trials from the middle region to be high confidence (when the criteria happen to be close to each other); decreased type II performance could thus result from unstable confidence criteria.

At this point, it is important to note that non-stationary estimates are characteristic of learning. When trying to learn the optimal value of a variable,

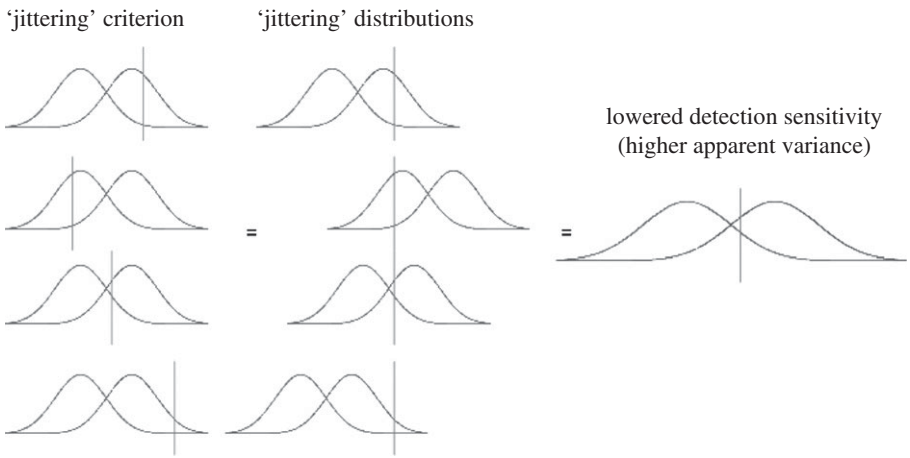


Figure 3. A graphical representation of why jittering the detection criterion leads to underestimation or effective lowering of detection sensitivity. Jittering the detection criterion is effectively identical to jittering the distributions, because with respect to detection behaviour, the scale of the horizontal axis is not critical. Therefore, given four conditions (left column) where different criteria are set, we can conceptualize them as equivalent to four conditions where the criterion is constant, but the distributions themselves move around (middle column). If we average these four conditions over a large number of trials, higher effective variances result for both distributions, thereby lowering detection sensitivity.

one strategy may be to adjust its value from its initial point to see if the results are desirable, and readjust according to the results of the previous step, and so on. Once the optimal value is reached, there is no more need for readjustment; the value becomes stable. Taking this into account, we offer the following possible explanation for jittering confidence criteria: in blindsight, subjects are, in fact, attempting to learn and update the representations for the new internal response distributions, although unsuccessfully. This could be because the brain mechanisms for optimal criterion adjustment may only be able to handle reasonably sized changes in the strength of the internal response. Drastic changes owing to lesions to the primary visual cortex, or perhaps damaged connections to higher cortical areas responsible for learning, could overwhelm the said mechanisms and cause them to fail to converge onto the optimal value; the jittering of the confidence criteria could reflect the mechanisms' continued futile attempts to achieve optimality.

Just as conservative criteria for detection and discrimination confidence are closely connected (because they are both due to failure to represent the internal distributions correctly), jittering of confidence criteria for discrimination should also be coupled with jittering of the detection criterion (because the jitter is caused by failed attempts to update representations of internal response distributions, which will naturally affect both kinds of criteria). Detection criterion jitter can explain the last psychophysical property of blindsight, that detection sensitivity ($d'$) is lower than expected based on the sensitivity of forced-choice discrimination tasks [21].

It is often stated that detection sensitivity ($d'$) and detection criterion are mathematically independent. The former reflects the extent of the overlap between the two distributions for internal response, whereas the latter reflects a decision strategy or rule. However, as we have explained earlier, jittering criteria for confidence ratings can effectively lower type II performance, because jitter decreases the consistency of the subject's decisions. A similar relationship holds between the detection criterion and detection sensitivity. If the criterion for detection jitters, trials that are supposed to be hits will occasionally become misses (when the criterion happens to be too high due to jitter), and trials that are supposed to be correct rejections will occasionally become false alarms (when the criterion happens to be too low). Figure 3 gives an intuitive graphical account of why a jittering detection criterion effectively lowers detection sensitivity.

Indeed, Azzopardi & Cowey [27], who reported that detection sensitivity was lower than expected in a blindsight patient, also reported behavioural evidence pointing towards criterion jitter in detection, and argued that this may be an explanation of why the patient's detection sensitivity was low [21]. It is worth noting, however, that although Azzopardi & Cowey proposed criterion jitter as an explanation for low detection sensitivity, they did not consider this explanation to be related to metacognition, as this aspect of blindsight was derived from type I receiver operator characteristic curves. We argue that these aspects are indeed metacognitive, at least, in a limited sense. Criterion setting depends on representations *about* other internal representations, i.e. knowledge about the statistical distributions of the internal perceptual response.

One may wonder, if criterion jitter effectively lowers detection sensitivity, would it not do the same to discrimination sensitivity as well? If it does, then it cannot account for the finding that detection sensitivity is too low *with respect to* the sensitivity of forced-choice discrimination, because both sensitivities would be reduced. The answer is that indeed, if the criterion for discrimination were to jitter, then sensitivity for discrimination would also be effectively lowered. As it turns out, however, even though detection criteria are unstable because of repeated misrepresentations of internal response distributions, this should not affect the criterion for discrimination, assuming that both stimulus distributions are misrepresented by the same amount. To see this, we can examine figure 1*c*, taking the mean of the noise

distribution to be the origin: assuming equal variances, as long as $d'(A) = d'(B)$, the optimal discrimination criterion always lies on the line $y = x$, that is, exactly halfway between the distributions on the discrimination axis, regardless of the value of $d'(A)$.

To summarize, we have described a possible account for the four psychophysical properties of blindsight described in §4. In short, the explanation is that after a lesion to the primary visual cortex, the internal response strength drops drastically. Perhaps because this drop is so drastic, and perhaps also because the connections between the visual areas and the higher cortical areas responsible for learning are also damaged, the brain fails to 'learn' the new representations of internal response distributions. Because of this, the brain continues to rely on its old, pre-lesion representations, and therefore generate conservative criteria for detection and discrimination confidence. As such a situation is suboptimal, however, the brain continues futilely to attempt to update the representations by trying different values ('jittering') and seeing if the outcome becomes more desirable, i.e. if the decisions and confidence ratings become more optimal. This leads to jittering discrimination confidence criteria, as well as to jittering detection criteria. These translate into lowered type II performance and lowered detection sensitivity ($d'$).

Alternative explanations for conservative detection criterion include the Neyman–Pearson lemma [26], as well as a reliance on distributions from the sighted hemifield, as proposed by Gorea & Sagi [31]. These explanations are not in contradiction to our account, but our account has the advantage of explaining all of the four psychophysical signatures we have discussed within the same mechanism. We do not argue that our account is the only possible explanation, but we submit that it is a parsimonious possibility, partially supported by empirical data [21]. Next, we show in a computer simulation that our account can produce results that are qualitatively similar to actual data.

### (c) *Computer simulations*

In our computer simulations, we sought to construct an algorithm, based on which an observer (the agent in the simulation) could try to set an optimal criterion based on feedback in visual tasks. The observer could thereby implicitly learn the distributions of internal responses. We show that under a small reduction in the average strength of internal responses for the target present condition, the algorithm can learn to update the relevant criteria to achieve optimality. However, if the reduction in internal response strength is too drastic, then the algorithm may fail to update the criteria and internal representations for the distributions. This results in detection and discrimination behaviour that are similar to those observed in blindsight, as described earlier.

Owing to the inefficiency and required resources presumably associated with keeping track of entire distributions (but see Zemel *et al.* [32], which argues that this can perhaps be done efficiently), we used a simple recursive heuristic as a possible mechanism for detection criterion learning. In particular, given a starting criterion, we had the simulation run a small number of detection trials with said criterion, and then calculate the logarithm of the estimated likelihood ratio ($\log \beta$) between the probabilities of the target being present and absent, respectively (see §2). When $\log \beta$ at some value is equal to zero, it means that at that internal response strength, it is equally likely that the target is present or absent. In tasks where the target is presented exactly 50 per cent of the time, a detection criterion set at this point can maximize detection accuracy [26]; the algorithm aims to move the criterion to this optimal point. In each iteration (or 'run'), the criterion was also shifted to the left and the right, and feedback information was collected at those neighbouring points. Log $\beta$ values for the starting criterion were then compared with those to its left and right; the criterion that was found to give a $\log \beta$ value closest to zero (the value at the optimum) was chosen to be the criterion for the next iteration. A certain number of runs were previously determined as the maximum memory capacity of the system; subsequent $\log \beta$ values were calculated using all available data at that criterion within memory (if any), in addition to the new trials performed per run. In reality, we expect that in most cases, any similar algorithm with a substantial memory capacity should be able to track down the optimal criterion with reasonable accuracy; alternative algorithms that should be able to perform the same learning objective include those that attempt to minimize prediction errors (see Rao [33]), but the simple heuristic we have described here does not involve predictive coding.

After each iteration, an implicit representation of the internal response distributions was updated, based on the detection criterion for that iteration. We call these distributions' representations 'implicit' because, as in previous research [34], only the detection criterion itself was explicitly represented. The mean values for the internal response distributions were estimated based on simplifying assumptions (see §2). Using a two-rating ('low' and 'high' confidence) system, confidence criteria for discrimination were also scaled based on the implicit estimates of the internal response distributions; in essence, this is based on the 'triangle' relationship between detection and discrimination as depicted in figure 2*b*. Below, we detail the parameters used in our simulation.

## 2. METHODS

### (a) *Computer simulation*

The computer simulation was programmed in MATLAB, v. 7.13.0 (R2011b), on a Macintosh computer (OS X v. 10.7).

An equal-variance Gaussian model with unit variance was assumed. Because the model assumes unit variance, true detection sensitivity ($d'$) is equal to the distance between the means of the two distributions, in units. Detection sensitivity was originally set at 6, and starting criterion was set halfway between the distributions at three units (value on an arbitrary scale of 'internal response', where zero is the mean for the target absent condition). In scenario 1, $d'$ was reduced three times by increments of 1.833, i.e. from 6 to

4.167, then 2.33, and finally 0.5, whereas in scenario 2, $d'$ was reduced directly to 0.5. True discrimination sensitivity was calculated as merely $\sqrt{2}$ times true detection sensitivity.

In each scenario, the simulation performed 2000 'runs' for each of the following three task types: detection of stimulus $A$, detection of stimulus $B$, and discrimination between $A$ and $B$, where $A$ and $B$ are independent and uncorrelated stimuli. (For detection, this corresponded to 500 runs per sensitivity level for scenario 1, whereas in scenario 2, the first 500 runs were performed at $d' = 6$ and the next 1500 runs were performed at $d' = 0.5$.) Runs were randomly interleaved, for a total of 6000 runs. In each detection run, 25 signal and 25 noise trials were randomly interleaved, and responses were collected using the most recently updated criterion. Fifty further trials (25 signal and 25 noise) were performed using criteria a small interval (0.05 units in this simulation) to the left and right of the most recent criterion, for a total of 150 trials per detection run.

Memory capacity for both scenarios was set at 250 runs. At the end of each detection run, $\log \beta$ (natural logarithm of the likelihood ratio) was calculated for each of the three criteria used using all of the data in memory for each criterion value from previous trials, as below (see Wickens [35] for derivation)

$$\log \beta = \frac{1}{2}(Z^2(f) - Z^2(h)),$$

where $f$ is the false alarm rate, $h$ is the hit rate and $Z$ is the inverse cumulative Gaussian function. The criterion that yielded the smallest absolute value of $\log \beta$ (i.e. closest to zero) was chosen for use in the next detection run. In other words, in each run, the observer moves the criterion among the three neighbouring points, and after a total of 150 trials, chooses the point among the three that is most likely to be optimal. Criteria were tracked separately for detection $A$ and detection $B$ runs.

In cases where $f$ and/or $h$ were zero (thereby rendering $\log \beta$ infinite or undefined), we interpreted this as the brain not receiving any worthwhile information from the run, and thus instructed the simulation to randomly select one of the three criteria as the criterion for the next run. In realistic terms, the interpretation is that at the tails of the distributions, hit rates and false alarm rates become extreme, making it difficult for the brain to collect useful information. The observer may have no other choice but to attempt to explore randomly along the decision dimension in search of the optimal value.

In discrimination runs, the discrimination criterion was placed exactly halfway between the signal distributions (i.e. at the optimal value) for reasons stated in the main text. For the purposes of this simulation, we used a two-confidence-rating system; taking the midpoint between the projected stimulus distributions to be zero, the discrimination confidence criteria were set as follows:

$$\lambda_{\text{conf}} = \pm \frac{\sqrt{2}}{2}\lambda_{\text{det}},$$

where $\lambda$ represent criteria for discrimination confidence ($\lambda_{\text{conf}}$) and detection ($\lambda_{\text{det}}$). Essentially, this equation places confidence criteria exactly halfway between the optimal discrimination criterion and the projected stimulus distributions. This divided the discrimination axis into four parts, with the outer parts representing 'high confidence' and the inner parts representing 'low confidence' (figure 1*b*).

With the confidence criteria in place, 50 discrimination trials (stimuli $A$ and $B$ were presented 25 times each, randomly interleaved) were performed in each run, and responses and confidence ratings were collected.

### (b) *Data analysis*

For both scenarios, we calculated detection and discrimination accuracies; detection and discrimination sensitivities; discrimination confidence–accuracy correlation; percentage of 'low' confidence-rating responses; and meta-$d'$, a bias-free measure that assesses the correspondence between accuracy and confidence [36]. All measures except for confidence–accuracy correlation and meta-$d'$ were assessed every 1000 trials. Degrees of freedom for Student's $t$-tests result from the fact that each measure was calculated a total of 100 times; comparisons in the main text between scenario 1 after the third reduction and scenario 2 after the first and only reduction thus entailed $25 + 75 = 100$ observations. Confidence–accuracy correlation and meta-$d'$ were calculated over all post-reduction trials in each scenario.

Accuracy was defined as the sum of the number of hits and the number of correct rejections divided by total number of trials. Measured detection and discrimination sensitivity were calculated using the well-known result, as follows

$$d' = Z(h) - Z(f).$$

Undefined or infinite $d'$ values were taken as missing data points. Confidence–accuracy correlation refers to the linear correlation between discrimination confidence ('low' = 1, 'high' = 2) and accuracy (incorrect = 0, correct = 1) across the trials in question using Pearson's product–moment correlation coefficient. 'Low' confidence percentage was defined as simply the number of 'low' confidence ratings occurring during the runs in question divided by the total number of trials during the said runs. Finally, meta-$d'$ was calculated using the type 2 SDT analysis functions found at http://www.columbia.edu/~bsm2105/type2sdt/.

## 3. RESULTS

We show that in scenario 1, the algorithm successfully learns to update the criterion to the optimal location after each small reduction (figure 4*a*), giving good values for detection sensitivity and measures of type II discrimination performance (figure 5). In scenario 2 ('blindsight'), however, we found that the algorithm fails to place the criterion in the optimal location, merely jittering it around its starting location (figure 4*b*). The effective detection sensitivity and type II discrimination performance are correspondingly reduced, and

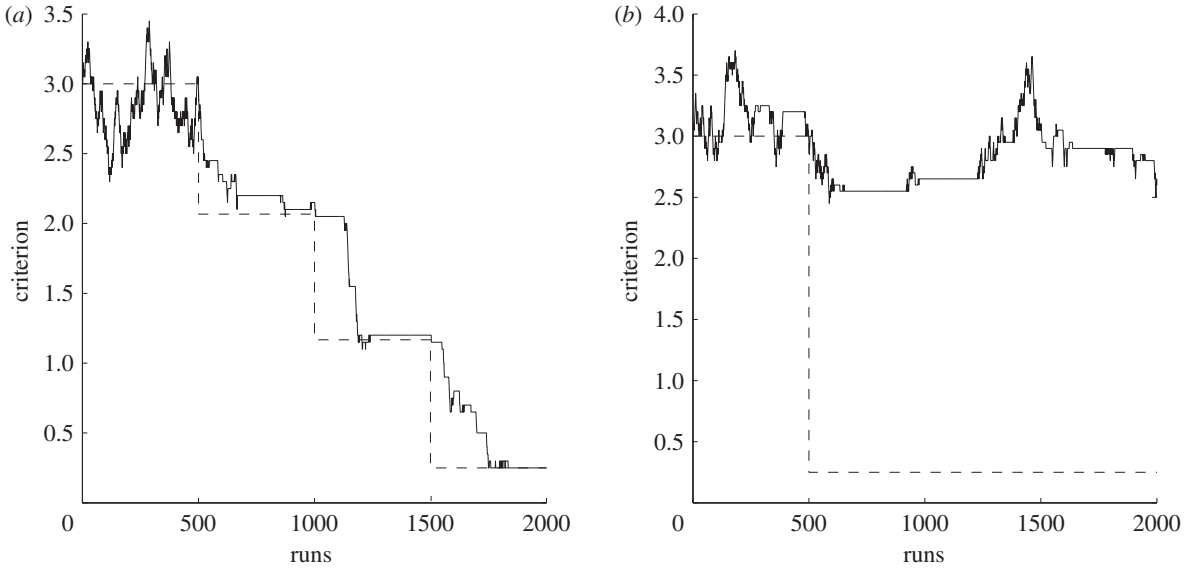


Figure 4. (*a*) Scenario 1: detection criterion (here for stimulus *A*) approaches and remains near optimal value with repeated iterations of the learning algorithm, following each small reduction in the strength of the target present distribution. (*b*) Scenario 2 (blindsight): detection criterion jitters around initial value even after repeatedly applying the learning algorithm, following a drastic reduction in the strength of the target present distribution. Dashed lines represent optimal value.

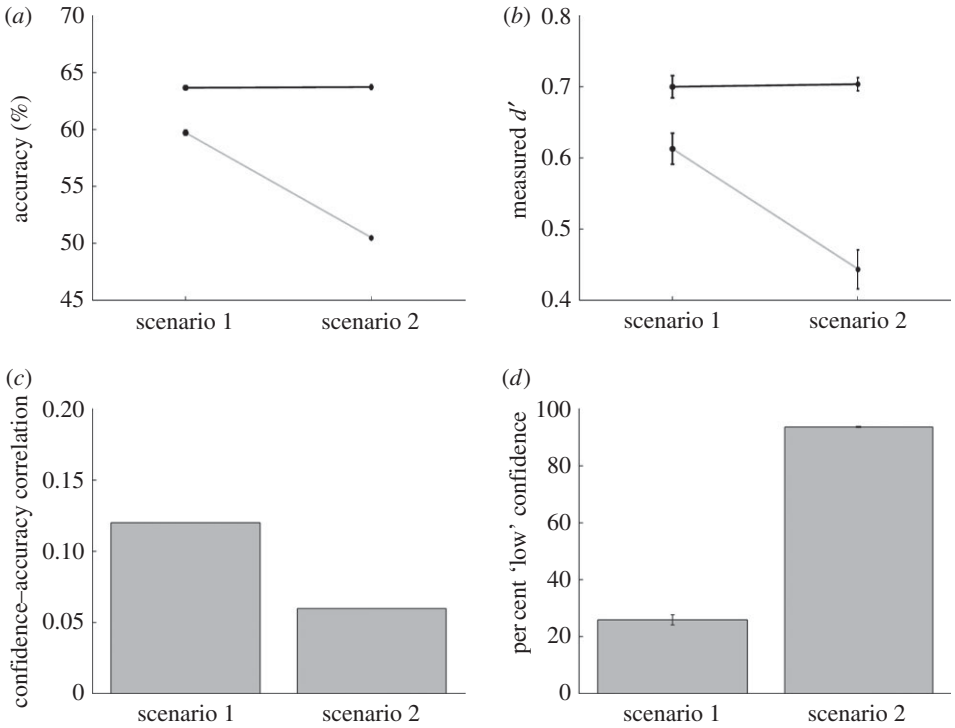


Figure 5. (*a*) Detection accuracy (grey line with circles) is high in scenario 1 and near chance in scenario 2, whereas discrimination accuracy (black line with circles) is significantly above-chance in both scenarios. (*b*) Mean-measured detection sensitivity (grey line with circles) is low with respect to discrimination sensitivity (black line with circles) in scenario 2 (see §3) but is as expected in scenario 1. (*c*) The correlation between confidence rating and response accuracy is significantly lower in scenario 2 compared with scenario 1. (*d*) Mean percentage of 'low' confidence-rating responses is much higher in scenario 2 compared with scenario 1. Error bars represent $\pm 1$ s.e.m.; not shown when smaller than data point.

average confidence rating is also much lower than in scenario 1 (figure 5).

Here, we compare the data for scenario 1 after the third reduction against the data for scenario 2 after the first and only reduction, thus matching for true detection and discrimination sensitivity. Detection accuracy was significantly higher in scenario 1 (59.7%) than in scenario 2 (50.5%), $t(98) = 42.55$,

$p < 10^{-64}$, whereas discrimination accuracy was virtually identical and remained above-chance in both scenarios (scenario $1 = 63.7\%$, $t(24) = 46.48$, $p < 10^{-24}$; scenario $2 = 63.7\%$, $t(74) = 66.18$, $p < 10^{-66}$; figure 5*a*). Measured discrimination sensitivity did not differ from the true underlying value ($d' = 0.71$) for either scenario, which was to be expected, as the discrimination decision criterion was hard-coded to be placed at the optimum. On the other hand, measured detection sensitivity was significantly lower than the value predicted by discrimination sensitivity ($d' = 0.5$) in scenario 2 ($d' = 0.44$, $t(62) = -2.07$, $p = 0.04$), whereas in scenario 1, measured detection sensitivity was indistinguishable from the predicted value ($d' = 0.61$, $t(24) = 1.25$, $p = 0.23$; figure 5*b*). For the discrimination task, the correlation between confidence rating and response accuracy over all post-reduction trials was lower in scenario 2 ($r = 0.06$) than in scenario 1 ($r = 0.12$; figure 5*c*). The percentage of 'low' confidence responses was significantly higher in scenario 2 (93.7%) than in scenario 1 (25.9%; figure 5*d*), $t(98) = 38.24$, $p < 10^{-59}$.

The fact that the confidence–accuracy correlation was small in both scenario 1 and 2 is probably because such correlation is largely dependent on discrimination sensitivity ($d'$) [24]. Although the values differed between scenario 1 and 2, it is not straightforward to assess whether the size of the effect was large or small, because the absolute value of the correlation is not easy to interpret. In order to also control for effects of type I and II criteria [24], as an additional measure of type II discrimination performance, we applied the bias-free measure meta-$d'$ [36], which also assesses the correspondence between accuracy and confidence. Because meta-$d'$ and $d'$ are measured on the same scale [36], we were able to compare meta-$d'$ trials with different discrimination $d'$ by calculating the ratio between meta-$d'$ and $d'$. A ratio of exactly 1 would mean that one's confidence ratings are perfectly diagnostic of one's response accuracy, whereas a ratio of less than 1 would indicate reduced type II capacity; a ratio of zero would indicate no metacognitive capacity whatsoever. After the reductions in signal strength, the ratio values were 0.93 and 0.75, respectively, for scenario 1 and 2. This suggests that, in scenario 2, there was considerable drop (*ca* 20%) in metacognitive efficiency.

The simulation results show that the model we have described in this study, with its conservative jittering criteria caused by signal strength reduction and failed learning (scenario 2), can qualitatively account for these observed properties of blindsight: (i) subjects perform virtually at chance for detection tasks, but can perform above-chance in a discrimination (2AFC) task; (ii) subjects' confidence ratings tend to be low in discrimination; (iii) are not very diagnostic of accurate discrimination responses; and (iv) subjects' detection sensitivity is lower than expected considering their discrimination sensitivity.

However, it is important to note that this is a very simplistic simulation. Some obvious limitations of the model are as follows: the learning algorithm calculates $\log \beta$ based on hit and false alarm rates (see §2); this assumes the presence of feedback. Moreover, in cases where the criterion is very conservative, in earlier iterations hit and/or false alarm rates may well be zero, causing $\log \beta$ to become infinite or undefined. To resolve this, we resorted to a simple work-around explained in §2. Also, the slow speed with which the $\log \beta$ algorithm (with our parameters) attains the optimal value suggests that the existence of multiple complementary learning mechanisms certainly cannot be discounted. Another point of concern is that the algorithm does not necessarily learn the optimal criterion at the same speed for two separate detection tasks (i.e. for stimulus *A* versus for stimulus *B*), and thus at some intermediate stage the criterion used for each task may be different. This, in turn, may confound the placement of confidence criteria for discrimination and affect measures of type II discrimination performance during learning. Finally, our model assumes equal variance for noise and stimulus distributions, as well as equal variance before and after signal reduction (i.e. lesion). This, of course, may not be the case in reality, and though it is still possible to project the location of the distributions given these variance parameters, the brain may not have access to such information (e.g. post-lesion signal variance), at least at first. With the appropriate heuristics, these can be learned, however, and so we believe that our model can still be adapted to unequal-variance cases.

With these limitations, we certainly do not claim that our model is the best or a biologically realistic account of blindsight. However, here we have demonstrated a basic 'proof of concept', to show that psychophysical properties of blindsight can be explained in terms of criterion learning within a unifying framework.

## 4. CONCLUDING REMARKS

Here, we have presented a signal detection theoretic account of some essential psychophysical features of blindsight. In this account, though blindsight is typically a result of lesion to the primary visual cortex, the underlying mechanism is essentially 'higher-order' in a certain sense: it is not damage to the primary visual cortex itself that drives the phenomenology of blindsight, but rather, the fact that damage leads to suboptimal criterion setting, which itself is a process that may depend on the prefrontal cortex [5].

Pasquali and co-workers [37] have also developed a formal model of blindsight that is similar in spirit to the model presented here. They trained a connectionist model to place wagers on its own performance (akin to giving confidence ratings). They showed that after damage to the early sensory system, the model can simulate the unusual pattern-wagering behaviour observed in an actual blindsight patient. Compared with our signal detection theoretic approach, their model is perhaps more mechanistic and accounts for the mechanism of how one learns to represent the statistical behaviour of the internal sensory response. However, our account deals with various psychophysical features, such as detection criteria and sensitivity, and is closer to certain data [27]. Thus, the two approaches complement each other and offer different levels of explanation.

Recent advances in anatomical and physiological studies in blindsight [10–17], especially those in non-human primates [8,9], shed light on the underlying mechanism. However, most of these studies focus on early sensory mechanisms, and/or how subcortical activity may support blindsight after the cortex is damaged. On the other hand, brain imaging studies suggest that in humans, blindsight may critically depend on activity in the prefrontal cortex [14,15,38]. Together with the account presented here, this suggests that future studies could perhaps focus more on 'higher-order' mechanisms that are responsible for criterion setting and perceptual decision-making, as they may be key to understanding some essential aspects of blindsight.

H.L. is supported by the Templeton Foundation (grant no. 21569). We thank the editors Steve Fleming and Chris Frith, an anonymous reviewer and Brian Maniscalco for helpful comments.

## REFERENCES


1 Weiskrantz, L. 2004 Roots of blindsight. *Prog. Brain Res.* **144**, 229–241.
2 Weiskrantz, L. 1997 *Consciousness lost and found: a neuropsychological exploration*. New York, NY: Oxford University Press.
3 Stoerig, P. & Cowey, A. 2007 Blindsight. *Curr. Biol.* **17**, R822–R824. (doi:10.1016/j.cub.2007.07.016)
4 Breitmeyer, B. G. 1984 *Visual masking: an integrative approach*. New York, NY: Oxford University Press.
5 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)
6 Lau, H. C. 2008 Are we studying consciousness yet? In *Frontiers of consciousness: Chichele Lectures* (eds L. Weiskrantz & M. David), pp. 245–258. New York, NY: Oxford University Press.
7 Weiskrantz, L., Barbur, J. L. & Sahraie, A. 1995 Parameters affecting conscious versus unconscious visual discrimination with damage to the visual cortex (V1). *Proc. Natl Acad. Sci. USA* **92**, 6122–6126. (doi:10.1073/pnas.92.13.6122)
8 Schmid, M. C., Mrowka, S. W., Turchi, J., Saunders, R. C., Wilke, M., Peters, A. J., Ye, F. Q. & Leopold, D. A. 2010 Blindsight depends on the lateral geniculate nucleus. *Nature* **466**, 373–377. (doi:10.1038/nature09179)
9 Isa, T. & Yoshida, M. 2009 Saccade control after V1 lesion revisited. *Curr. Opin. Neurobiol.* **19**, 608–614. (doi:10.1016/j.conb.2009.10.014)
10 Yoshida, M., Takaura, K., Kato, R., Ikeda, T. & Isa, T. 2008 Striate cortical lesions affect deliberate decision and control of saccade: implication for blindsight. *J. Neurosci.* **28**, 10 517–10 530. (doi:10.1523/JNEUROSCI.1973-08.2008)
11 Radoeva, P. D., Prasad, S., Brainard, D. H. & Aguire, G. K. 2008 Neural activity within area V1 reflects unconscious visual performance in a case of blindsight. *J. Cogn. Neurosci.* **20**, 1927–1939. (doi:10.1162/jocn.2008.20139)
12 Ptito, A. & Leh, S. E. 2007 Neural substrates of blindsight after hemispherectomy. *Neuroscientist* **13**, 506–518. (doi:10.1177/1073858407300598)
13 Stoerig, P., Kleinschmidt, A. & Frahm, J. 1998 No visual responses in denervated V1: high-resolution functional magnetic resonance imaging of a blindsight patient. *Neuroreport* **9**, 21–25. (doi:10.1097/00001756-199801050-00005)
14 Sahraie, A., Weiskrantz, L., Barbur, J. L., Simmons, A., Williams, S. C. & Brammer, M.J. 1997 Pattern of neuronal activity associated with conscious and unconscious processing of visual signals. *Proc. Natl Acad. Sci. USA* **94**, 9406–9411. (doi:10.1073/pnas.94.17.9406)
15 Persaud, N., Davidson, M., Maniscalco, B., Mobbs, D., Passingham, R. E., Cowey, A. & Lau, H. 2011 Awareness-related activity in prefrontal and parietal cortices in blindsight reflects more than superior visual performance. *Neuroimage* **58**, 605–611. (doi:10.1016/jn.neuroimage.2011.06.081)
16 Bridge, H., Thomas, O., Jbabdi, S. & Cowey, A. 2008 Changes in connectivity after visual cortical brain damage underlie altered visual function. *Brain* **131**, 1433–1444. (doi:10.1093/brain/awn063)
17 Leh, S. E., Johansen-Berg, H. & Ptito, A. 2006 Unconscious vision: new insights into the neuronal correlate of blindsight using diffusion tractography. *Brain* **129**, 1822–1832. (doi:10.1093/brain/awl111)
18 Lamme, V. A. 2003 Why visual attention and awareness are different. *Trends Cogn. Sci.* **7**, 12–18. (doi:10.1016/S1364-6613(02)00013-X)
19 Dehaene, S. 2008 Conscious and nonconscious processes. Distinct forms of evidence accumulation? In *Strüngmann forum report. Better than conscious? Decision making, the human mind, and implications for institutions* (eds C. Engel & W. Singer), pp. 21–49. Cambridge, MA: MIT Press.
20 Green, D. M. & Swets, J. A. 1966 *Signal detection theory and psychophysics*. Oxford, UK: John Wiley.
21 Azzopardi, P. & Cowey, A. 2001 Why is blindsight blind? In *Varieties of unconscious processing: new findings and models* (eds B. de Gelder, E. de Haan & C. Heywood), pp. 3–18. Oxford, UK: Oxford University Press.
22 Morgan, M. J. 2002 Detecting the wrong signals? *Trends Cogn. Sci.* **6**, 443–445. (doi:10.1016/S1364-6613(02)01899-5)
23 Campion, J. & Latto, R. 1983 Apperceptive agnosia due to carbon monoxide poisoning. An interpretation based on critical band masking from disseminated lesions. *Behav. Brain Res.* **15**, 227–240. (doi:10.1016/0166-4328(85)90177-9)
24 Galvin, S. J., Podd, J. V., Drga, V. & Whitmore, J. 2003 Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychon. Bull. Rev.* **10**, 843–876. (doi:10.3758/BF03196546)
25 Persaud, N., McLeod, P. & Cowey, A. 2007 Post-decision wagering objectively measures awareness. *Nat. Neurosci.* **10**, 257–261. (doi:10.1038/nn1840)
26 Macmillan, N. A. & Creelman, C. D. 2005 *Detection theory: a user's guide*, 2nd edn. Cambridge, UK: Cambridge University Press.
27 Azzopardi, P. & Cowey, A. 1997 Is blindsight like normal, near-threshold vision? *Proc. Natl Acad. Sci. USA* **94**, 14 190–14 194. (doi:10.1073/pnas.94.25.14190)
28 Timmermans, B., Schilbach, L., Pasquali, A. & Cleeremans, A. 2012 Higher-order thoughts in action: consciousness as an unconscious re-description process. *Phil. Trans. R. Soc. B* **367**, 1412–1423. (doi:10.1098/rstb.2011.0421)
29 Lau, H. C. 2008 A higher order Bayesian decision theory of consciousness. *Prog. Brain Res.* **168**, 35–48. (doi:10.1016/S0079-6123(07)68004-2)
30 Azzopadi, P., Fallah, M., Gross, C. G. & Rodman, H. R. 2003 Response latencies of neurons in visual areas MT

and MST of monkeys with striate cortex lesions. *Neuropsychologia* **41**, 1738–1756. (doi:10.1016/S0028-3932 (03)00176-3)
31 Gorea, A. & Sagi, D. 2000 Failure to handle more than one internal representation in visual detection tasks. *Proc. Natl Acad. Sci. USA* **97**, 12 380–12 384. (doi:10.1073/pnas.97.22.12380)
32 Zemel, R.S., Dayan, P. & Pouget, A. 1998 Probabilistic interpretation of population codes. *Neural Comput.* **10**, 403–430. (doi:10.1162/089976698300017818)
33 Rao, R. P. N. 2010 Decision making under uncertainty: a neural model based on partially observable Markov decision processes. *Front. Comput. Neurosci.* **4**, 1–18. (doi:10.3389/fncom.2010.00146)
34 Treisman, M. & Williams, T. C. 1984 A theory of criterion setting with an application to sequential dependencies. *Psychol. Rev.* **91**, 68–111. (doi:10.1037/0033-295X.91.1.68)
35 Wickens, T. D. 2001 *Elementary signal detection theory.* New York, NY: Oxford University Press.
36 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)
37 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)
38 Azzopardi, P. & Cowey, A. 2002 Cerebral activity related to guessing and attention during a visual detection task. *Cortex* **38**, 833–836. (doi:10.1016/S0010-9452(08)70051-0)