A model of subjective report and objective discrimination as categorical decisions in a vast representational space

Jean-Rémi King, Stanislas Dehaene

▶ **To cite this version:**

Jean-Rémi King, Stanislas Dehaene. A model of subjective report and objective discrimination as categorical decisions in a vast representational space. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences (1934–1990), 2014, 369 (1641), 10.1098/rstb.2013.0204 . hal-01212024

HAL Id: hal-01212024
https://hal.science/hal-01212024

Submitted on 12 Oct 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Model of Subjective Report and Objective Discrimination as Categorical Decisions in a Vast Representational Space

King, J.R.[1,2,3], Dehaene, S.[1,2,4,5]

Affiliations:

1. Cognitive Neuroimaging Unit, Institut National de la Santé et de la Recherche Médicale, U992, F-91191 Gif/Yvette, France ; 2. NeuroSpin Center, Institute of BioImaging Commissariat à l'Energie Atomique, F-91191 Gif/Yvette, France ; 3. Institut du Cerveau et de la Moelle Épinière Research Center, Institut National de la Santé et de la Recherche Médicale, U975 Paris, France ; 4. Université Paris 11, Orsay, France ; 5. Collège de France, F-75005 Paris, France

Corresponding authors:

Jean-Rémi KING: jeanremi.king@gmail.com

Stanislas DEHAENE: stanislas.dehaene@cea.fr

PROOFS

Abstract

Subliminal perception studies have shown that one can objectively discriminate a stimulus without subjectively perceiving it. We show how a minimalist framework based on Signal Detection Theory and Bayesian inference can account for this dissociation, by describing subjective and objective tasks with similar decision-theoretic mechanisms. Each of these tasks relies on distinct response classes, and therefore distinct priors and decision boundaries. As a result, they may reach different conclusions. By formalizing, within the same framework, forced-choice discrimination responses, subjective visibility reports and confidence ratings, we show that this decision model suffices to account for several classical characteristics of conscious and unconscious perception. Furthermore, the model provides a set of original predictions on the non-linear profiles of discrimination performance obtained at various levels of visibility. We successfully test one such prediction in a novel experiment: when varying continuously the degree of perceptual ambiguity between two visual symbols presented at perceptual threshold, identification performance varies quasi-linearly when the stimulus is unseen and in an "all-or-none" manner when it is seen. The present model highlights how conscious and non-conscious decisions may correspond to distinct categorizations of the same stimulus encoded by a high-dimensional neuronal population vector.

Keywords: Signal Detection Theory, subliminal, subjective reports, metacognition, consciousness, 2 alternative forced choice.

INTRODUCTION

Since Helmholtz (1867/1910)'s proposal of perception as unconscious inference, several computational models have been put forward to describe the mechanisms of this process (KERSTEN ET AL., 2004; KNILL AND RICHARDS, 2008). The hypothesis that perception corresponds to an inferential decision on sensory data has received support from neurophysiological recordings during perceptual tasks (POUGET ET AL., 2002; FRISTON, 2010). For instance, intracranial (GOLD AND SHADLEN, 2000) and scalp recordings (DE LANGE ET AL., 2011; WYART ET AL., 2012) have revealed a neural response seemingly reflecting the accumulation of sensory evidence following the presentation of a stimulus and which may predict how subjects perceive the stimulus (SHADLEN ET AL., 2008).

Nevertheless, superficially at least, conscious perception does not always seem to obey the logic of optimal perceptual inference. For instance, one can objectively discriminate a stimulus at above-chance level while subjectively claiming not to have seen it (DEHAENE ET AL., 2006; DEHAENE AND CHANGEUX, 2011). This paradoxical dissociation, referred to as "subliminal perception", has nourished a vast body of philosophical and scientific proposals on the nature of conscious and unconscious perception. For instance, Tononi and Edelman (TONONI AND EDELMAN, 1998) have argued that conscious processes are *quantitatively* more complex, integrated and differentiated, than unconscious processes. Lau (LAU, 2008) and Rosenthal (ROSENTHAL, 1997) claim that conscious perception is *qualitatively* different from unconscious perception as it relies on higher-order metacognitive representations. Recent empirical studies challenge these accounts, however. First, subliminal stimuli can recruit complex semantic and integrative processes (GREENWALD ET AL., 1996; DEHAENE ET AL., 1998; KOUIDER AND DEHAENE, 2007). Second, even second-order metacognitive inferences can apparently be performed above chance on unseen stimuli (KANAI ET AL., 2010; CHARLES ET AL., 2013).

Here, building upon earlier proposals (LAU, 2008; SHADLEN ET AL., 2008), we explore a simple theoretical idea: objective and subjective tasks rely on the same inference principles, but they differ in the nature and the size of the decision space. Our proposal stems from Signal Detection Theory (SDT) and outlines how a minimal extension of the classic unidimensional depiction of SDT to multiple dimensions provides geometrical intuitions on several empirical findings in conscious and unconscious perception.

Specifically, we identified six major sets of empirical findings that should be accounted for:

- Stimuli which are subjectively reported as "unseen" can nevertheless be objectively discriminated above chance in a two-alternative forced-choice task (WEISKRANTZ, 1986; MARSHALL AND HALLIGAN, 1993; DRIVER ET AL., 2001; DEHAENE ET AL., 2006; KOUIDER AND DEHAENE, 2007; STOERIG AND COWEY, 2009).
- Discrimination performance is typically better on seen than on unseen trials, even when sensory stimuli are physically identical (LAU AND PASSINGHAM, 2006; DEL CUL ET AL., 2007; NEUROSCIENCE ET AL., 2012).
- Experimental paradigms can be designed in which objective discrimination performance is identical, while subjective visibility differs (LAU AND PASSINGHAM, 2006; LAU, 2008; RAHNEV ET AL., 2011).

PROOFS

- Subjective reports vary non-linearly as a function of sensory strength. For instance, brief or faint visual stimuli are generally reported as "completely unseen", but once their duration or contrast reaches a threshold level, subjects tend to report items as "clearly seen" (SERGENT AND DEHAENE, 2004; SERGENT ET AL., 2005; DEL CUL ET AL., 2007; MELLONI ET AL., 2011; NEUROSCIENCE ET AL., 2012).
- Prior knowledge increases the subjective visibility of physically identical stimuli (SIMONS AND CHABRIS, 1999; CUL ET AL., 2006; MELLONI ET AL., 2011; PITTS ET AL., 2012).
- Attention generally increases subjective visibility, but has also been found to decrease it (DEHAENE ET AL., 2006; RAHNEV ET AL., 2011).

MODEL

GENERAL ASSUMPTIONS

Our first assumption is that incoming stimuli are encoded as *continuous vectors in a vast representational space*. In the visual domain, for instance, a hierarchy of specialized visual processors decompose any visual scene into a broad variety of features that range from low-level (line orientation, contrast, colour etc) to higher-level attributes (face/non-face, etc). Each of these features may be encoded by the firing rate of a group of neurons. Mathematically, each stimulus is therefore encoded by a set of coordinates, one for each feature dimension (Figure 1.a).

Second, *stimulus strength* is assumed to be directly reflected in the length (*i.e.* the norm) of the input vector. This assumption corresponds to the observation that the depth of sensory encoding varies with the quality of the incoming stimulus: a briefly flashed and masked stimulus only evokes modest activity in higher visual cortices (SERGENT ET AL., 2005; DEL CUL ET AL., 2007), and thus, its internal vector has a small projection, particularly on high-level dimensions. Conversely, an unmasked high-contrasted image results in a long internal vector (Figure 1.a).

Our third assumption is that each behavioural task imposes, in a top-down manner, a *categorical structure of classes* to this continuous vector space (*e.g.* "click left for faces, and right for non-faces"). Performing the task consists in identifying, on every trial, the class in which the input vector falls. Formally, this is a statistical inference problem: in order to perform optimally, given a sensory input and prior knowledge, subjects should attempt to compute the posterior probability of each of the classes in order to select the class with the maximum *a posteriori* (MAP), which is the one most likely to be correct. Each task imposes distinct, possibly overlapping response classes, and may therefore lead to different answers.

Our fourth assumption is that the *content of conscious perception*, which can be reported verbally, is the outcome of such an inferential decision process, but with the specific characteristic of having a very rich set of classes. While simple binary decisions may be performed non-consciously (*e.g.* press right or press left (DEHAENE ET AL., 1998)), the inference system that underlies conscious perception must remain constantly open to myriads of possible contents, including unexpected ones (*e.g.* a fire alarm). We propose that what the subject experiences as a conscious

PROOFS

percept is the class with the highest posterior probability, amongst *all* possible classes. As we shall see, "negative" classes, such as "I didn't see anything", must be considered too.

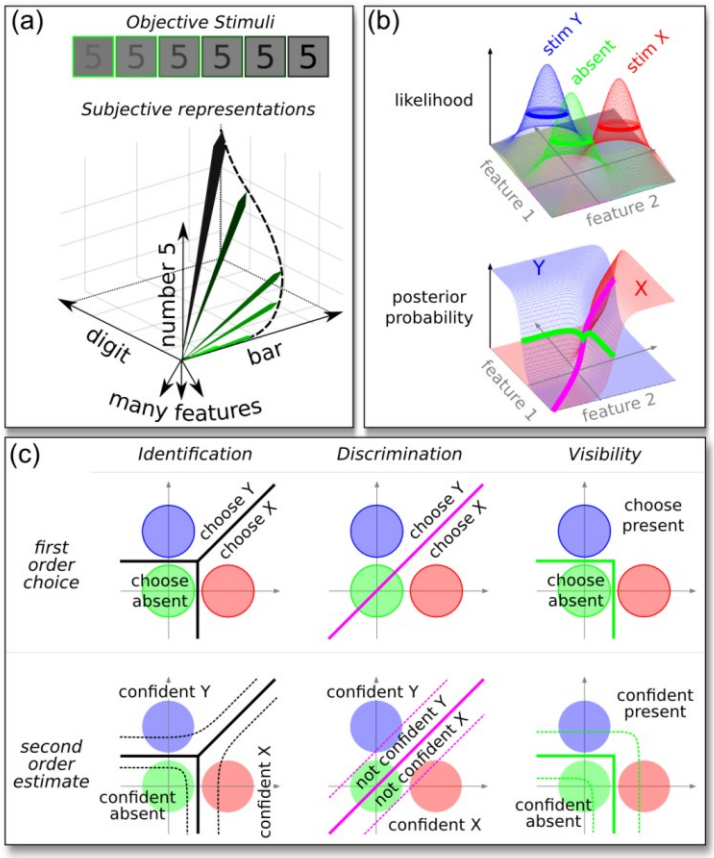

FIGURE 1. A MULTIDIMENSIONAL DECISION-THEORY FRAMEWORK FOR OBJECTIVE DISCRIMINATION AND SUBJECTIVE REPORTS.

(a) Stimulus information is represented in a vast vector space, in which each dimension encodes the evidence about a particular feature. Each sensory stimulus thus corresponds to an input vector whose length and direction changes depending on the quality of the stimulus.

(b) When considering binary decisions (e.g. perceiving stimulus X or stimulus Y), the huge dimensionality of the representational space can be approximated by a 2D feature space. In this space, assuming that the true stimulus distributions are known, the likelihood (top), the prior and the posterior probability (bottom) of belonging to a given class ("Absent" trial in green, stimulus X in red, or stimulus Y in blue) can be computed for each input vector (here, the posterior probabilities of the Absent class have been removed for readability.)

(c) Posteriors can be used to perform different tasks. In each case, the regions of the problem space corresponding to a fixed decision are delineated by a boundary. ***Identification*** *consists in finding the maximum* a posteriori *(MAP) across all classes (Absent, X, or Y ; black lines).* ***Discrimination*** *consists in determining the MAP amongst a restricted set of classes (X or Y; purple line).* ***Visibility judgment*** *consists in determining whether the absent class is the most likely amongst all classes ("absent" or not "absent"; green line). Each of these first-order decisions can be supplemented by a second-order confidence judgment task, which is modelled as the estimation of the likelihood of a correct response in the primary task. Samples far away from the decision border are associated with higher posterior probabilities of the corresponding class and can thus be classified as more "confident" than samples close to the border. This geometrical representation makes it clear that each confidence judgment is always attached to a specific task, and is thus not necessarily identical to visibility judgment.*

Note that the present colour coding (classes, tasks, etc.) will be used throughout the figures.

GEOMETRICAL APPROXIMATION IN TWO DIMENSIONS

The vast number of input features, classes and tasks makes the present proposal difficult to apprehend in its full generality. However, most of its properties can be approximately captured by projecting the large vector space onto a plane defined by the two main axes of interest (Figure 1.b-c). These axes are chosen to be two features or feature bundles that are most relevant to the task under consideration (*e.g.* the mean vectors of neuronal activity evoked by face and by non-face stimuli, if the task is face/non-face discrimination). Each circle represents the top of the distribution of a particular class of stimuli (*i.e.* likelihood function, given sensory and internal noise). The lines delimit the regions of space where response decisions change. Although one should not forget that this is just a considerable simplification of the underlying multidimensional space and stimuli distribution, this 2D representation brings the present model closer to the classic two-class problem of Signal Detection Theory. Indeed, although Signal Detection Theory is not limited to a single dimension, it is often depicted as a binary problem with two Gaussian distributions plotted along a single axis. We argue that this classic diagram fails to capture the interaction between multiple features, classes and tasks, whereas a 2D depiction fulfils these requirements (see (GREEN AND SWETS, 1966; KLEIN, 1985; KO AND LAU, 2012) for similar proposals using 2D representations to dissociate tasks such as discrimination and detection).

MATHEMATICAL FORMULATION

Bayesian theory describes the optimal way of selecting the most likely model of the environment, referred to as "hypothesis" (H, here the response class), in the presence of sensory evidence (E), here the input vector. Each class is characterized by a likelihood function P(E|H) and a prior probability P(H). P(E|H) indicates the probability that the evidence E was generated by the class H, and therefore captures how sensory samples from a given class are distributed within the vector space. The prior probability P(H) defines the probability of H to occur independently of any evidence. Bayes' theorem stipulates that the posterior probability of H is a function of its prior probability and of its likelihood: $P(H|E) = P(E|H) * P(H) / P(E)$. Finally, decisions result from the selection of the class that has the maximum posterior probability (MAP). This MAP criterion results in the segregation of representational space into distinct regions separated by sharp decision boundaries (importantly, the placement of these boundaries does not constitute an additional hypothesis of the model, but derives directly from the hypothesis that decisions are based on a MAP criterion).

In the following simulations, we use a series of computational simplifications. First, we neglect the cost function associated with each decision – sometimes referred to as "loss" or "utility" function. In the presence of costs, the optimal decision is the ones which minimizes the expected loss and may differ from the MAP. Mathematically, however, priors and costs play a similar role and were thus merged in the present paper for simplicity. Second, the present model assumes that priors are fixed in a given context, rather than continuously updated after each decision. Assuming modifiable priors would lead to important new predictions, but would also increase the number of ad-hoc parameters in the models (*e.g.* learning rate, estimated world volatility, creation or deletion of classes etc.). Third, we assume Gaussian distributions in order to facilitate the computations. Fourth, importantly, we assume that subjects have an accurate estimate of stimulus distributions – although following Lau (LAU, 2008; KO AND LAU, 2012), we will discuss the

PROOFS

important consequences that ensue when subjects' priors and likelihood functions are inappropriately calibrated. Fifth, we assume that, on a given trial, the same input vector enters into different tasks, thus neglecting the possibility that the internal evidence evoked by a fixed stimulus may vary with the task, due for instance to decay (GREENWALD ET AL., 1996; DUPOUX ET AL., 2008), noise level (DEL CUL ET AL., 2007), attention (SERGENT ET AL., 2013) or other top-down changes. Finally, we treat stimulus evidence on a given trial as a single discrete point in the n-dimensional space. In the discussion, we briefly examine the additional properties that arise if these simplifying assumptions are relaxed.

THE FUNDAMENTAL THREE-CLASS PROBLEM

Given these assumptions, binary decision experiments can be simplified to a stereotypical three-class problem: either nothing is presented ("Absent" trial), or one of two stimuli X or Y is displayed (Figure 1.b). Absent trials are assumed to correspond to a null vector whose likelihood function peaks at the origin of vector space. X and Y trials are represented by two base vectors which are chosen as the axes of the 2D representation.

In this typical setup, three different tasks can be performed (Figure 1.c):

i) Identification consists in determining which hypothesis has the highest posterior probability (Absent, X, or Y?).
ii) Forced-choice discrimination consists in restricting the responses to a subset of classes (*e.g.* X or Y, excluding the Absent class).
iii) Visibility judgment consists in reporting whether the stimulus is seen or unseen. We assume that this instruction is interpreted as a decision, whether the stimulus is most likely to be absent or present (*i.e. Absent* or not Absent?).

Formally, these are all first-order tasks, because they all ask a simple question: which class (or set of classes) could have led to the observed input vector? For each of them, a second-order "confidence" judgment can also be performed by setting additional response classes, corresponding to whether the first-order decision has a high or a low probability of being correct. As shown graphically in Figure 1.c, there is a distinct confidence judgment associated with each primary task. At expense with Persaud *et al.* (PERSAUD ET AL., 2007) and Lau *et al.* (LAU, 2008), we note that second-order tasks need not coincide with visibility judgment. Also note that, for both first- and second-order decisions, the decision boundaries can be derived directly from the definition of the task, the priors, and the likelihood functions for each class, and therefore do not constitute additional assumptions of the model.

EMPIRICAL CONSEQUENCES OF THE DECISION FRAMEWORK

We shall now see how this framework accounts for the six fundamental empirical properties listed above.

PROOFS

1. ABOVE-CHANCE DISCRIMINATION OF STIMULI REPORTED AS "UNSEEN"

Empirical finding 1 is that perceptual decisions can be performed at above-chance level even when subjects report not seeing any stimulus (MARSHALL AND HALLIGAN, 1993; COWEY AND STOERIG, 1995; DRIVER ET AL., 2001; PERSAUD ET AL., 2007; TAMIETTO ET AL., 2007). For example, blindsight patients can perform simple discriminations on visual stimuli they report not seeing (STOERIG AND COWEY, 2009). This paradoxical ability also exists in healthy subjects whose discrimination performances have been repeatedly shown to be dissociated from subjective reports (see review in (KOUIDER AND DEHAENE, 2007; OVERGAARD AND SANDBERG, 2012)).

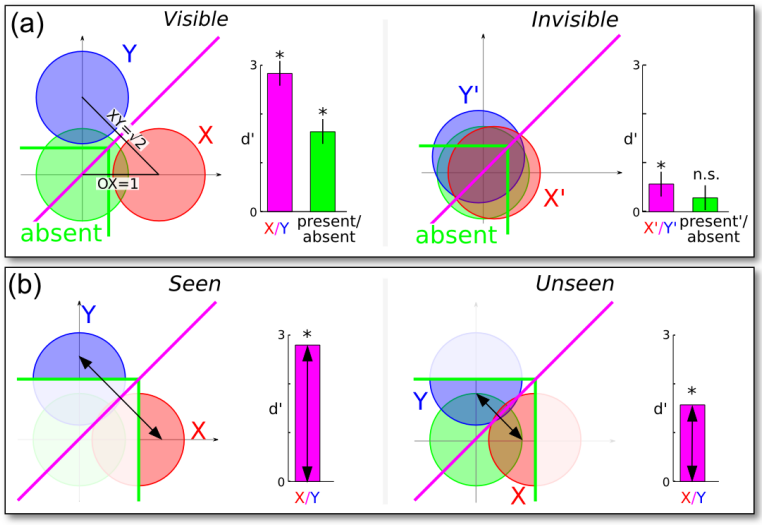

FIGURE 2. AN ACCOUNT OF UNCONSCIOUS AND CONSCIOUS DISCRIMINATION PERFORMANCE IN TWO TYPES OF EXPERIMENTAL DESIGNS.

A. In the stimulus degradation design, stimuli are made invisible by reducing the evidence (e.g. lowered contrast, masking, inattention, etc). This manipulation makes the stimuli more similar to the Absent class (right). As the XY distance can be longer than the distances separating the Absent class and the stimuli classes (OX, OY), discrimination performance (purple) can remain significant while detection sensitivity (green) is not detectable better than chance.

B. In the fixed-stimulus design, near-threshold stimuli are sorted as a function of whether they are reported as "seen" or "unseen". Unseen stimuli can be discriminated at above-chance levels, but discrimination performance improves drastically on seen trials.

For simplicity, we only consider here the case in which two stimuli (X and Y) become undetectable when they are visually degraded (X' and Y'). We assume that the degraded stimuli are generated from the same class as X and Y, yet with lower evidence (*i.e.* shorter vector length). As shown in Figure 2.a, it is quite possible for degraded stimuli X' and Y' to fall in the region reported as "unseen" during visibility judgment (*i.e.* the most likely class is Absent), and yet to yield above-chance performance in a forced-choice task when discrimination is restricted to classes X and Y. This finding could be trivial if the visibility judgment was systematically biased towards the "unseen" response (and indeed such response bias has often been proposed as an interpretation of subliminal perception experiments (HOLENDER, 1986)). However, our simulations assume a Bayes-optimal inference process. Thus, we show that there are conditions under which the Absent or "unseen" response is the most probable one, and yet X *versus* Y can still be discriminated.

The geometry of the 2D model reveals why discrimination performance (*i.e. d'* of X/Y discrimination) can be higher than detection sensitivity (*i.e. d'* of Absent/not-Absent judgment): the distance separating the X and Y vectors is larger than the distance separating them from the Absent class. In the two-dimensional case, discrimination performance is √2 higher than detection performance (Figure 2.a). Consequently, given adequate statistical power, discrimination may be significantly above chance when detection sensitivity is not.

The above account can also be extended to second-order judgments such as confidence rating and post-decision wagering on the first-order forced-choice X/Y discrimination task. Because such second-order judgments rely on similar decisional principles as the first-order tasks (Figure 1.c), confidence in discrimination can be above-chance on "unseen" trials, and confidence in visibility can be lower than confidence in discrimination. This conclusion fits with two recent experiments in which subjects performed above-chance in their confidence judgments, even on trials reported as unseen (KANAI ET AL., 2010; CHARLES ET AL., 2013).

2. DISCRIMINATION PERFORMANCE GENERALLY IMPROVES WITH SUBJECTIVE VISIBILITY

Empirical finding 2 is that, although objective discrimination can be above chance with subjectively invisible stimuli, such unconscious performance is generally mediocre. In many studies, objective discrimination performance improves dramatically when the stimuli are reported as "seen" compared to "unseen", even when sensory stimulation is identical (SERGENT AND DEHAENE, 2004; DEL CUL ET AL., 2007; NEUROSCIENCE ET AL., 2012).

How does the model account for these findings? In experiments that compare high-contrast visible stimuli with degraded invisible stimuli, the improvement in discrimination performance with subjective visibility is trivial (Figure 2.a): stimulus degradation diminishes the evidence for X and Y, and thus worsens both visibility judgment and X/Y discrimination. The two tasks are thus necessarily correlated (LAU AND PASSINGHAM, 2006; LAU, 2008). Less trivially, however, the model predicts the same effect for fixed stimuli presented at perceptual threshold. Even when the stimuli are physically identical, internal variability can explain why ~50% of them are reported as "unseen" (those which are most similar to the Absent class). As a consequence of this variability, sensory inputs reported as "unseen" are associated with a shorter input vector and are therefore closer to the X/Y discrimination border than samples reported as "seen" (Figure 2.b). The simple hypothesis of a noisy input vector, together with non-orthogonal discrimination and detection tasks, suffices to explain why unseen trials generally exhibit a lower discrimination performance than seen trials.

3. DISCRIMINATION PERFORMANCE CAN BE EQUATED ON "SEEN" AND "UNSEEN" TRIALS

Empirical finding 3 is that it is possible to find experimental conditions in which discrimination performance is equated while visibility varies. For instance, blindsight patients do not always show different discrimination performance in their blind and healthy visual fields (WEISKRANTZ, 1986; COWEY, 2010; KO AND LAU, 2012). In healthy subjects, using metacontrast masking and

PROOFS

inattention, stimuli have been created that differ in visibility, but are equated for objective discrimination performance (LAU AND PASSINGHAM, 2006; RAHNEV ET AL., 2011, 2012).

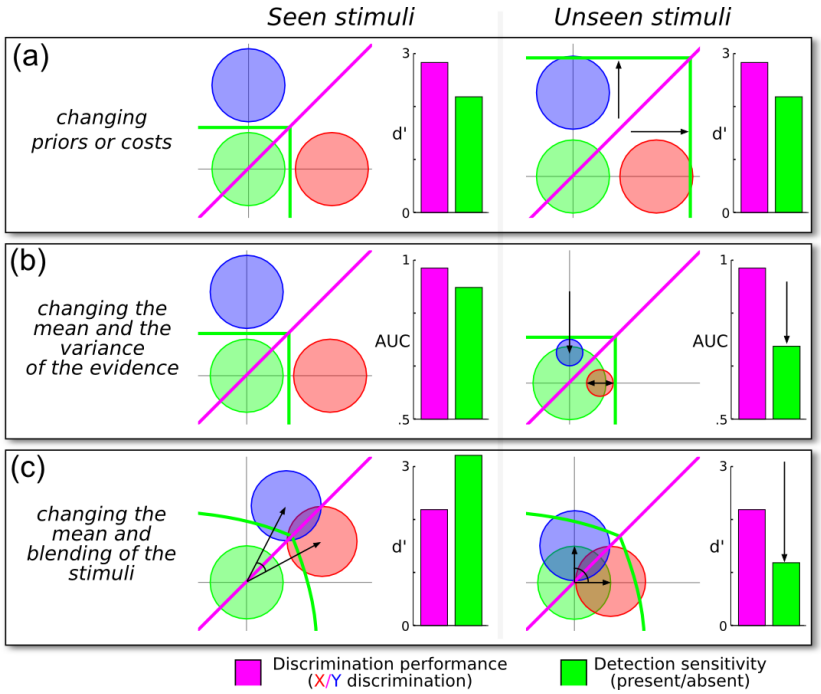

FIGURE 3. THREE WAYS IN WHICH STIMULUS VISIBILITY CAN BE MANIPULATED INDEPENDENTLY OF STIMULUS DISCRIMINABILITY.

A. Changing the prior (or the cost) of the Absent class affects the placement of the criterion for subjective visibility reports and can thus lead to a systematic report of invisibility, without affecting objective discrimination performance (purple) or detection sensitivity (green).

B. Simultaneously changing the length and the variance of the input vectors jointly affects detection sensitivity and subjective visibility reports while preserving objective discrimination performance. (The Area Under the Curve (AUC) is an equivalent of d' for continuous measures.)

C. Simultaneously changing the length (e.g. contrast) and the angle (e.g. ambiguity) of the input vectors can lead to a similar pattern of results.

In the model, three major circumstances (and mixtures of them) may lead to identical discrimination performance for seen and unseen stimuli.

First, for fixed stimuli X and Y, an increase in the prior probability (or cost) of the Absent class may lead to an increase in "unseen" responses while leaving X/Y discrimination unaffected (Figure 3.a). This account formalizes the hypothesis that blindsight patients have an inappropriate "criterion" for visibility judgment (*e.g.* (LAU, 2008; COWEY, 2010; KO AND LAU, 2012)). Note however that the concept of criterion can be misleading because it incorrectly suggests a single scalar value. In the present framework, the "criterion" emerges as a set of decisional boundaries that delimit the categorical regions in the representational space, and that are specific to the selected task. A change in the task or in the priors may thus impose a different division of space, and hence a shift in decision boundaries.

Second, consider experiments in which, within each class, the experimenter presents two visible targets X and Y, and two invisible targets X' and Y'. If both the length and the variance of the input vectors X' and Y' are reduced compared to X and Y, their visibility can drop without

affecting discrimination performance (Figure 3.b). This case could correspond to a simultaneous manipulation of stimulus strength (length of input vector) and of attention (variance of the input vector) as proposed by Rahnev and collaborators (RAHNEV ET AL., 2011).

Third, if both the amplitude and the angle of the input vectors X' and Y' are decreased compared to X and Y, then X/Y discrimination performance could be manipulated independently of visibility (Figure 3.c). This case could correspond to a simultaneous change in contrast and in stimulus ambiguity, for instance using morphing or blending to reduce the difference between X and Y stimuli.

The present account provides no less than three mechanisms by which blindsight, metacontrast and inattention could produce their effects. Each mechanism could be explicitly tested by experimentally manipulating the contrast, the variance and/or the blending of sensory stimuli as well as the prior probability associated with each class.

4. SUBJECTIVE REPORTS ARE OFTEN NON-LINEARLY RELATED TO SENSORY STRENGTH

Empirical finding 4 is that a non-linear curve often relates the strength of sensory stimulation and visibility ratings (SERGENT AND DEHAENE, 2004; DEL CUL ET AL., 2007; MELLONI ET AL., 2011). For example, when the stimulus onset asynchrony (SOA) separating a briefly flashed digit and its subsequent mask is varied linearly, a sharp transition in visibility occurs around an SOA of 50 ms: below this duration, subjects tend to report the stimulus as completely unseen, whereas above it, stimuli are reported as clearly visible (SERGENT AND DEHAENE, 2004; DEL CUL ET AL., 2007). However, this all-or-none visibility pattern does not characterize all types of subjective reports (SERGENT AND DEHAENE, 2004; SANDBERG ET AL., 2011; OVERGAARD AND SANDBERG, 2012; WINDEY ET AL., 2013). For example, Sergent and Dehaene (SERGENT AND DEHAENE, 2004) showed that the attentional blink leads to a much sharper non-linear pattern than backward masking.

We consider two classes X and Y, within which the stimuli can vary parametrically in strength from trial to trial (Figure 4.a). This parametric variation is assumed to have a linear effect on the amount of sensory evidence in favour of the corresponding stimulus (*i.e.* the length of the input vector). In such cases, the model predicts that visibility responses are non-linearly related to stimulus evidence, as the MAP criterion imposes a decision boundary that sharply delineates the regions of space respectively responded with the "seen" and "unseen" labels. Interestingly, although the fraction of "seen" responses is always a sigmoid, its slope may vary from a step-wise "all-or-none" pattern to a shallow and near-linear function. The parameter driving this change in sigmoid slope is the variance in representational space. With higher variance, visibility becomes more linearly related to sensory evidence (Figure 4.a right). This is because when variance increases, a greater number of Absent samples fall outside of the region responded classified as Absent, and, analogously, a greater number of present trials (X or Y) fall outside their respective regions – ultimately leading to a flat relationship between stimulus evidence and discrimination performance. This change is also accompanied by an increased proportion of unseen responses. Contrarily, the sigmoid becomes sharper and the number of seen responses increases when the variance of the stimulus diminishes (Figure 4.a left).

PROOFS

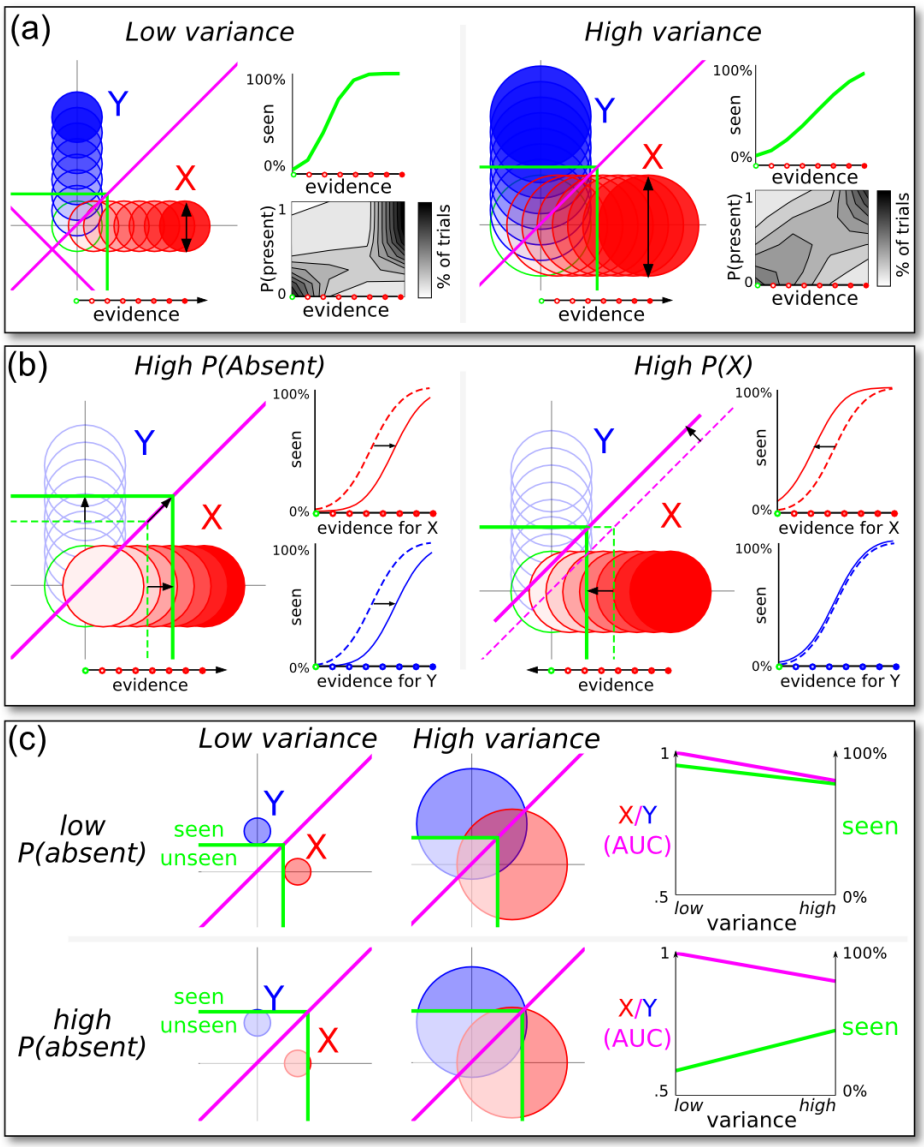

FIGURE 4. INPUT VARIANCE AND PRIOR KNOWLEDGE CAN AFFECT THE NON-LINEARITY AND THE THRESHOLD OF SUBJECTIVE VISIBILITY REPORTS.

(a) Parametrically varying stimulus strength directly changes the amplitude of the input vector and leads to a non-linear pattern of subjective visibility reports. The slope and the intercept of the resulting sigmoid depend on stimulus variance: low variance leads to an all-or-none relationship between the evidence and the visibility reports (left), whereas high variance leads to a more linear relationship as well as an increase in the visibility threshold (right).

(b) Prior knowledge can also affect the visibility threshold. Increasing the prior probability of the Absent class increases the visibility threshold for all stimuli, thus lowering subjective visibility reports. When only the prior probability of X is increased (capturing "hysteresis" experiments where subjects come to expect the next stimulus), then the visibility threshold is lowered for X alone, while the visibility threshold for Y barely changes.

(c) Visibility and discrimination interact when both priors and stimuli variance are varied. If the probability of the Absent class is relatively low (or similarly if the evidence is relatively high), increasing the variance reduces both visibility ratings and discrimination performance. However, when P(Absent) is high (or similarly, if the evidence is low), increasing the variance can diminish discrimination performance while increasing visibility ratings. This diagram captures the paradoxical finding that increased attention can lead to reduced visibility (RAHNEV ET AL., 2011).

The present model thus shows how both near-linear and non-linear visibility patterns can be produced by a single type of decision. The model also predicts that unseen trials should tend to be characterized by linear patterns, and seen trials with all-or-none patterns – an empirically verified phenomenon (SERGENT AND DEHAENE, 2004; DEL CUL ET AL., 2007; DE GARDELLE ET AL., 2011; DE LANGE ET AL., 2011; MELLONI ET AL., 2011). Because there is no unequivocal way of determining the internal variance of sensory inputs in existing experiments, the present account remains speculative. Nevertheless, stimulus variance could be explicitly manipulated in future experiments.

5. PRIOR KNOWLEDGE CAN LOWER THE VISIBILITY THRESHOLD

Empirical finding 5 is that the subjective visibility threshold is affected by prior knowledge (MOONEY, 1957; RODRIGUEZ ET AL., 1999; SIMONS AND CHABRIS, 1999; TAILLON-BAUDRY AND BERTRAND, 1999; CUL ET AL., 2006; MELLONI ET AL., 2011). Prior exposure to a given word increases its objective identification and subjective visibility when the same word is later presented under stronger masking (CUL ET AL., 2006). Similarly, Melloni *et al.* (MELLONI ET AL., 2011) recently used a hysteresis paradigm in which letters were embedded in white noise. Across a series of trials, the identity of the letter was fixed while its signal-to-noise gradually increased and then gradually decreased. Subjects reported seeing the letter better in the descending than in the ascending condition (*i.e.* once they knew the identity of the letter), even for identical physical stimulation.

In the present model, these effects arise from changes in the priors for classes X and Y. At the beginning of the ascending condition, stimulus evidence is low, and the X and Y classes are equally likely. Once the stimulus has been identified, at the beginning of the descending condition, its prior probability P(X) is increased, and consequently P(Absent) and P(Y) are decreased. Because the decision boundary for the "seen" response is partly determined by P(X), the "seen" response is more likely in the descending sequence than in the ascending one (Figure 4.b).

Although this account captures the influence of prior knowledge on visibility reports (CUL ET AL., 2006), it oversimplifies the hysteresis paradigm (MELLONI ET AL., 2011). Indeed, subjects are also likely to learn the structure of the ascending and of the descending sequences, and expect a higher frequency of absent trials towards the beginning of the ascending sequence and towards the end of the descending sequence. This expectation, if present, would again increase the prior probability of the "unseen" response, thus leading to increased reports of invisibility for these stimuli compared to physically identical stimuli presented in a random sequence. The model further predicts that X/Y discrimination should remain identical in ascending and descending sequences. During the descending sequence, subjects should exhibit a bias towards X reports, due to the increased prior for X, but no change in *d'*. These predictions offer a way to test the validity of the present model.

6. ATTENTION CAN EITHER INCREASE OR DECREASE VISIBILITY

Empirical finding 6 is that attention and visibility can be paradoxically decorrelated. In many studies, attention increases detection sensitivity and subjective visibility (*e.g.* (KIM AND BLAKE, 2005; RAHNEV ET AL., 2011; SERGENT ET AL., 2013)). However, attention can also lead to *decreased* subjective visibility (RAHNEV ET AL., 2011). In Rahnev *et al.*'s study (RAHNEV ET AL., 2011), subjects performed a basic detection task on a target whose location was validly cued on 70% of trials. Crucially, the

PROOFS

contrast of the unattended target was adjusted to yield the same level of objective performance as the attended target. Remarkably, subjects reported that unattended trials were more visible than attended ones.

If we assume that attention affects the variance of the input vector, the present model predicts that attention can lead to opposite visibility effects depending on the proportion of trials reported as seen or unseen (Figure 4.c). If P(absent) is low, so that most trials are reported as seen, then increasing the variance diminishes both discrimination performance and visibility, because it increases the proportion of input vectors that fall close to the Absent class. This captures the classical effect that inattention increases noise and thus reduces both objective performance and subjective visibility. Importantly, however, if P(absent) is high, so that most trials are reported as unseen, then increasing the stimulus variance still diminishes discrimination performance, but may paradoxically *increase* visibility ratings. This is because with higher variance, a greater number of samples fall outside of the region responded as "unseen" and thus become subjectively visible (see Figure 4.c).

The model therefore predicts that attention can induce opposite effects on visibility and discrimination performance even when the mean evidence is unchanged. Contrarily to Rahnev *et al.* (RAHNEV ET AL., 2011), who argue that attention induces a conservative visibility bias by changing the inter-trial variance of the stimulus, we predict that visibility ratings are influenced by an interaction between the variance and the initial visibility threshold (determined by prior knowledge or stimulus evidence). Once again, this prediction could be tested in an experiment explicitly manipulating stimulus variance, contrast and priors.

PROOFS

EXPERIMENTAL TEST OF THE MODEL

Most the above arguments account for empirical observations only in retrospect. We thus opted to confront the present model to a novel experimental setup. The model critically predicts that linear and non-linear profiles of behavioural responses arise from the *same decision mechanism*. In particular, it predicts that the discrimination profile of *physically identical* stimuli will increasingly become non-linear as visibility increases (Figures 2.b & 4.a).

We tested this prediction by linearly varying a parameter λ to create a continuum between two perceptual classes X and Y (Figure 5). For λ=0, the stimulus is X, for λ=1, the stimulus is Y, but we can create an arbitrary series of intermediate stimuli S(λ)= λX+ (1- λ)Y. Whereas de Gardelle *et al.* (DE GARDELLE ET AL., 2011) used a linear morph between two faces, here we varied the contrast of a single line to create a continuum between two different digits (*e.g.* 55555555). Geometrically, such a continuum can be represented as a line joining the prototypical vectors of each class (Figures 6.a left). We presented the stimuli at perceptual threshold, such that for a fixed stimulus, there was a large number of both "seen" and "unseen" subjective reports.

The model predicts that the steepness of discrimination performance should increase as subjective visibility increases. Stimuli rated as "unseen" could be categorized better than chance (Figure 2.b), but with a shallow slope because such stimuli are necessarily close to the "Absent"

class (Figure 6.a right). Conversely, highly visible stimuli should yield a steeper sigmoidal function. Thus, we expected significantly better identification performance on "seen" compared to "unseen" trials (Figure 2.b), and an increasingly "all-or-none" response pattern as a function of stimulus ambiguity λ (Figure 6.a right).

METHOD

Nineteen healthy volunteers, with normal or corrected-to-normal vision, participated after giving informed consent (29% males, Age: 25 ± 5 years old, 88% right handed). Each trial began with the presentation of an ambiguous digit (target) presented for 83 ms and subsequently masked by pseudo-random black surrounding letters displayed for 67 ms (Figure 5). Subjects were asked to identify in less than 2 s which of four digits was presented (5, 6, 8 or 9), using their left and right index and middle fingers. Visual feedback was given for non-ambiguous trials (morphs at 0% or 100%): misidentifications were followed by a 100 ms red fixation cross, whereas correct identifications were followed by 100 ms green fixation cross. Subjects subsequently reported subjective visibility using a 10-point vertical rating scale (bottom: not seen, top: clearly visible). Subjects used the two middle fingers to change the location of the randomly-placed visibility cursor, and pressed the space bar with their thumb to validate the visibility rating. The inter-trial interval was fixed at 300 ms. Subjects performed a total of 1000 trials divided into 25 blocks, at the end of which their median reaction times and their accuracy were displayed. The experiment lasted approximately one hour.

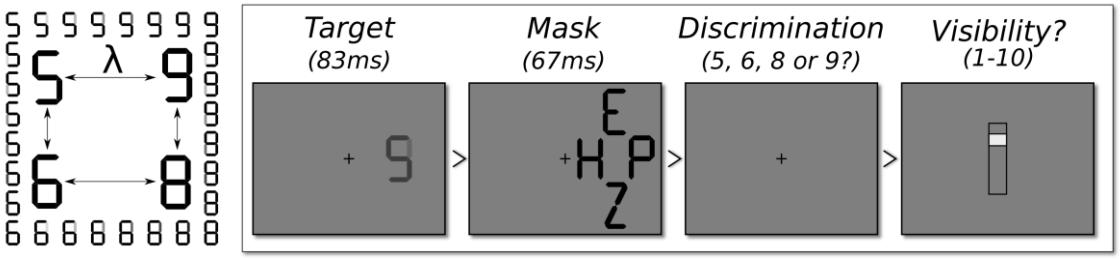

FIGURE 5. EXPERIMENTAL DESIGN

To test whether linear and non-linear subjective reports could be accounted by a single type of decision, we parametrically varied the evidence (λ) favouring four different stimuli (5, 6, 8, 9) by creating morphs between pairs of these digits (left). For each morph, on each trial, subjects performed a forced-choice identification task and provided a subjective visibility report (right).

Prior to the main experiment, subject performed a staircase procedure similar to the main task (100 trials with unambiguous targets, no visibility ratings, and no time limit). The contrast was lowered to reach an accuracy of ~70% (KAERNBACH, 1991). Target contrast then remained fixed throughout the main experiment. The staircase procedure was repeated up to five times in case of an unstable perceptual threshold. Two subjects who failed to converge to a stable threshold were excluded.

All stimuli were generated on a computer using INKSCAPE, MATLAB 2009b and the Psychophysics Toolbox and were displayed on a 17" computer CRT screen (1600 x 900 refreshed at 60 Hz). The screen background colour was 50% gray throughout the whole experiment and a black fixation-cross was constantly presented in the middle of the screen. Targets were morphs

between two digits (5-6, 5-9, 6-8, 9-8) each made of 5 to 7 black bars (Figure 5). In each pair, a single bar varied between gray (background colour) to maximal contrast in eight linear steps (parameter λ varying from 0 to 1 in steps of 0.143). Masks were composed of four pseudo-random capital letters constructed from the same basic visual features as the digits and were located at the top (E, O, U, Z), at the bottom (A, F, P Z), to the left (A, H, O, U) and to the right (E, F, P H) of the target digit. Symbols subtended 0.45° x 0.85° and were presented to the left or to the right side of the fixation (2.12°). Masks were centred on the previously presented target (1.23° x 2.27°). Targets, masks and their respective location were randomly selected at each trial. On 15% of trials, the target was absent and replaced by a gray background.

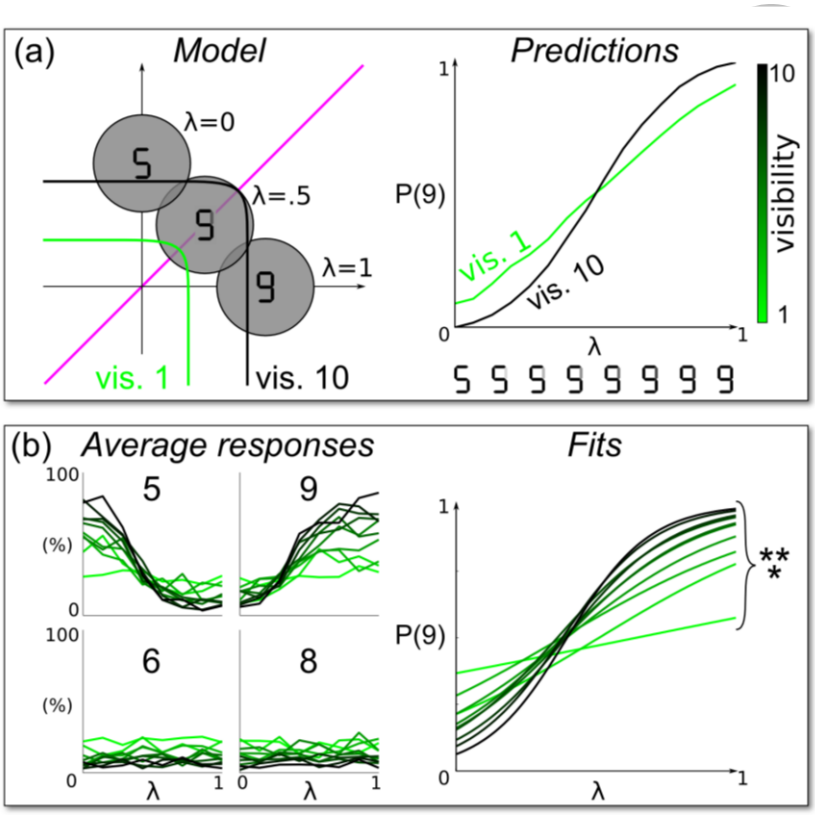

FIGURE 6. EMPIRICAL TEST OF THE PREDICTED VARIATION IN NON-LINEAR CATEGORIZATION AS A FUNCTION OF VISIBILITY.

The present framework predicts that the steepness of the sigmoid characterizing discrimination performance as a function of λ should increase with visibility reports. In particular, the discrimination performance of unseen stimuli should follow a quasi linear trend.

The results (n=17) confirm that i) stimuli could be identified above chance even at the lowest visibility ratings ii) discrimination performance correlated with visibility ratings, and iii) increasingly steeper sigmoids indicated that, unlike unseen stimuli, visible stimuli were associated with a nearly all-or-none identification performance.

RESULTS

Unambiguous targets were accurately identified on 67.7% of trials (SD = 14.1%, $t(16) = 5.00$, $p < 0.001$) confirming that the staircase procedure was efficient (targeted accuracy: 70%). Subjects used the visibility scale appropriately, as indicated by their more frequent use of the 0% visibility response on target-absent trials than on target-present trials (36.7% *vs.* 16.9% of trials,

$t(16) = 4.867$, $p = < 0.001$). Subjects used the entire visibility scale on target-present trials, from 0% visibility (16.9% of trials) up to 100% visibility (18.7% of trials).

We sorted trials as a function of reported visibility (10 levels), and within each level, examined how identification responses varied as a function of bar contrast (parameter λ). We only focused on the two adequate responses to a given morph (*e.g.* responses 5 or 9 for the 5-9 morph), and computed the fraction of these responses that corresponded to reporting the presence of a bar. We used R software to fit a binomial distribution as a function of bar intensity, separately for each subject and each visibility level. As seen in Figure 6.b, subjects' choices varied significantly as a function of bar contrast at all visibility ratings (all $p < 0.001$). Thus, subjects discriminated digits at above-chance level even on trials when they reported no subjective perception. Furthermore, as predicted, the slope of the sigmoid function increased significantly with visibility ratings (r^2 (15) = 0.79, $p = 0.004$). Thus, discrimination performance improved with subjective visibility ratings. Trials rated as invisible had such a shallow slope that the response proportion were nearly linearly related to the intensity of the bar, while trials rated as highly visible resulted in a nearly stepwise, "all-or-none" response function.

DISCUSSION OF THE EXPERIMENT

Although subjects were presented with identical stimuli, subjective reports varied considerably from trial to trial, from total invisibility to maximal visibility. Furthermore, three predictions were verified: (1) identification scores were always higher than chance level; (2) they increased with visibility; (3) when varying the degree of ambiguity λ, objective identification became increasingly non-linear as subjective visibility increased. These results confirm that, for physically identical stimuli, visibility is associated with a greater degree of "all-or-none" perception, a finding that the framework can explain without any additional assumption (*i.e.* no need to postulate a qualitative difference between conscious and unconscious processing).

Our results extend a previous study by de Gardelle *et al.* (DE GARDELLE ET AL., 2011), which examined the amount of masked repetition priming elicited by a morphed face when the prime was unmasked (SOA = 300 ms) or heavily masked stimuli (SOA = 43 ms). As in the present experiment, they observed linearly increasing priming for invisible morphs, and categorical priming for visible morphs. Although the authors proposed that this dissociation reflected two distinct processes (unconscious analogue versus conscious discrete), the present model suggests that this interpretation is unnecessary: even within a single decision process, response patterns may vary in their degree of non-linearity depending on the mean and variance of the stimulus evidence.

The model further predicts that, when conscious perception occurs, subjects perceive the stimuli strictly categorically (digit 5 or 9, but no intermediate percept). According to Harnad's definition (HARNAD, 2003), categorical perception is defined by "within-category compression and between-category separation". In another paper (King *et al.*, *in preparation*), we present additional evidence that the conscious experience of our morphs follows Harnad's definition of categorical perception (HARNAD, 2003). First, discriminability is indeed enhanced for pairs of digits presented near the perceptual boundary. Second, when presented with two identical ambiguous morphs, subjects frequently judge that the stimuli differ, as predicted if each has a ~50% chance of falling in either of two discrete perceptual categories. Third, when the present identification task is repli-

PROOFS

cated using a continuous response scale, subjects respond bimodally and barely use the intermediate levels to report perceiving a mixture of two digits. Thus, at least for this type of stimuli, and as postulated in our theoretical premises, what we consciously perceive seems to result from a categorical decision among a limited number of classes (see also (Moreno-Bote et al., 2011; Gershman et al., 2012)).

General discussion

We have shown how a simple geometrical framework for subjective report and objective discrimination tasks, based on Signal Detection and Bayesian Theories, can account for six fundamental findings in behavioural studies of conscious and unconscious perception. The present model subsumes a series of frameworks describing both conscious and unconscious perception as statistical inferences (Kersten and Yuille, 2003; Kersten et al., 2004; Knill and Pouget, 2004; Doya, 2007; Knill and Richards, 2008; Lau, 2008; Shadlen et al., 2008). The core of our hypothesis is that, during perception, the brain is faced with a massive classification problem. Each task, including conscious identification and subjective report, imposes, in a top-down manner, a set of classes along which the stimuli can be classified. Contrarily to most laboratory tasks, open-ended subjective reports are typically based on numerous features and classes. A picture naming task, for instance, typically involves tens of thousands of classes. Like others before us (Green and Swets, 1966; Klein, 1985; Ko and Lau, 2012), we thus insist on the necessity to conceptualize decisions within a multidimensional framework. This conceptualization leads to several important methodological and theoretical consequences.

Firstly, the present model goes against the idea that subjective reports of "not seeing" are necessarily unreliable because they can be affected by conservative response biases (Eriksen, 1960; Green and Swets, 1966; Merikle, 1982; Holender, 1986), and that objective measures such as detection sensitivity (*d'*) should be favoured (see review in (Kouider and Dehaene, 2007)). On the contrary, we show that subjective reports cannot be reduced to objective measures (Eriksen, 1960; Merikle, 1982; Holender, 1986) nor to second-order measures such as confidence rating and post-decision wagering (Lau and Passingham, 2006; Persaud et al., 2007; Lau, 2008; Rahnev et al., 2011). In particular, the present model predicts that visibility and confidence should be partially correlated (Figure 1.c) but experimentally dissociable. This prediction is well supported by recent empirical findings showing that second-order judgments can be performed above chance on unseen stimuli (Kanai et al., 2010; Sandberg et al., 2011; Overgaard and Sandberg, 2012; Charles et al., 2013). In the present model, subjective visibility reports reflect a legitimate decision process whose details (including response bias) can and should be accounted for. As recently demonstrated (Cul et al., 2006; Schwiedrzik et al., 2009; Melloni et al., 2011), a shift in visibility criterion reflects the underlying prior probabilities and cost functions of the subjects' internal model of the world, and, consequently, should not be disregarded as an experimental confound. What we call a "subjective" report may simply be the brain's best attempt at solving a difficult perceptual decision problem with myriad of potential classes, each with different costs and prior probabilities that depend on the subject's prior experience.

PROOFS

Secondly, the model shows, in a principled manner, how experimental conditions can be designed to equate discrimination performance between seen and unseen trials (Figure 3). In a series of behavioural experiments, Lau and collaborators have equated objective discrimination performance between seen and unseen responses, in an attempt to isolate conscious processing independently of other pre- or post-perceptual increases in information processing (Lau and Passingham, 2006; Lau, 2008; Rahnev et al., 2011). The present geometrical analysis suggests that Lau's experiments have adopted only a subset of the possible solutions: masking the stimuli at different levels (Lau and Passingham, 2006) or changing the amount of attention they receive (Rahnev et al., 2011) may both change the signal-to-noise ratio of the incoming evidence. However, under such conditions, discrimination performance is equated at the expense of introducing physical differences between the visible and invisible stimuli. It is therefore unclear whether contrasting the two reflects an effect of visibility or of the stimulus' physical properties. Consequently, it may be preferable to use physically identical stimuli and alter subjective visibility by changing the priors (Figure 3.a) – a solution indeed adopted in several recent studies (Cul et al., 2006; Schwiedrzik et al., 2009; Melloni et al., 2011).

The empirical finding of a non-linear sigmoidal relationship between subjective visibility reports and the physical properties of a stimulus (Vorberg et al., 2003; Dehaene et al., 2006; Deco et al., 2007; Kouider and Dehaene, 2007; Quiroga et al., 2008; De Gardelle et al., 2011; de Lange et al., 2011) has led to the notion that conscious perception is an all-or-none phenomenon (Sergent and Dehaene, 2004; Del Cul et al., 2007; Melloni et al., 2011). The present model readily reproduces this non-linear pattern (Figure 4.a), but it also predicts exceptions in cases of high stimulus variance or low signal-to-noise ratio. These predictions remain untested, but may offer potential explanations to studies revealing a continuous relationship between stimulus evidence and subjective reports (Sergent and Dehaene, 2004; Sandberg et al., 2011; Overgaard and Sandberg, 2012; Windey et al., 2013). In the future, directly manipulating the mean and the variance of stimulus evidence could clarify the role of each of these factors in linear and non-linear response patterns to sensory manipulations.

According to the present model, the reason why unconscious responses tend to be linearly related to stimulus evidence is simple: when perceptual evidence is low enough to be categorized as "unseen", the evidence necessarily lies close to the origin of the multidimensional space and therefore leads to shallow (though above-chance) forced-choice curves. We tested this idea in an original experiment, and the results confirmed that fixed stimuli presented at threshold lead to quasi-linear discrimination when reported as unseen, but to a sharp sigmoid discrimination curve when reported as seen. Contrarily to previous proposals (Del Cul et al., 2007; De Gardelle et al., 2011; Charles et al., 2013) the present model accounts for these findings without having to postulate that distinct processes operate below and above the threshold for conscious perception.

Limits of the model and possible extensions

For simplicity we postulated that the very same representational vector is used for different tasks. The idea is that the same input vector is "resampled" several times with different response classes (*e.g.* a discrimination task followed by a visibility task on the same trial). This resampling assumption is supported by a recent experiment (Vul et al., 2009) in which, within a rapid stream of letters, subjects were asked to identify the one that was circled by a visual cue. On

PROOFS

each trial, subjects provided as many as four mutually exclusive guesses about the target letter. The results showed that all guesses were sampled from an identical distribution centred on the position and/or the time of the cue. This experiment suggests that the posterior probability of each letter was computed once and for all, and that successive guesses corresponded to the maximum a-posteriori (MAP) after excluding the previous answers, exactly as expected from the present model.

Nevertheless, in other contexts, the hypothesis that the input vector remains unchanged and identically available for a series of successive judgments may turn out to be simplistic. Temporal decay may affect the quality of decisions made after a delay (SPERLING, 1960), particularly for unconscious stimuli (GREENWALD ET AL., 1996; DUPOUX ET AL., 2008). A recent study suggests that an attentional cue presented *after* a sensory stimulus can retroactively improve its visibility (SERGENT ET AL., 2013). The task set imposed by a first task may also change the quality of the evidence available for a second task (JAZAYERI AND MOVSHON, 2007). Similarly, the order in which two questions are presented may influence the subject's answers (GILOVICH ET AL., 2002). Busemeyer *et al.* (BUSEMEYER ET AL., 2011) have proposed to account for the latter phenomenon with a computational principle inspired from quantum mechanics, according to which each successive judgment alters the input vector by projecting it onto a subspace defined by the task. As projections are not commutative, the order of successive questions can change the successive decisions. It remains to be seen whether such non-commutativity is a fundamental principle that should be added to the present model.

Another limit of the present model lies in its assumption, shared with SDT, that decisions are based on a single input vector. A natural extension of the model would represent a sensory input as a series of samples, *i.e.* a trajectory in multidimensional space. Indeed, SDT has been superseded by sequential sampling models (RATCLIFF AND ROUDER, 1998; GOLD AND SHADLEN, 2001; LEITE AND RATCLIFF, 2010), according to which each decision is based on an accumulation of noisy samples arising from the stimulus. Whichever accumulator first reaches a fixed threshold is selected as the winner of the perceptual decision. Models of this kind are supported by a large set of empirical findings, (RATCLIFF ET AL., 1999; KNILL AND RICHARDS, 2008; SHADLEN ET AL., 2008; DEHAENE, 2009; LIU AND PLESKAC, 2011) and account, not only for response proportions, but also for response times and their distributions (RATCLIFF ET AL., 1999; RATCLIFF AND MCKOON, 2008; LEITE AND RATCLIFF, 2010). Extending the present model in this direction, as attempted by Del Cul *et al.* (DEL CUL ET AL., 2007), would lead to precise predictions about subjects' reaction times in objective and subjective tasks.

In the tradition of "ideal observer" analyses, we also assumed that the decision system is fully informed of the stimulus distributions and uses optimal priors and likelihood functions to compute the posterior probability of each response class. This is undoubtedly an idealization. A dynamic model in which the likelihood functions, priors and costs would be learned by updating them after each trial, and may therefore be ill-estimated, may go a long way towards explaining a variety of human deviations from optimality. For instance, using a model similar to the present one, Ko and Lau (KO AND LAU, 2012) proposed an account of blindsight as an inadequate revision of priors following the radical decrease in visual input strength caused by a lesion to area V1 (similar to Figure 3.a). Confidence judgments and visibility ratings would be particularly affected by inadequate priors and likelihoods, because the present model assumes that these tasks require

PROOFS

a quantitative estimation of the posterior probabilities (Figure 1). In agreement with this idea, Rahnev, Lau and collaborators (Rahnev et al., 2011, 2012) performed a series of experiments in which human observers deviated radically from optimality in their confidence judgments. Their findings could be explained by assuming that subjects used a single estimate of input variance for distinct experimental conditions (*e.g.* for attended versus unattended trials). This interpretation is compatible with the present model, and with the general idea that there are sharp limits to the number of decision criteria that subjects may deploy on a given trial (Gorea and Sagi, 2000, 2010).

Neural mechanisms

The present model was framed at an abstract mathematical level of description. While this approach provides useful geometrical intuitions and a simple testable framework, an important future endeavour will be to flesh it out at the neural level. The vast representational space may correspond to the function of posterior unimodal and multimodal sensory areas, where many neurons render explicit dimensions of the stimuli that are only encoded implicitly and in a distributed form in the sensory periphery. Their role may be to augment the dimensionality of sensory inputs and therefore facilitate decision making by turning decisions into linearly separable problems (DiCarlo et al., 2012). The categorical decision system, in turn, could be subserved by areas of the dorsolateral and inferior prefrontal cortices as well as anterior temporal and superior parietal cortices. These areas have been proposed to form a "global workspace" where conscious information is maintained and broadcasted to additional processes (Dehaene and Changeux, 2011). They receive the necessary convergence of multimodal inputs and are known to contribute to both decision making and to all-or-non conscious perception (Freedman et al., 2002; Wood and Grafman, 2003; Dehaene and Changeux, 2011). Explicit simulations of such recurrent networks with winner-take-all dynamics show how they tend to quickly converge to a discrete stable attractor (Dehaene et al., 2003) which approximates the maximum-likelihood estimate (Deneve et al., 1999; Wang, 2008). The dynamics of such networks may therefore account for perceptual categorizations, which the present model considers as inherent to conscious perception.

PROOFS

Acknowledgements

This work was supported by a DGA grant to J.R.K., by INSERM, CEA, a European Research Council senior grant 'Neuro-Consc' to S.D., and the European Union Seventh Framework Programme (FP7/2007:2013) under grant agreement no. 604102 (Human Brain Project).We are grateful to Claire Sergent, Christelle Larzabal and Simon van Gaal for their help in the task, Patrick Cavanagh, Catherine Wacongne, Florent Meyniel & Victor Lamme for helpful comments, as well as to Lionel Naccache, Laurent Cohen, Laurence Labruna, Giovanna Santoro and Isabel Seror for their daily support. Finally, we thank our two anonymous reviewers for their constructive criticisms.

REFERENCES

BUSEMEYER JR, POTHOS EM, FRANCO R, TRUEBLOOD JS (2011) A quantum theoretical explanation for probability judgment errors. Psychological Review 118:193–218.

CHARLES L, VAN OPSTAL F, MARTI S, DEHAENE S (2013) Distinct brain mechanisms for conscious versus subliminal error detection. NeuroImage 73C:80–94.

COWEY A (2010) The blindsight saga. Experimental brain research Experimentelle Hirnforschung Expérimentation cérébrale 200:3–24.

COWEY A, STOERIG P (1995) Blindsight in monkeys. Nature 373:247–249.

CUL A DEL, NACCACHE L, VINCKIER F, COHEN L, DEHAENE S, GAILLARD RR, DEL CUL A (2006) Nonconscious semantic processing of emotional words modulates conscious access. Proceedings of the National Academy of Sciences of the United States of America 103:7524–7529.

DE GARDELLE V, CHARLES L, KOUIDER S (2011) Perceptual awareness and categorical representation of faces: Evidence from masked priming. Consciousness and Cognition 20:1–10.

DE LANGE FP, VAN GAAL S, LAMME V A F, DEHAENE S (2011) How awareness changes the relative weights of evidence during human decision-making. PLoS biology 9:e1001203.

DECO G, PÉREZ-SANAGUSTÍN M, DE LAFUENTE V, ROMO R, PE M (2007) Perceptual detection as a dynamical bistability phenomenon: A neurocomputational correlate of sensation. Proceedings of the National Academy of Sciences of the United States of America 104:20073–20077.

DEHAENE S (2009) Conscious and Nonconscious Processes: Distinct Forms of Evidence Accumulation? Engel C, Singer W, eds. Seminaire Poincare XII:89 – 114.

DEHAENE S, CHANGEUX J-P (2011) Experimental and theoretical approaches to conscious processing. Neuron 70:200–227.

DEHAENE S, CHANGEUX JP, NACCACHE L, SACKUR J, SERGENT C (2006) Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends in Cognitive Sciences 10:204–211.

DEHAENE S, NACCACHE L, LE CLEC'H G, KOECHLIN E, MUELLER M, DEHAENE-LAMBERTZ G, VAN DE MOORTELE PF, LE BIHAN D (1998) Imaging unconscious semantic priming. Nature 395:597–599.

DEHAENE S, SERGENT C, CHANGEUX J-P (2003) A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proceedings of the National Academy of Sciences of the United States of America 100:8520–8525.

DEL CUL A, BAILLET S, DEHAENE S, CUL A DEL (2007) Brain dynamics underlying the nonlinear threshold for access to consciousness. PLoS biology 5:e260.

DENEVE S, LATHAM PE, POUGET A (1999) Reading population codes: a neural implementation of ideal observers. Nature neuroscience 2:740–745.

DICARLO JJ, ZOCCOLAN D, RUST NC (2012) How does the brain solve visual object recognition? Neuron 73:415–434.

DOYA K (2007) Bayesian Brain: Probabilistic Approaches to Neural Coding (Doya K, Ishii S, Pouget A, Rao RPN, eds). MIT Press.

DRIVER J, VUILLEUMIER P, EIMER M, REES G (2001) Functional magnetic resonance imaging and evoked potential correlates of conscious and unconscious vision in parietal extinction patients. NeuroImage 14:S68–S75.

DUPOUX E, DE GARDELLE V, KOUIDER S (2008) Subliminal speech perception and auditory streaming. Cognition 109:267–273.

ERIKSEN CW (1960) Discrimination and learning without awareness: A methodological survey and evaluation. Psychological Review 67:279–300.

FREEDMAN DJ, RIESENHUBER M, POGGIO T, MILLER EK (2002) Visual categorization and the primate prefrontal cortex: neurophysiology and behavior. Journal of neurophysiology 88:929–941.

FRISTON K (2010) The free-energy principle: a unified brain theory? Nature reviews Neuroscience 11:127–138.

PROOFS

GERSHMAN SJ, VUL E, TENENBAUM JB (2012) Multistability and Perceptual Inference. Neural Computation 24:1–24.
GILOVICH T, GRIFFIN D, KAHNEMAN D (2002) Heuristics and Biases: The Psychology of Intuitive Judgment Gilovich T, Griffin D, Kahneman DTSB, eds. System 29:695.
GOLD JI, SHADLEN MN (2000) Representation of a perceptual decision in developing oculomotor commands. Nature 404:390–394.
GOLD JI, SHADLEN MN (2001) Neural computations that underlie decisions about sensory stimuli. Trends in Cognitive Sciences 5:10–16.
GOREA A, SAGI D (2000) Failure to handle more than one internal representation in visual detection tasks. Proceedings of the National Academy of Sciences of the United States of America 97:12380–12384.
GOREA A, SAGI D (2010) Using the unique criterion constraint to disentangle transducer nonlinearity from lack of noise constancy. Journal of Vision 1:437–437.
GREEN DM, SWETS JA (1966) Signal Detection Theory and psychophysics. Wiley.
GREENWALD AG, DRAINE SC, ABRAMS RL (1996) Three cognitive markers of unconscious semantic activation. Science 273:1699.
HARNAD S (2003) Categorical Perception. Wiley Interdisciplinary Reviews Cognitive Science 1:69–78.
HOLENDER D (1986) Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. Behavioral and Brain Sciences 9:1–66.
JAZAYERI M, MOVSHON JA (2007) A new perceptual illusion reveals mechanisms of sensory decoding. Nature 446:912–915.
KAERNBACH C (1991) Simple adaptive testing with the weighted up-down method. Perception And Psychophysics 49:227–229.
KANAI R, WALSH V, TSENG C-H (2010) Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness. Consciousness and cognition:2005–2009.
KERSTEN D, MAMASSIAN P, YUILLE A (2004) Object perception as Bayesian inference Knill D, Richards W, eds. Annual Review of Psychology 55:271–304.
KERSTEN D, YUILLE A (2003) Bayesian models of object perception. Current Opinion in Neurobiology 13:150–158.
KIM C, BLAKE R (2005) Psychophysical magic: rendering the visible "invisible". Trends in Cognitive Sciences 9:381–388.
KLEIN S A (1985) Double-judgment psychophysics: problems and solutions. Journal of the Optical Society of America A, Optics and image science 2:1560–1585.
KNILL DC, POUGET A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in Neurosciences 27:712–719.
KNILL DC, RICHARDS W (2008) Perception as Bayesian Inference. Cambridge Univ Pr.
KO Y, LAU H (2012) A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. Philosophical transactions of the Royal Society of London Series B, Biological sciences 367:1401–1411.
KOUIDER S, DEHAENE S (2007) Levels of processing during non-conscious perception: a critical review of visual masking. Philosophical transactions of the Royal Society of London Series B, Biological sciences 362:857–875.
LAU H (2008) A higher order Bayesian decision theory of consciousness. Progress in Brain Research 168:35–48.
LAU HC, PASSINGHAM RE (2006) Relative blindsight in normal observers and the neural correlate of visual consciousness. Proceedings of the National Academy of Sciences of the United States of America 103:18763–18768.
LEITE FP, RATCLIFF R (2010) Modeling reaction time and accuracy of multiple-alternative decisions. Attention, perception & psychophysics 72:246–273.
LIU T, PLESKAC TJ (2011) Neural correlates of evidence accumulation in a perceptual decision task. Journal of neurophysiology 106:2383–2398.

PROOFS

MARSHALL JC, HALLIGAN PW (1993) Visuo-spatial neglect: a new copying test to assess perceptual parsing. Journal of neurology 240:37–40.
MELLONI L, SCHWIEDRZIK CM, MÜLLER N, RODRIGUEZ E, SINGER W (2011) Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness. The Journal of neuroscience 31:1386–1396.
MERIKLE PM (1982) Unconscious perception revisited. Perception and Psychophysics 31:298–301.
MOONEY CM (1957) Age in the Development of Closure Ability in Children. Canad J Psychol 11:219–226.
MORENO-BOTE R, KNILL DC, POUGET A (2011) Bayesian sampling in visual perception. Proceedings of the National Academy of Sciences of the United States of America 108:1–6.
NEUROSCIENCE H, LIU Y, PARADIS A, YAHIA-CHERIF L, TALLON-BAUDRY C, STRANGE BA (2012) Activity in the lateral occipital cortex between 200 and 300 ms distinguishes between physically identical seen and unseen stimuli. Frontiers in human neuroscience 6:211.
OVERGAARD M, SANDBERG K (2012) Kinds of access: different methods for report reveal different kinds of metacognitive access. Philosophical transactions of the Royal Society of London Series B, Biological sciences 367:1287–1296.
PERSAUD N, MCLEOD P, COWEY A (2007) Post-decision wagering objectively measures awareness. Nat Neurosci 10:257–261.
PITTS M A, MARTÍNEZ A, HILLYARD S A (2012) Visual processing of contour patterns under conditions of inattentional blindness. Journal of cognitive neuroscience 24:287–303.
POUGET A, DENEVE S, DUHAMEL JR (2002) A computational perspective on the neural basis of multisensory spatial representations. Nature Reviews Neuroscience 3:741–747.
QUIROGA RQ, MUKAMEL R, ISHAM EA, MALACH R, FRIED I (2008) Human single-neuron responses at the threshold of conscious recognition. Proceedings of the National Academy of Sciences of the United States of America 105:3599–3604.
RAHNEV D A, BAHDO L, DE LANGE FP, LAU H (2012) Prestimulus hemodynamic activity in dorsal attention network is negatively associated with decision confidence in visual perception. Journal of neurophysiology 108:1529–1536.
RAHNEV D, MANISCALCO B, GRAVES T, HUANG E, DE LANGE FP, LAU H (2011) Attention induces conservative subjective biases in visual perception. Nature neuroscience 14:1513–1515.
RATCLIFF R, MCKOON G (2008) The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks.
RATCLIFF R, ROUDER JN (1998) Modeling Response Times for Two-Choice Decisions. Psychological Science 9:347–356.
RATCLIFF R, VAN ZANDT T, MCKOON G (1999) Connectionist and diffusion models of reaction time. Psychological Review 106:261–300.
RODRIGUEZ E, GEORGE N, LACHAUX JP, MARTINERIE J, RENAULT B, VARELA FJ (1999) Perception's shadow: long-distance synchronization of human brain activity. [London: Macmillan Journals], 1869-.
ROSENTHAL DM (1997) A theory of consciousness. The nature of consciousness:729–753.
SANDBERG K, BIBBY BM, TIMMERMANS B, CLEEREMANS A, OVERGAARD M (2011) Measuring consciousness: task accuracy and awareness as sigmoid functions of stimulus duration. Consciousness and cognition 20:1659–1675.
SCHWIEDRZIK CM, SINGER W, MELLONI L (2009) Sensitivity and perceptual awareness increase with practice in metacontrast masking. Journal of vision 9:18.1–18.
SERGENT C, BAILLET S, DEHAENE S (2005) Timing of the brain events underlying access to consciousness during the attentional blink. Nature Neuroscience 8:1391–1400.
SERGENT C, DEHAENE S (2004) Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. Psychological Science 15:720–728.
SERGENT C, WYART V, BABO-REBELO M, COHEN L, NACCACHE L, TALLON-BAUDRY C (2013) Cueing attention after the stimulus is gone can retrospectively trigger conscious perception. Current Biology 23:150–155.

PROOFS

SHADLEN MN, KIANI R, HANKS TD, CHURCHLAND AK (2008) Neurobiology of Decision Making An Intentional Framework. In: Better than Conscious?, pp 71–101.
SIMONS DJ, CHABRIS CF (1999) Gorillas in our midst: sustained inattentional blindness for dynamic events. Perception 28:1059–1074.
SPERLING G (1960) The information available in brief visual presentations. Psychological Monographs General and Applied 74:1–29.
STOERIG P, COWEY A (2009) Blindsight. The Oxford Companion to Consciousness:112–115.
TALLON-BAUDRY C, BERTRAND O (1999) Oscillatory gamma activity in humans and its role in object representation. Trends in Cognitive Sciences 3:151–162.
TAMIETTO M, GEMINIANI G, GENERO R, DE GELDER B (2007) Seeing fearful body language overcomes attentional deficits in patients with neglect. Journal of Cognitive Neuroscience 19:445–454.
TONONI G, EDELMAN GM (1998) Consciousness and Complexity. Science 282:1846–1851.
VORBERG D, MATTLER U, HEINECKE A, SCHMIDT T, SCHWARZBACH J (2003) Different time courses for visual perception and action priming. Proceedings of the National Academy of Sciences of the United States of America 100:6275–6280.
VUL E, HANUS D, KANWISHER N (2009) Attention as inference: selection is probabilistic; responses are all-or-none samples. Journal of experimental psychology General 138:546–560.
WANG X-J (2008) Decision making in recurrent neuronal circuits. Neuron 60:215–234.
WEISKRANTZ L (1986) Blindsight: A case study and implications. Oxford University Press, USA.
WINDEY B, GEVERS W, CLEEREMANS A (2013) Subjective visibility depends on level of processing. Cognition 129:404–409.
WOOD JN, GRAFMAN J (2003) Human prefrontal cortex: processing and representational perspectives. Nature reviews Neuroscience 4:139–147.
WYART V, DE GARDELLE V, SCHOLL J, SUMMERFIELD C (2012) Rhythmic Fluctuations in Evidence Accumulation during Decision Making in the Human Brain. Neuron 76:847–858.

PROOFS