Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation

Stephen M Fleming[1], Nathaniel D Daw[2]

[1]*Wellcome Trust Centre for Neuroimaging, University College London, London, UK*

[2]*Princeton Neuroscience Institute and Department of Psychology, Princeton University, New York, USA*

Number of pages: 59
Number of figures: 9
Number of tables: 1

Manuscript accepted for publication in *Psychological Review.*

Correspondence:

Stephen M. Fleming,
Wellcome Trust Centre for Neuroimaging
University College London
12 Queen Square
London
WC1N 3BG

E: stephen.fleming@ucl.ac.uk

Acknowledgements

We thank participants of the Fondation des Treilles workshop on "Subjective Confidence: Psychology, Physiology, Theory" (June, 2015) for helpful discussions. SMF was supported by a Sir Henry Wellcome Fellowship from the Wellcome Trust (WT096185). NDD was supported by a Scholar Award from the MacDonnell Foundation. Code supporting simulations found in this article can be downloaded from https://github.com/smfleming/Self-evaluation-paper

ABSTRACT

People are often aware of their mistakes, and report levels of confidence in their choices that correlate with objective performance. These metacognitive assessments of decision quality are important for the guidance of behaviour, particularly when external feedback is absent or sporadic. However a computational framework that accounts for both confidence and error detection is lacking. In addition, accounts of dissociations between performance and metacognition have often relied on ad hoc assumptions, precluding a unified account of intact and impaired self-evaluation. Here we present a general Bayesian framework in which self-evaluation is cast as a "second-order" inference on a coupled but distinct decision system, computationally equivalent to inferring the performance of another actor. Second-order computation may ensue whenever there is a separation between internal states supporting decisions and confidence estimates over space and/or time. We contrast second-order computation against simpler first-order models in which the same internal state supports both decisions and confidence estimates. Through simulations we show that second-order computation provides a unified account of different types of self-evaluation often considered in separate literatures, such as confidence and error detection, and generates novel predictions about the contribution of one's own actions to metacognitive judgments. In addition, the model provides insight into why subjects' metacognition may sometimes be better or worse than task performance. We suggest that second-order computation may underpin self-evaluative judgments across a range of domains.

Keywords: metacognition; decision-making; confidence; computation

INTRODUCTION

People are often aware of their mistakes, and report levels of confidence in their choices that correlate with objective performance. These assessments of decision quality are important for the guidance of behaviour, particularly when external feedback is absent or sporadic, and such metacognitive abilities are particularly well-developed in humans (Beran, Brandl, Perner, & Proust, 2012; Metcalfe, 1996; Norman & Shallice, 1986; Shea et al., 2014). Understanding the relationship between self-evaluations and performance is a key goal for multiple interlocking research areas including judgment and decision-making (Lichtenstein, Fischhoff, & Phillips, 1982), education (Veenman, Wilhelm, & Beishuizen, 2004), social psychology (Heatherton, 2011), consciousness science (Lau & Rosenthal, 2011) and clinical disorders (David, Bedford, Wiffen, & Gilleen, 2012; Goldstein et al., 2009). However, an appropriate computational framework that subsumes both confidence and error detection is lacking (Yeung & Summerfield, 2012). In addition, accounts of dissociations between performance and metacognition have often relied on ad hoc assumptions, precluding a unified account of intact and impaired metacognition.

In the lab, the mechanisms underpinning self-evaluation of performance have often been investigated by asking subjects to judge their confidence in simple decisions. As we will outline in further detail below, decision confidence can be defined as a subjective probability of a decision being correct (Aitchison et al., 2015; Pouget et al., 2016), and is one of many forms of uncertainty that the brain may encode (Meyniel et al., 2015a; Bach & Dolan, 2012). Decision confidence can be elicited through a variety of measures including self-reports, post-decision wagers and opt-out responses (see Kepecs & Mainen, 2012, for a review), and previous studies show that variability in decision confidence tracks changes in objective performance (Henmon, 1911; Nelson & Narens, 1990; Peirce & Jastrow, 1885) and supports the recognition of task errors (Gehring, Goss, Coles, Meyer, & Donchin, 1993; Rabbitt, 1966; Rabbitt & Rodgers, 1977; Yeung, Botvinick, & Cohen, 2004).

Formal models of decision confidence have focused on the role played by the decision variable – an internal subjective state that is influenced by incoming sensory evidence (Kepecs, Uchida, Zariwala, & Mainen, 2008; Kiani & Shadlen, 2009; Merkle & Van Zandt, 2006; Ratcliff & Starns, 2009; Vickers, 1979). For instance, in signal detection theoretic models, the absolute distance of the decision variable from a criterion is a proxy for confidence (Cartwright & Festinger, 1943; Ferrell & McGoey, 1980; Kepecs et al., 2008; Macmillan & Creelman, 2005; Suantak, Bolger, & Ferrell, 1996; Treisman & Faulkner, 1984). Dynamic extensions of signal detection theory accumulate evidence for or against a particular choice (Link & Heath, 1975; Gold & Shadlen, 2002), and several variants of this approach have linked the state of the decision variable at decision time to confidence (Kiani & Shadlen, 2009; Merkle & Van Zandt, 2006; Moreno-Bote, 2010; Ratcliff & Starns, 2009; Vickers, 1979; see Fetsch, Kiani, & Shadlen, 2015; Yeung & Summerfield, 2012 for reviews). Empirically, putative neural correlates of decision variables are also correlated with subjective confidence (De Martino, Fleming, Garrett, & Dolan, 2013; Gherman & Philiastides, 2015; Kiani & Shadlen, 2009; Komura, Nikkuni, Hirashima, Uetake, & Miyamoto, 2013; Zizlsperger, Sauvigny, Händel, & Haarmeier, 2014).

However, a close coupling between decision variables and confidence is potentially in tension with a burgeoning literature identifying dissociations between performance and metacognition. There are systematic differences between factors affecting task performance and confidence in perceptual decisions, including attentional or stimulus manipulations (Bona & Silvanto, 2014; Graziano & Sigman, 2009; Lau & Passingham, 2006; Rahnev et al., 2011; Vlassova, Donkin, & Pearson, 2014; Wilimzig, Tsuchiya, Fahle, Einhäuser, & Koch, 2008), individual differences (Baird, Cieslak, Smallwood, Grafton, & Schooler, 2015; Barttfeld et al., 2013; Fleming, Weil, Nagy, Dolan, & Rees, 2010; McCurdy et al., 2013; Song et al., 2011), developmental trajectory (E. C. Palmer, David, & Fleming, 2014; L. G. Weil et al., 2013) and brain lesions or reversible inactivation in both humans (Del Cul, Dehaene, Reyes, Bravo, & Slachevsky, 2009; Fleming, Ryu, Golfinos, & Blackmon, 2014; Rounis, Maniscalco, Rothwell, Passingham, & Lau, 2010), non-human primates (Komura et al., 2013) and rodents (Lak et al., 2014).

In addition, psychiatric and neurological disorders are often associated with impairments in self-evaluation (David et al., 2012; Fleming et al., 2014; Goldstein et al., 2009; Pannu & Kaszniak, 2005; Schmitz & Johnson, 2007; Weiskrantz, Warrington, Sanders, & Marshall, 1974).

Such dissociations may arise for a number of reasons. First, the evidence contributing to decisions may be subject to further processing that introduces additional variability into confidence reports. This further processing may occur over space and/or time. For instance, metacognitive reports may require a neural "read-out" of confidence from decision circuitry (Insabato, Pannunzi, Rolls, & Deco, 2010; Maniscalco & Lau, 2010; Shimamura, 2000). Alternatively confidence may be affected by continued processing of pre-decision evidence in time (Baranski & Petrusic, 1998; Moran, Teodorescu, & Usher, 2015; Rabbitt & Vyas, 1981; Resulaj, Kiani, Wolpert, & Shadlen, 2009; S. Yu, Pleskac, & Zeigenfuse, 2015) or the receipt of new post-decision evidence (Bronfman et al., 2015; Kvam, Pleskac, Yu, & Busemeyer, 2015; Navajas, Bahrami & Latham, 2016). Second, evidence contributing to decisions may be inaccessible to confidence reports. A canonical example is blindsight, in which cortically blind individuals may perform visual discrimination tasks well above chance but be unable to self-evaluate their performance, having a poor impression of whether they performed well or badly on individual trials (Ko & Lau, 2012; Persaud, McLeod, & Cowey, 2007; Persaud et al., 2011; Weiskrantz, 1998; Weiskrantz et al., 1974). Third, evidence contributing to confidence reports may be inaccessible to decision-making. A classic example of this phenomenon is error detection, in which human subjects rapidly signal errors made in simple laboratory tasks (Rabbitt, 1966; Rabbitt & Rodgers, 1977). The presence of the "error-related negativity" (ERN) in the scalp EEG signal around the time of the response is consistent with a rapid evaluation that one's impending response is likely to be incorrect (Gehring et al., 1993). Together these findings suggest an architecture in which evidence supporting decisions and confidence is maintained at least partly separately and in parallel (Baranski & Petrusic, 2001; Charles, King, & Dehaene, 2014; Del Cul et al., 2009; Ro, Shelton, Lee, & Chang, 2004; Schmid et al., 2010).

This variety of performance-confidence dissociations has hitherto precluded a unified account of metacognition in decision-making. Here we set out to account for such dissociations in a general framework in which confidence operates as a second-order computation about one's own performance. Our core proposal is that within a single individual, samples of sensory evidence underpinning decisions and confidence judgments are distinct but coupled[1]. Such a distinction between decision and confidence variables arises necessarily in many of the situations considered above, and once this is formally recognized, sound statistical inference differs in key ways from that prescribed by first-order signal detection theory. In our analysis, self-evaluation of decision performance is achieved by leveraging the confidence sample and one's own actions to infer the performance of the coupled decision system, over time and/or space. We develop these ideas in a Bayesian ideal observer model, at Marr's computational level, jumping off from the standard signal detection theory framework that has served as the foundation for much work in perception and metacognition. These more abstract computational considerations would, of course, be complimented by more implementational considerations at Marr's other levels of analysis, as indeed has proved a highly synergistic program in the case of signal detection theory and its real-time generalizations such as the sequential likelihood ratio test (Link & Heath, 1975; Gold & Shadlen, 2002).

It will turn out that this framework, inspired by the dissociations reviewed above, holds key implications for metacognitive computation in general. First, second-order computation naturally accommodates different behavioural manifestations of metacognition such as confidence and error detection within a common framework. The intuition, which will be formalised below, is that a secondary view on the decision problem is required for a system to view itself in error (Charles et al., 2014; James, 1950; Pasquali, Timmermans, & Cleeremans, 2010; Rabbitt, 1966). Error monitoring and

[1] See Jang, Wallsten & Huber (2012) for a related model of metamemory judgments which posits two correlated internal states contributing to memory recall and judgments of learning.

confidence have typically been studied in separate literatures (Yeung & Summerfield, 2012), but here a continuum of confidence ranging from being certain of committing an error to being sure of being correct emerges naturally from the model architecture. Second, a second-order account predicts that one's own actions will contribute to self-evaluation. The intuition here is that rather than actions simply signalling the output of a decision pathway, they may themselves carry information about the subject's internal states that is otherwise inaccessible to confidence reports.

In the sections that follow we compare the qualitative predictions of second-order computation to those made by first-order accounts with and without post-decisional processing, and evaluate these predictions against the empirical literature on decision confidence and error monitoring. We will show that first-order models are special cases of second-order computation that arise under particular noise conditions (Figure 1). Our analysis thus clarifies the situations in which these simpler architectures are suitable, and the sorts of approximations being made by adopting them when these conditions are not satisfied. We go on to demonstrate how a second-order perspective accounts for individual differences in metacognitive bias and accuracy, and may explain cases in which metacognition is sometimes better than task performance. We close by outlining the implications of this framework for future empirical studies and discuss possible neural implementations of second-order computation.

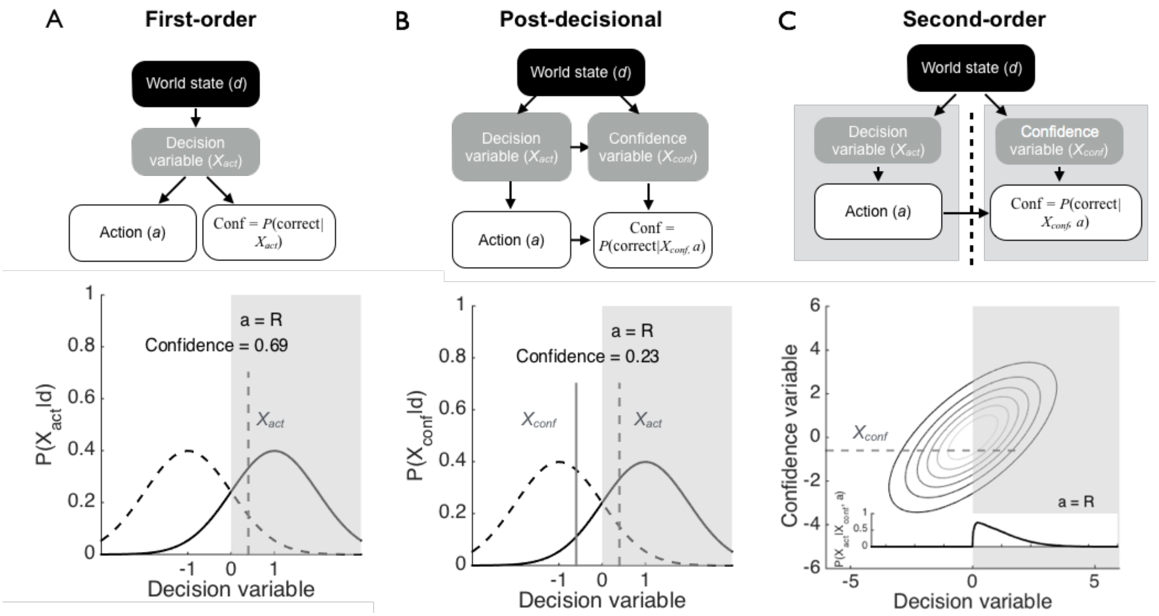

Figure 1. Schematic graphical models of self-evaluation. *Upper panels show graphical models (with variance/covariance parameters omitted for clarity). In each model, a categorical world state (e.g. stimulus = left [-1] or right [1]) gives rise to a binary action (left or right). Building on signal detection theory, we assume both stimuli give rise to internal decision variables that are Gaussian distributed along a unitary decision axis. To make an action, the observer choose "right" if the decision variable is greater than 0, and "left" otherwise. Lower panels depict a computation of confidence on a single trial of each model, in which the observer responds "right". A) First-order model. The world state generates a decision variable* X_{act} *that supports both actions and confidence reports. B) Post-decisional first-order model. As in (A), but allowing the confidence variable (*X_{conf}*) to sample additional evidence about the world state. C) Second-order model. The decision and confidence variables are represented as two correlated hidden states. A computation of decision confidence proceeds by first inferring the distribution of possible decision variables conditional on the confidence variable (shown by the probability distribution in the inset), and marginalizing conditional on the subject's action to arrive at an appropriate confidence level.*

Model overview

We consider three classes of model of how a subject generates a report of confidence in his or her decision. All models have the same basic ingredients. First, we define a categorical world state, d, such as whether a stimulus is moving left ($d = -1$) or right ($d = 1$). Second, the subject makes a response a to indicate their perceived state of the world (i.e. left, $a = -1$, or right, $a = 1$). On each "trial" internal states $X = [X_{act}\ X_{conf}]$ denote the decision and confidence variables. To make a decision, the subject chooses "right" if $X_{act} > 0$, and left otherwise:

$$a = \begin{cases} 1 \text{ if } X_{act} > 0 \\ -1 \text{ otherwise} \end{cases} \qquad (1)$$

We define the subject's confidence z as a degree of belief that a particular choice was correct (i.e. choice a reflected the true state of the world d), given a particular set of internal states X, model m and model parameters ϑ:

$$z = P(a = d|X, \vartheta, m) \qquad (2)$$

In all model simulations we assume Gaussian noise for how internal states X are generated from world states d. However, the models differ in how these states are coupled, and how confidence is computed, as described in the following sections.

First-order model

In the simplest "first-order" model we assume that the decision and confidence variables are identical, such that the same internal state supports both choices and confidence. First, the decision variable X_{act} is obtained from a Gaussian distribution conditional on the world state:

$$X_{act} \sim N(d, \sigma^2) \qquad (3)$$

The confidence variable $X_{conf} = X_{act}$. Confidence is then a transformation of the posterior belief in d conditional on the action taken (or equivalently, the sign of X_{act}):

$$z = P(a = d | X_{conf}, \sigma^2) = \begin{cases} P(d = 1 | X_{conf}, \sigma^2) \text{ if } a = 1 \\ 1 - P(d = 1 | X_{conf}, \sigma^2) \text{ if } a = -1 \end{cases} \quad (4)$$

where Bayes' rule provides the posterior probability of a particular world state (assuming flat priors on d):

$$P(d | X_{conf}, \sigma^2) = \frac{P(X_{conf} | d, \sigma^2)}{\sum_d P(X_{conf} | d, \sigma^2)} \quad (5)$$

Post-decisional model

In the post-decisional model, the confidence variable $X_{conf}$is derived from X_{act} plus additional information about the world state, X_{new}:

$$X_{conf} = X_{act} + X_{new} \quad (6)$$

For ease of exposition here we define X_{new} as an additional sample of evidence, $X_{new} \sim N(d, \sigma^2)$[2]. One can imagine different generative models – the key property here is that the true world state d is conditionally independent from the action a (and its decision variable X_{act}), given the confidence decision variable X_{conf}. Informally, X_{conf}

[2] We do not explicitly consider the within-trial dynamics of the decision variable here though we appreciate their importance for a complete account of confidence (Fetsch et al., 2015). Just as sequential sampling models represent dynamic extensions of signal detection theory (Gold & Shadlen, 2002; Link & Heath, 1975; Pleskac & Busemeyer, 2010; Ratcliff, 1978), the framework we consider here may be naturally extended to incorporate sequential samples of evidence. Because the primary aim of this paper is to contrast first- and second-order computation we restrict ourselves to the simpler, static cases, returning in the Discussion to consider the issue of dynamics in greater detail.

should provide all the information contained in X_{act}. This will be satisfied, for instance, if X_{act} and X_{conf} are both states of a perfect accumulator (with X_{conf} read out later, see e.g. Resulaj et al., 2009; van den Berg et al., 2016), but not if the accumulator is lossy or if X_{conf} arises from a noisy readout of X_{act}, degrading the signal with additional noise.

The observer then derives confidence in a similar fashion to the first-order model above:

$$z = P\left(a = d \middle| X_{conf}, 2\sigma^2\right) = \begin{cases} P\left(d = 1 \middle| X_{conf}, 2\sigma^2\right) \text{ if } a = 1 \\ 1 - P\left(d = 1 \middle| X_{conf}, 2\sigma^2\right) \text{ if } a = -1 \end{cases} \quad (7)$$

Note that the first-order model is a special case of the post-decisional model when X_{act}= X_{conf}.

Second-order model

The second-order model is subtly but importantly different from the first-order and post-decisional models. Unlike in the first-order case, confidence is not derived directly from X_{conf} – instead X_{conf} is leveraged, together with the observed action *a* and knowledge of the covariance between X_{conf} and X_{act}, to infer the state of the decider at the time of choice.

We first describe a second-order model of confidence in another individual's performance to provide the intuition for the within-subject case, and to demonstrate the symmetry between evaluating one's own actions and those of another actor. Consider two individuals, an Actor (*act*) and Confidence-rater (*conf*). The actor is carrying out a two-choice discrimination task as described above. Both receive internal samples X_{act} and X_{conf} generated from binary world state *d* (e.g. a stimulus moving left or right). We model these samples as draws from a bivariate Gaussian with covariance matrix Σ:

$$\begin{bmatrix} X_{act} \\ X_{conf} \end{bmatrix} \sim N(\boldsymbol{d}, \Sigma) \quad (8)$$

$$\Sigma = \begin{bmatrix} \sigma_{act}^2 & \rho\sigma_{act}\sigma_{conf} \\ \rho\sigma_{act}\sigma_{conf} & \sigma_{conf}^2 \end{bmatrix} \tag{9}$$

The covariance matrix has 3 parameters: σ_{act}, σ_{conf} and ρ. σ_{act} and σ_{conf} control the noise of the signal for the Actor and the Confidence-rater, respectively. The correlation parameter ρ governs the association between the two samples: capturing, for instance, the fact that the variance in the two observers' samples of the stimulus will be partly common (due to objective variation in the stimulus) and partly distinct (due e.g. to distinct sensory and neural noise). The Confidence-rater's job is to say how confident she is in the Actor responding correctly, or the posterior probability that the Actor's action *a* was appropriate for the inferred state of the world *d*, conditional on beliefs about different sources of variability. To do this, the observer infers (for the purpose of marginalizing) the state of the decision variable driving choice (X_{act}) from the confidence variable (X_{conf}):

$$z = P\left(a = d \middle| X_{conf}, a, \mathit{\Sigma}\right) = \begin{cases} P\left(d = 1 \middle| X_{conf}, a, \mathit{\Sigma}\right) & \text{if } a = 1 \\ 1 - P\left(d = 1 \middle| X_{conf}, a, \mathit{\Sigma}\right) & \text{if } a = -1 \end{cases} \tag{10}$$

where $P\left(d \middle| X_{conf}, a, \mathit{\Sigma}\right) \propto P\left(d \middle| X_{conf}, \mathit{\Sigma}\right) P\left(a \middle| X_{conf}, d, \mathit{\Sigma}\right)$

$$= P\left(d \middle| X_{conf}, \mathit{\Sigma}\right) \int P(a | X_{act}, \mathit{\Sigma}) P(X_{act} | X_{conf}, d, \mathit{\Sigma})\, dX_{act} \tag{11}$$

The core of our proposal is that individuals generate confidence in their own performance by applying an analogous computation to their own actions (Fig. 1C). Importantly, in Equation 10 the probability of being correct is determined not only by X_{conf} but also one's own action *a* and beliefs about the fidelity of the decision and confidence variables, captured by Σ. In other words, second-order inference reflects an active process of inferring the state of the decider, rather than a passive sensitivity to the difficulty of the decision. In Appendix A we derive analytic solutions to this equation for two-choice decision scenarios assuming Gaussian noise.

In the between-subject case, we might expect limited correlation between the confidence and decision variables, as depicted in Figure 2A. In the within-subject case, this correlation may be higher, although one evidence stream may be noisier than the other, thereby weakening the information that either the Actor or the Confidence-rater has about the true world state (Fig. 2B). The model architecture is agnostic about how the relationship between X_{act} and X_{conf} arises: it may be that they remain segregated in the brain (e.g. in parallel pathways); X_{conf} may depend on the same neural activity as X_{act} at a later time point, or X_{conf} may reflect a noisy read-out of X_{act}. The many possible relationships between X_{act} and X_{conf} are flexibly accommodated via the parameters of the covariance matrix Σ. In the special case in which $\rho = 1$ and $\sigma_{act} = \sigma_{conf}$, the second-order model reduces to the first-order case, as on any given trial the same evidence supports both actions and confidence (Fig. 2C).

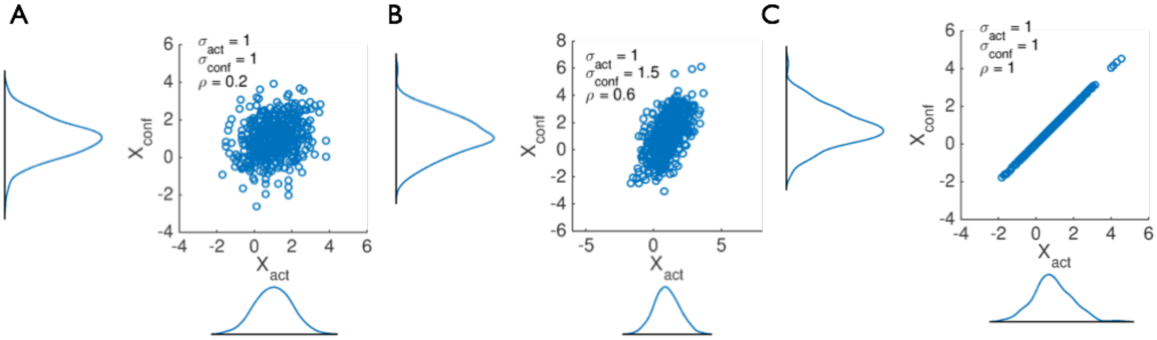

Figure 2. Illustration of effects of second-order model parameters on decision and confidence variables. *Each panel shows samples of the decision variable (X_{act}) and the confidence variable (X_{conf}) drawn from models with different parameter settings. The correlation coefficient ρ increases from (A) to (C). Panel (B) shows the effect of selectively increasing the variability in the confidence variable (compare the width of the marginal distributions of X_{conf} and X_{act}). The parameter settings in panel (C) mimic a first-order model in which X_{act} and X_{conf} are identical.*

We note that these model variants are naturally nested, with each representing an extension of the previous case. The first-order model is a special case of the post-decisional model in which the decision and confidence variables are identical, and the post-decisional model is a special case of the second-order model in which X_{conf} is a sufficient statistic for X_{act} with respect to d (e.g. when evidence is accumulated without forgetting). Indeed, variants of the first-order or post-decisional models outlined above are optimal under limited cases in which the confidence computation has direct access to the actor's decision variable. However the computational considerations we highlight here apply to all but the simplest cases in which internal states underpinning performance are transparently accessible to those underpinning confidence.

RESULTS (1): FEATURES OF SECOND-ORDER COMPUTATION

In this section we describe qualitative features of first- and second-order computation, and relate these to key findings in the empirical literature.

Relationship between decision confidence, accuracy and stimulus strength

We begin with internal representations supporting decision confidence. Decision confidence typically increases with stimulus evidence for correct judgments, but decreases with stimulus evidence for errors ("X-pattern", Figure 3; Kepecs et al., 2008; Lak et al., 2014; Sanders et al., 2016; although see Kiani et al., 2014). Here we show that all three model variants are able to reproduce this pattern, and therefore observing an X-pattern in behaviour is not diagnostic of first- or second-order computation.

First-order model

To simulate confidence as a function of stimulus strength we modified all models such that the sample mean depends on stimulus strength θ (varying between 0 and 1; $\mu = d\theta$; see Appendix B for details of this and other simulations). The upper panel in Fig. 3A shows that the first-order model reproduces the qualitative X-pattern observed in the behavioural data despite the confidence and decision variables being identical. The intuition for this pattern is as follows. A given direction d and stimulus strength θ leads to a range of samples X_{act}, and the possibility of erroneous responses. As θ increases, the likely values of $|X_{conf}|$ $(= |X_{act}|)$ following an incorrect response therefore decrease in magnitude. To take a concrete example, suppose we have a leftward trial (d = -1). If the subject's sample X_{act} is +0.05, she will erroneously respond "right" and derive confidence from a monotonic transformation of $|X_{conf}|$. But this subjective sample may have arisen from many different objective stimulus strengths θ, including both correct and error trials, and occur more often with some than others. When the experimenter then plots the subject's confidence as a function of the externally-manipulated variable θ, a divergent pattern of confidence emerges for correct and error trials. In other words, the X-pattern is due to the necessity of relating observed confidence to θ (which is unknown to the subject) rather than to X_{conf} (which is unknown to the experimenter).

However, if it were possible to determine the decision variable X_{conf} on individual trials, we would predict that confidence always scales monotonically with $|X_{conf}|$ for both correct and error trials in the first-order case (Fig. 3A, lower panel). The internal state representation of a first-order model does not show the X-pattern[3].

Post-decisional model

The same X-pattern is obtained for confidence derived from simulations of the post-decisional model (Figure 3B). However, in this case the model's internal state diverges as a function of choice accuracy due to cases in which the decision and confidence variables dissociate (cf. Figure 1B). In other words, if it were possible for the experimenter to know X_{conf} on a single trial, a post-decisional account would predict divergent

[3] Kepecs and colleagues have shown that spike rates of single neurons in orbitofrontal cortex, a putative neural correlate of confidence, show an X-pattern as a function of external stimulus strength (Kepecs et al., 2008). Given that the internal state of a first-order model does not show this pattern, it is tempting to instead conclude that a more complex model is needed to account for these findings. However this conclusion does not necessarily follow: again, the experimenter has access only to stimulus strength rather than the animal's decision variable, and similar considerations apply as when interpreting behavioural data.

relationships between confidence and X_{conf} on correct and error trials (Figure 3B, lower panel).

Second-order model

Finally, the behavioural X-pattern also emerges from a second-order computation of confidence, but for different reasons (Figure 3C). Here the model detects its own errors by applying second-order inference. Specifically, given a sample X_{conf}, the model generates a probability that its action matched the most likely state of the world. In this case, confidence decreases on error trials with increasing θ because there tends to be increasing evidence (from X_{conf}) that the action taken was inappropriate. As in the post-decisional model, an interaction with choice accuracy is also observed in the model's internal state (Figure 3C, lower panel).

In summary, all models are able to account for the X-pattern relating confidence to stimulus strength as a function of accuracy, but do so for different reasons. The pattern emerges in the first-order model due to an imprecise mapping between the experimenter-observed variable θ and internal state X_{act}; it emerges in the second-order model due the effect exerted by beliefs counter to one's choice on the posterior probability of having made a correct action. The internal states of the post-decisional and second-order models also show the X-pattern observed in behavior.

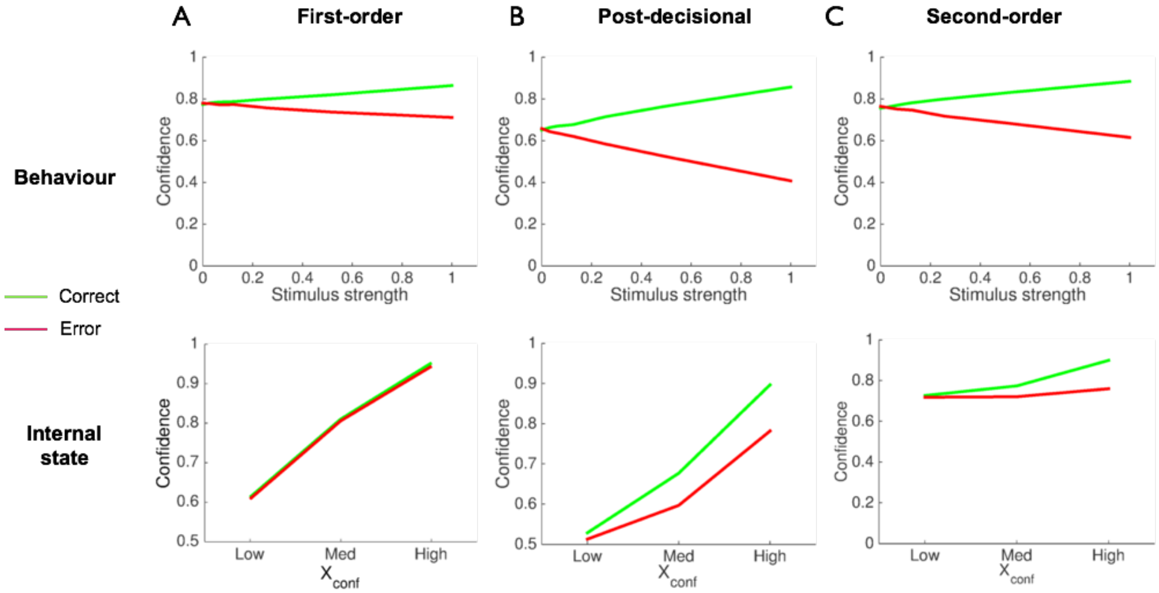

Figure 3. Internal representations supporting decision confidence. *Simulations of first-order (A), post-decisional (B) and second-order (C) models showing how confidence changes as a function of stimulus strength and decision accuracy. The upper panels show confidence as a function of objective stimulus strength; the lower panels show confidence as a function of the internal state of each model.*

Relationship between confidence and error detection

Human subjects are able to rapidly detect errors made in simple laboratory tasks (Rabbitt, 1966; Rabbitt & Rodgers, 1977). Other work has investigated the dynamics of changes of mind – a switch from an initial, often erroneous response to an alternative, correct response (Resulaj et al., 2009). Both error detection and changes of mind can be formalized as a subjective probability of success for a chosen action being lower than that of an alternative action, which in two-choice discrimination corresponds to a decision confidence level less than 0.5.

It is notable that such representations are precluded in the simplest first-order model because the same evidence drives both choices and confidence, resulting in a lower bound on confidence of 0.5 (Fig. 4A). In other words, if a single decision variable indicates that the alternative option is preferable, then the action also follows suit;

dissociations between actions and confidence do not occur and confidence is monotonic in $|X_{conf}|$. In contrast, in both the post-decisional and second-order models (Fig. 4B, C), confidence maps out a space from being sure that an error has been committed to being sure of a correct response, due to regimes in which the model infers that its action a was at odds with the most probable direction *d*, and there is no longer a monotonic mapping between $|X_{conf}|$ and confidence. Finally, Figure 4C illustrates a feature of second-order computation that we will return to below: even when the confidence variable provides equivocal evidence about the world (X_{conf} = 0), the model's confidence is not necessarily at chance (0.5). Instead, for the parameters used in this simulation, confidence when X_{conf} = 0 is around 0.7, due to the confidence computation also incorporating knowledge about the average reliability of actions, i.e. σ_{act} (Drugowitsch, Moreno-Bote, & Pouget, 2014). In summary, post-decisional and second-order models are able to reproduce error-detection-like behavior (*P*(correct) < 0.5), but the simplest first-order model cannot.

The internal representations of the second-order model that support error detection are illustrated in Figure 4D. Here we sampled moderately correlated samples of X_{act} and X_{conf} from world state *d* = 1 (i.e. the true stimulus class is "right"). By applying a neutral decision criterion, the observer erroneously responds "left" whenever X_{act} is less than zero. However, whether this error will be detected depends on whether X_{conf} provides enough (positive) evidence in support of the alternative, correct response (orange samples in Figure 4D). The proportion of detected errors is itself governed by the covariance of X_{conf} and X_{act}. Figure 4E simulates the proportion of detected errors for a constant performance level (σ_{act} = 1; ~84% correct). Error detection is highest when σ_{conf} is low, due to the confidence variable providing accurate information about the true world state. Notably error detection also depends on the correlation between the samples – as ρ approaches 1 (lower right quadrant of the heatmap) the model reduces to the first-order case and error detection is again precluded.

These simulations of error detection are of course an over-simplification – the criterion for whether to report an error is itself under subject control, and may be adjusted above or below 0.5 in the face of changing incentives (Neyman & Pearson, 1933; Steinhauser & Yeung, 2010). The aim here is simply to show that both post-decisional and second-order models naturally handle error detection and changes of mind by modeling cases in which the confidence and decision variables disagree.

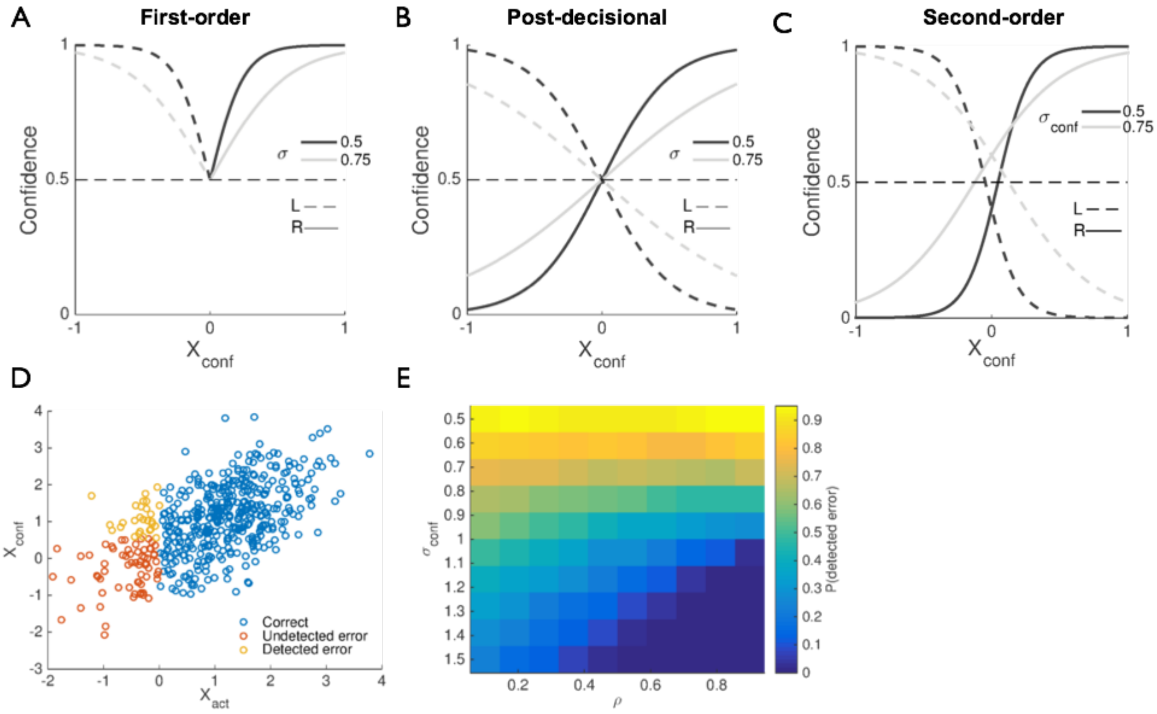

Figure 4. Internal representations supporting error detection. *A) Confidence as a function of the decision variable and uncertainty parameter σ in the first-order model. B, C) Confidence as a function of the confidence variable, chosen action and uncertainty parameter σ_{conf} in the post-decisional model (B) and second-order model (C) D) Simulation of how error detection emerges from correlated samples in the second-order model. Samples are generated from a true world state $\mathrm{d} = 1$ with parameter settings $\sigma_{act} = 1$, $\sigma_{conf} = 1$ and $\rho = 0.6$. The model makes errors when X_{act} falls to the left of the neutral (0) criterion. A subset of these objective errors are "detected" due to the confidence variable providing evidence that the alternative action is preferred, generating a confidence level of less than 0.5. D) Heat map revealing how the proportion of detected errors in (C) varies according to model parameters σ_{conf} and ρ . Objective accuracy (governed by σ_{act}) is constant.*

Influences of self-generated actions on confidence

A counterintuitive but important feature of second-order computation is that one's own actions may causally affect subsequent confidence ratings, particularly if X_{act} and X_{conf} are only weakly coupled. This influence arises because actions carry information about the subject's internal states, leading a rational observer to incorporate her own actions as additional data when computing confidence. Consider Figure 5A. Plotted on the y-axis is the posterior probability that the current world state is rightward (*d* = 1) as a function of confidence variable X_{conf}. Intuitively, as X_{conf}becomes more positive, the model gains greater evidence that *d* = 1. However, having taken an action a, this inference is modulated, such that a leftward action reduces the belief in rightward world states, whereas a rightward action boosts it.

To further explore this effect, we simulated the model's confidence after "clamping" X_{conf} at 0. In the absence of any action (grey line in Figure 5B and C), the model is equivocal about the world state and confidence remains at 0.5. However, after an action is made, the model leverages this new information to modulate its belief in *d*. The extent to which this modulation occurs is dependent on (beliefs about) the covariance of X_{act} and X_{conf}. As the confidence variable becomes more noisy (σ_{conf} increases), the information provided by $X_{conf}$is less reliable and confidence is more influenced by actions (Fig. 5B). Conversely, as the correlation between X_{act} and X_{conf} increases (ρ increases), actions provide less new information about the possible values of *d*, and the modulation of confidence by action decreases (Fig. 5C).

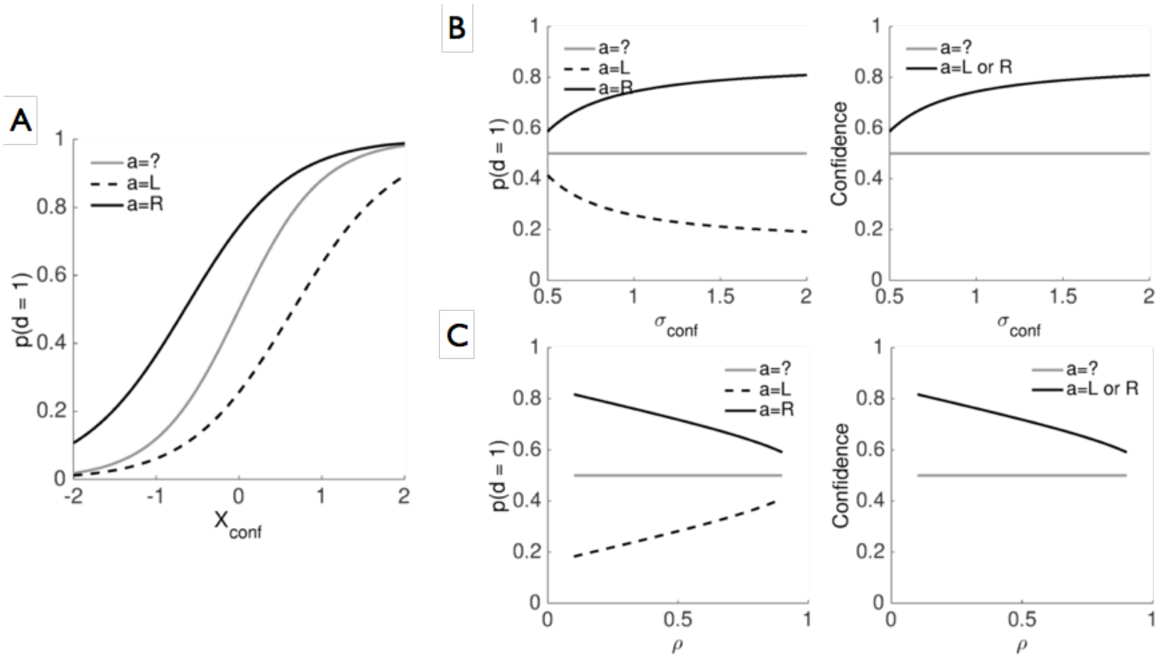

Figure 5. Influence of choices on second-order model confidence. *A) Posterior probability of a rightward world state as a function of confidence variable X_{conf} and the chosen action. B, C) The lefthand panels show the influence of actions on the posterior probability of d = 1 for a constant, uninformative sample ($X_{conf} = 0$). The righthand panels show the corresponding confidence level. B) As the confidence variable becomes less informative (σ_{conf} increases), actions have a greater effect on posterior beliefs. C) As the correlation between X_{act} and X_{conf} increases, actions provide less new information about the possible values of d, and their influence on confidence reduces. Constant parameters in all panels are set at $\sigma_{act} = 1$, $\sigma_{conf} = 1$, $\rho = 0.4$.*

This feature of the model leads to a counter-intuitive empirical prediction: elicitation of actions should affect confidence judgments. For instance, if subjects are asked to rate their confidence *before* their response (i.e. confidence in making a hypothetical response), then they may compute their confidence without conditioning on self-action (which is precluded in this case unless subjects covertly choose and then rate; Fig. 6A). This leads to two effects (Fig. 6B). First, the difference in confidence between correct and error trials should be greater (metacognitive sensitivity should increase) when ratings are given after a decision than before, due to the additional diagnostic information provided by the action. Second, ratings given after a decision should be systematically lowered compared to those given before (Figures 7A and B show that these qualitative effects are obtained across a large range of second-order model parameters). In contrast, actions do not provide any additional diagnostic information about hidden states in first-order accounts, and in the absence of additional post-decision evidence, confidence levels are equivalent whether elicited pre- or post-decision (Fig. 6C).

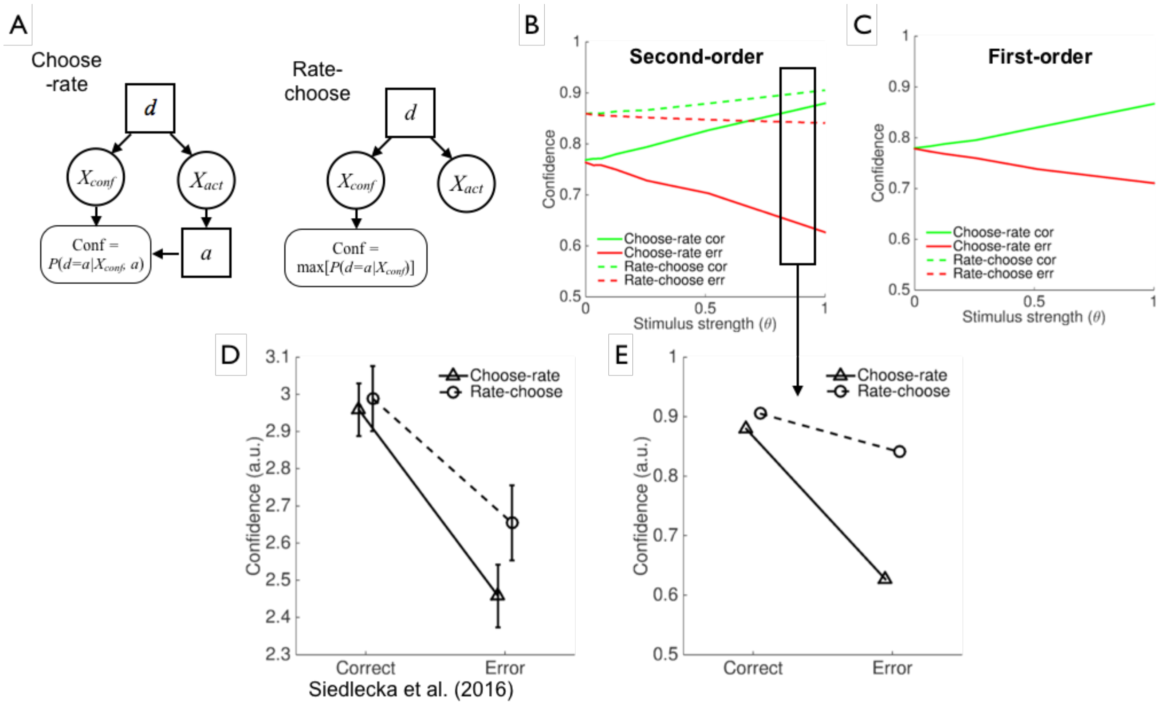

Figure 6. Predicted effects of choice on confidence. *A) Graphical models for choose-rate and rate-choose experiments illustrating the influence of actions on confidence in the choose-rate condition. B) Simulation of confidence from choose-rate and rate-choose experiments as a function of stimulus strength and decision accuracy for the second-order model (*$\sigma_{act} = 1$*,* $\sigma_{conf} = 1$*,* $\rho = 0.6$*). Overall confidence (bias) decreases relative to the rate-choose condition when choices are made before confidence ratings (choose-rate), whereas the difference in confidence between correct and error trials (metacognitive sensitivity) increases. C) As in (B) for the first-order model (*$\sigma_{act} = 1$*). Here the predictions for confidence from the choose-rate and rate-choose models are identical and the dotted lines are obscured. D) Data replotted from Siedlecka et al. (2016), with permission, in which choice and rating order were manipulated. E)*

Simulations of second-order model predictions at constant stimulus strength, plotted using same conventions as (D).

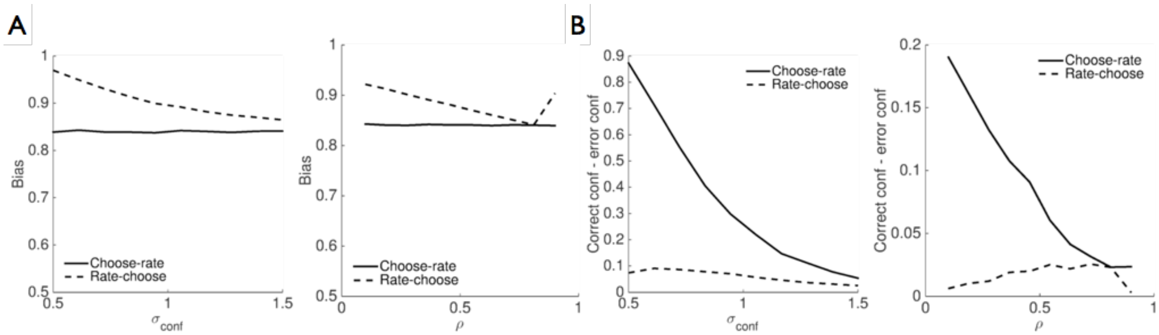

Figure 7. Effects of choice on confidence across a range of second-order model parameter settings. *A) Plots of bias as a function of model parameters σ_{conf} (left panel) and ρ (right panel). Across a range of parameter settings confidence is decreased in the choose-rate condition. In the σ_{conf} simulation, $\rho = 0.6$, whereas in the ρ simulation, $\sigma_{conf} = 1$. B) Similar to (A) for metacognitive sensitivity (the difference between correct and error confidence). Across a range of parameter settings metacognitive sensitivity is increased in the choose-rate condition.*

Empirical observation of a pattern similar to that depicted in Figure 6B would therefore provide support for a second-order model of confidence. While revising our manuscript for publication (and after developing these simulations) we became aware of a published dataset that directly tested and confirmed our predictions (Figure 6D). Siedlecka et al. (2016) asked subjects to provide confidence ratings about whether a target word presented on the screen was the solution to a previously-studied anagram. In a between-subjects design, participants were assigned to one of three conditions: deciding if a target word was an anagram and then judging confidence (target-decision-metacognitive judgment, tDM); judging confidence after seeing the target but before making a decision (tMD); or rating the confidence of their decision before seeing the target word (MtD). Here we focus on the difference between the tDM and tMD conditions, as they represent direct analogues of our choose-rate and rate-choose simulations. In Figure 6D we replot their data alongside the second-order model simulation at constant stimulus strength (Figure 6E). Siedlecka et al. (2016) found that metacognitive sensitivity was greater in the tDM than the tMD conditions, in accordance with the predictions of a second-order model in which actions inform confidence ratings. In addition, confidence was overall lower in the rate-choose case, although unlike the effect on metacognitive sensitivity, this was not statistically significant. As can be seen by comparing Figure 6D and E, the second-order model simulation qualitatively captures the patterns observed in Siedlecka et al's experiment.

Conclusions

In this section we have explored features of first- and second-order models of confidence, and compared their qualitative predictions against empirical findings on confidence and error detection. We find that while all models can reproduce relationships between stimulus strength, accuracy and confidence, only post-decisional and second-order models permit levels of confidence that may support error detection, and only a second-order account naturally accommodates findings that actions themselves influence confidence judgments. These results are summarized in Table 1.

	First-order	**Post-decisional**	**Second-order**
X-pattern in confidence	Yes	Yes	Yes
Error detection	No	Yes	Yes
Effects of choice on confidence	No	No	Yes

Table 1. Summary of model variants and their ability to accommodate qualitative features of empirical data.

RESULTS (2): DISSOCIATIONS BETWEEN PERFORMANCE AND CONFIDENCE

Modeling dissociations between performance and confidence

Metacognitive accuracy refers to the relationship between self-evaluation and performance, and is comprised of two components: sensitivity and bias[4] (Fleming & Lau, 2014). Metacognitive sensitivity refers to the extent to which a subject can discriminate correct from incorrect performance on a first-order task, and can be assessed with type 2 receiver operating characteristic (ROC) analysis (Clarke, Birdsall, & Tanner, 1959; Galvin, Podd, Drga, & Whitmore, 2003) or meta-*d'*, which indexes metacognitive sensitivity in units of decision *d'* (Maniscalco & Lau, 2012; 2014). The logic is that if an observer has good sensitivity, she will be able to discriminate between her own correct and incorrect responses through offering up suitable confidence reports – lower confidence when incorrect, and higher confidence when correct. Metacognitive bias is the tendency to give higher overall confidence ratings, all else being equal. Note that bias is potentially independent of sensitivity – a subject might have high overall confidence but be unable to discriminate between correct and error trials.

In this section we show that second-order computation naturally accommodates changes in metacognitive sensitivity and bias through alterations in covariance parameters and beliefs about covariance parameters (hyperparameters), respectively, and handles cases in which metacognitive sensitivity is either better or worse than performance.

Metacognitive sensitivity

Two distinct (but not mutually exclusive) sets of parameter changes may lead to reductions in the second-order model's metacognitive sensitivity. In the first, metacognitive sensitivity is impoverished (type 2 ROC area is reduced) as the noise in the confidence variable σ_{conf} is increased (Fig. 8A). In the second, σ_{conf} remains constant but the correlation between X_{conf} and X_{act} is increased, leading to decreased metacognitive sensitivity despite task performance remaining constant (Fig. 8B). In other words, while the precision of the confidence variable remains constant, increased coupling between the confidence and decision variables reduces the model's ability to detect when its behavior may have been inappropriate (cf. Fig 4D).

[4] The related terms resolution and calibration are often employed in studies of probability judgments.

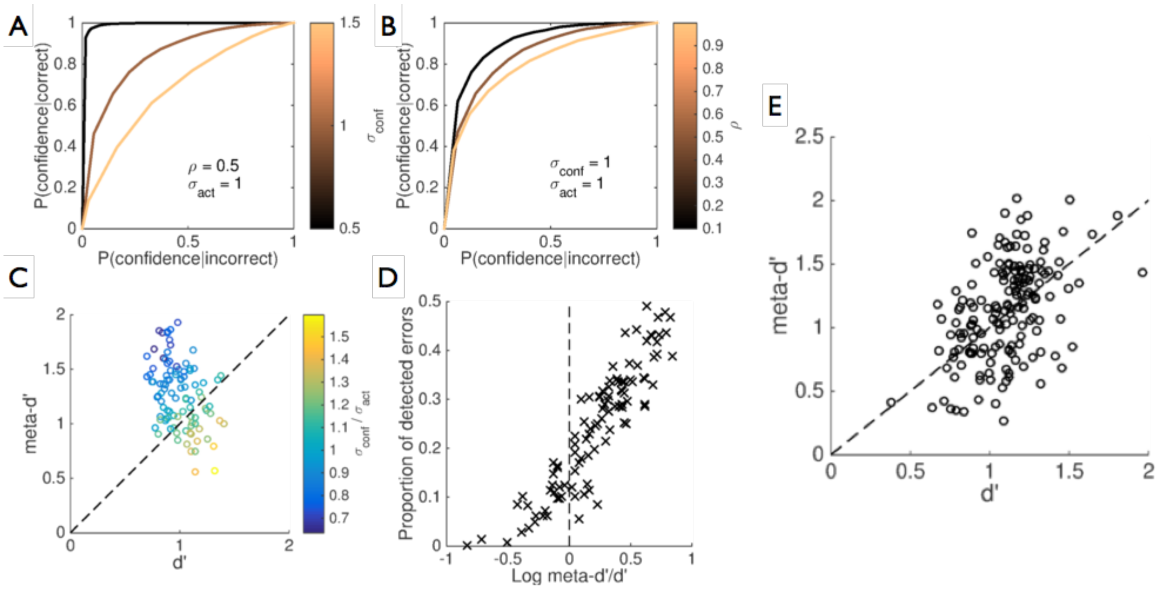

Figure 8. Modeling changes in metacognitive sensitivity in a second-order framework. *A) Simulated type 2 ROCs for different levels of noise in the confidence variable,* σ_{conf}*. As* X_{conf} *becomes more variable, metacognitive sensitivity is reduced despite task performance remaining constant. B) Simulated type 2 ROCs for different levels of* ρ*. As the correlation between the confidence and decision variables is increased, metacognitive sensitivity is decreased. C) Relationship between d' and meta-d' of simulated datasets colour-coded by settings of model parameters* σ_{conf} *and* σ_{act} *(*$\rho = 0.5$*). Cases of "hyper"-metacognitive sensitivity in which meta-d' > d' are associated with parameter ratios less than 1, indicating greater variability in the decision variable compared to the confidence variable. D) Relationship between meta-d'/d' of simulated datasets and proportion of detected errors in each dataset. Cases of meta-d'/d' > 1 (log(meta-d'/d') > 0) are associated with an increase in the number of detected errors. E) Plot of d' against meta-d' obtained from data pooled across a number of empirical studies* (Fleming et al., 2010; Fleming, Huijgen, & Dolan, 2012; E. C. Palmer et al., 2014; L. G. Weil et al., 2013)*, demonstrating the substantial frequency of hyper-metacognitive sensitivity observed in these datasets.*

Accounting for hyper- and hypo-metacognitive sensitivity

In signal detection theoretic approaches to metacognition, type 1 performance provides a theoretical upper bound on the type 2 ROC (Galvin et al., 2003). In other words, it is not possible, under these accounts, for more signal to be available to the Confidence-rater than available to the Actor. Maniscalco and Lau provided an elegant method for comparing metacognitive sensitivity and performance by characterizing metacognitive sensitivity in units of type 1 *d'*, which they label meta-*d'* (Maniscalco & Lau, 2012). In this approach, an ideal observer's meta-*d'* equals *d'*, or the ratio meta-*d'/d'* = 1. Suboptimal or hypo-metacognitive sensitivity results in values of meta-*d'/d'* < 1 (Barrett, Dienes, & Seth, 2013; Maniscalco & Lau, 2014). Maniscalco and Lau suggested that empirical values of meta-*d'/d'* > 1 ("hyper"-metacognitive sensitivity) may be due to artifacts of estimation error or criterion variability. But in our experience, such values are routinely observed in empirical studies (see Figure 8E), and recent work has highlighted that in certain circumstances hyper-metacognitive sensitivity may be more common than previously assumed (Charles, Van Opstal, Marti, & Dehaene, 2013; Scott, Dienes, Barrett, Bor, & Seth, 2014).

Building on the simulations of error detection considered above, we can understand how hyper-metacognitive sensitivity may naturally arise as a consequence of post-decisional and/or second-order computation. If the confidence variable provides additional valid information about the world state (in the second-order model, when $\rho < 1$ and σ_{conf} is low), the model reliably detects its own errors (Figs. 4D, 8A and 8B). This may lead to circumstances in which metacognition is "better" than performance, i.e. meta-*d'* > *d'*. To demonstrate this we randomly sampled simulated datasets generated from a particular combination of σ_{act} (*d'*) and σ_{conf} (holding ρ constant at 0.5), and fitted meta-*d'* to each dataset. Figure 8C plots *d'* against meta-*d'*, colour-coded according to the ratio of model parameters $\sigma_{conf}/\sigma_{act}$. It can be seen that when this ratio is small, values of meta-*d'* > *d'* are routinely obtained. Furthermore, when we interrogate the relationship between the proportion of detected errors (i.e. errors with confidence < 0.5), hyper-metacognitive sensitivity is associated with the emergence of error detection in the model (Figure 8D).

These results demonstrate that both hypo- and hyper-metacognitive sensitivity are accommodated by a second-order framework.

Bias/calibration

Up until now we have assumed that the covariance parameters associated with internal states are identical to those entering into the model inversion step when computing confidence. This is presumably an over-simplification. Instead, a subject's beliefs (hyperparameters) about these parameters may be malleable, leading to systematic over- or underconfidence (Adams, Stephan, Brown, Frith, & Friston, 2013; Drugowitsch et al., 2014), and potentially accounting for systematic biases in self-evaluation.

To illustrate how changing hyperparameters leads to bias, in Figure 9 we plot the model's aggregate performance (proportion correct) conditioned on 10 levels of confidence for different settings of beliefs about parameters σ_{act}, σ_{conf} and ρ. Importantly, for all simulations the actual parameters used to generate internal samples and decisions were fixed at σ_{act}= 1.5, σ_{conf}= 1, ρ = 0.6. The deviation of the curves from the identity line show that subtly different beliefs about the true underlying parameters are sufficient to produce a range of patterns of systematic over- or under-confidence, typical of the probability distortions observed in the experimental literature (Drugowitsch et al., 2014; Harvey, 1997; Zhang & Maloney, 2012).

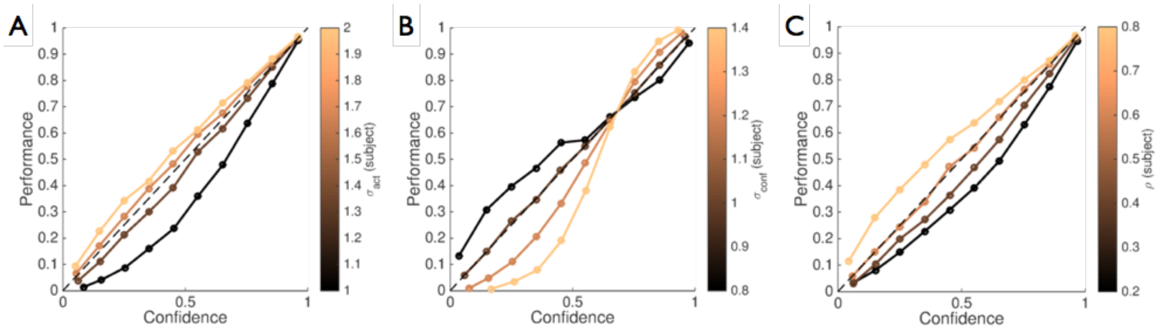

Figure 9 Modeling changes in metacognitive bias in a second-order framework. *Simulated performance levels conditioned on 10 equally spaced confidence bins for different beliefs about parameters (A) σ_{act} , (B) σ_{conf} or (C) ρ. In each panel we manipulated beliefs about the relevant parameter while holding the other two parameters*

constant. For all simulations the actual parameters used to generate samples were fixed at σ_{act}= 1.5, σ_{conf}= 1, ρ = 0.6.

DISCUSSION

We have proposed that metacognitive judgments of decision-making may depend on second-order computation about behaviour, computationally equivalent to inferring the performance of another actor. A key insight is that as soon as one recognizes a distinction between the decision variable controlling behavior, versus the information guiding the confidence judgment, then except in special cases, correctly judging confidence requires inferring the causes of one's own behavior. This general formalism subsumes several cases in which the internal states underlying performance and confidence may differ, such as dissociations over space and time. Second-order computation accounts for different behavioural manifestations of metacognition such as confidence and error detection within a single computational scheme. Furthermore, by positing coupled hidden states, a second-order framework naturally handles dissociations between performance and metacognition.

Nested within a second-order framework are simpler first-order accounts. We find that while first-order models can reproduce the empirical inter-relationship of confidence, stimulus strength and accuracy, only post-decisional and second-order models reproduce confidence levels that support error detection, and only the second-order model accommodates findings that actions themselves influence confidence judgments. Thus while we do not wish to propose that second-order computation always underpins confidence reports, some features of empirical data are at least consistent with the operation of second-order computation in a subset of cases. While intentionally broad in scope, a second-order framework nevertheless makes concrete empirical predictions, including the influence of actions upon decision confidence and the commonality between neural mechanisms supporting confidence and error detection. Here we consider in greater detail how our model relates to previous models of error detection and confidence, and explore possible neural implementations of second-order computation.

Relationship to previous models of metacognition in decision-making

Models of error detection

A second-order framework suggests that errors are detected as a mismatch between an inference on the world state and the selected action. This approach is consistent with earlier accounts of error monitoring that emphasize the comparison between intentions and actions (Charles et al., 2014; Coles, Scheffers, & Holroyd, 2001; Holroyd & Coles, 2002; Holroyd, Yeung, Coles, & Cohen, 2005; Rabbitt & Rodgers, 1977). While initially this literature focused on binary error signaling, there has been increasing recognition that similar principles may also underpin graded confidence judgments (Boldt & Yeung, 2015; Scheffers & Coles, 2000; Yeung & Summerfield, 2012). One influential model of error detection suggests that activation of two competing responses leads to conflict (and associated activation in the anterior cingulate cortex), and this conflict triggers the detection of an impending error (Yeung et al., 2004). An alternative perspective is that error detection relies instead on computing the likelihood of an error occurring in a given context (Alexander & Brown, 2011; Brown & Braver, 2005). The current framework provides a potential bridge between these accounts – error detection relies on "conflict" between two streams of evidence (see Fig. 4C), but rather than the model signaling this conflict *per se*, it harnesses this disagreement to infer a probability that an error will occur.

Holroyd and colleagues proposed a neural network model of error detection which assigned value to state-action conjunctions by reinforcement learning (Holroyd et al., 2005; Holroyd & Coles, 2002). Once the model has been trained, actions that are inappropriate for a given state became associated with negative values, leading to a negative prediction error (and associated error-related negativity) at the time of response. This scheme also shares commonalities with second-order computation in that confidence is conditional on both state and action variables. However it differs in that second-order computation does not explicitly represent stimulus-response conjunctions. Instead such associations are implicit in inverting a generative model of action when evaluating one's performance.

Models of confidence

Several previous models of confidence have built upon evidence accumulation models of decision-making, accounting for key inter-relationships between choice, confidence and response time (De Martino et al., 2013; Kiani & Shadlen, 2009; Kiani, Corthell, & Shadlen, 2014; Merkle & Van Zandt, 2006; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009; Vickers, 1979). One instance where decoupling of information underlying decision and confidence arises is when a single representation of decision evidence evolves over time, as in our post-decisional model simulations. This idea – a sort of bridge along the way from first to second-order models – has been used to model confidence and changes of mind (Moran et al., 2015; Pleskac & Busemeyer, 2010; Resulaj et al., 2009; van den Berg et al., 2016), and can also be seen as a special case of the framework we present here. (In particular, as we discuss further below, our analysis indicates that even in this first-order-like case, a confidence judgment should be informed by the chosen action, unless the accumulation is perfect and without decay). We note the relationship between decision time and confidence is likely to be complicated, and dependent on the task and goal of the observer (Pleskac & Busemeyer, 2010; Koizumi et al., 2015). However, a compelling avenue for future work is to unfold second-order computation in time, propagating multiple hidden states, just as the drift-diffusion model represents a temporal unfolding of classical signal detection (Ratcliff, 1978). Initial work along these lines has explored how the propagation of multiple internal decision variables holds promise for unifying accounts of decisions and subjective reports (Del Cul et al., 2009; Fuss & Navarro, 2013; Kvam et al., 2015; Zandbelt, Purcell, Palmeri, Logan, & Schall, 2014). Such models may provide computational insights not only into the dynamics of self-evaluation, but also the evaluation of the decisions of others (Patel, Fleming, & Kilner, 2012).

More broadly, there is an ongoing debate over whether confidence computation is best accommodated by serial or parallel architectures (Fleming & Dolan, 2012; Maniscalco & Lau, 2014; Pleskac & Busemeyer, 2010). Maniscalco and Lau found that a signal detection model in which confidence is derived from a noisy hierarchical representation of evidence supporting a choice provided a better fit to rating data than alternatives in which evidence for choices and confidence evolved in parallel (Maniscalco & Lau, 2016). Similarly, Pleskac & Busemeyer's 2-stage dynamic signal detection (2DSD) model proposes that a decision variable continues accumulating beyond the decision time, at which point confidence is determined by its relation to a set of response criteria (Pleskac & Busemeyer, 2010). This model accounts for a number of relationships between decision time, post-decision time and confidence. However serial accumulation may not be sufficient to account for cases in which error detection is very fast, consistent with a parallel representation of evidence against the decision (Charles et al., 2014; Rabbitt, 1966). Del Cul and colleagues accordingly suggested that information for decisions and subjective reports is accumulated in parallel, and this architecture was able to mimic a selective alteration in subjective reports due to prefrontal brain damage (Del Cul et al., 2009).

A second-order approach offers a broader perspective on this debate, subsuming several special cases. Specifically, depending on the covariance of the model's internal states, confidence ratings may appear to be determined by a hierarchical or parallel architecture. For instance, if $\sigma_{act} < \sigma_{conf}$ and ρ is high, the model will appear hierarchical, in that confidence depends on the same evidence as actions, albeit with added noise. Conversely, if ρ is low, the model operates in a parallel fashion, and as σ_{act} approaches zero, cases of "blind insight" may occur in which the model is aware of making erroneous or correct actions despite performing at or near chance (Scott et al., 2014). Finally, there may be domains or tasks in which confidence reports show a particularly high degree of sophistication in tracking task performance, which would suggest that decision and confidence variables are tightly coupled, with little opportunity for dissociations (e.g. Barthelme & Mamassian, 2010; Meyniel et al., 2015b; Peters & Lau, 2015).

A further implication of second-order computation is that common mechanisms should support both confidence judgments and monitoring of errors. Most previous work on error monitoring has focused on discrete cases in which actions diverge from intentions under time pressure. The canonical finding is that an error-related negativity (ERN) originating in the anterior cingulate cortex is observed time-locked to the onset of the erroneous response (Dehaene, Posner, & Tucker, 1994; Gehring et al., 1993). In contrast, studies of confidence have tended to focus on cases in which perceptual uncertainty is manipulated but response requirements are trivial (although see Faisal & Wolpert, 2009; Fleming, Maloney, & Daw, 2013). There is now increasing recognition that multiple sources of variability affect the strength of error- and confidence signals in the brain; for instance, neural signatures of error detection are also modulated by the degree of sensory uncertainty of the subject (Charles et al., 2013; Navarro-Cebrian, Knight, & Kayser, 2013; Scheffers & Coles, 2000). In support of this idea, Boldt and Yeung recently provided direct evidence for a common neural substrate for confidence and error detection. By applying multivariate decoding analyses to EEG data recorded during a visual discrimination task, they showed that neural markers of error detection were also predictive of varying levels of confidence in correct choices (Boldt & Yeung, 2015).

Varieties of metacognitive inaccuracy

The ability to discriminate one's own correct and incorrect responses can be quantified by type 2 ROC analysis (Clarke et al., 1959; Galvin et al., 2003; Maniscalco & Lau, 2012; 2014). Recently Maniscalco and Lau developed an elegant measure of metacognitive sensitivity, meta-*d'*, that quantifies the type 2 ROC area in units of first-order *d'* (Maniscalco & Lau, 2012). As shown in Figure 8, there may be a number of reasons for low meta-*d'* in the current framework. Increased noise in the confidence variable may impair inference on world states and therefore impair recognition of correct or incorrect responses. Conversely, an increase in correlation between the decision and confidence variables may lead to impaired insight, due to the model not being able to "recognize" when it may have been in error.

It is instructive to contrast the signal detection model underpinning meta-*d'* with the Bayesian framework outlined here. While meta-*d'* is primarily a tool for estimating metacognitive sensitivity, second-order computation provides an underlying generative model for confidence and an explanatory framework for different types of dissociation between performance and confidence. In addition, while confidence in the meta-*d'* model is specified in arbitrary units, second-order computation models decision confidence as a probability, thus allowing specification of parameters determining not only metacognitive sensitivity but also about the extent of over- or under-confidence (Drugowitsch et al., 2014; Fleming & Lau, 2014; Lichtenstein et al., 1982; Moore & Healy, 2008). It is therefore useful to view meta-*d'* as complementary to our framework. Just as *d'* provides a bias-free measure of perceptual sensitivity that may depend on a number of underlying processes, meta-*d'* provides a summary of an individual's metacognitive sensitivity that may be determined by the joint contribution of internal states and the computations applied to those states.

Multiple drivers of metacognitive sensitivity are also recognized by the stochastic detection and retrieval model (SDRM) of confidence in memory (Jang et al., 2012), which assumes that two samplings of evidence occur per stimulus, one leading to memory retrieval, and the other leading to a confidence rating. One important difference between second-order computation and the SDRM is that in the former, decision confidence is a probability of success derived from inverting a generative model of action, whereas in the latter, confidence is generated by comparing samples to additional criterion parameters. An intriguing consequence is that in the SDRM, an increase in ρ leads to increased metacognitive sensitivity, due to a tighter association between confidence and performance, whereas in second-order computation, an increase in ρ leads to a decrease in sensitivity, due to the model being unable to see itself in error (Figs. 3D and 8B). Empirical work combined with model comparison could test these predictions.

Our model accommodates dissociations between decision-making and metacognition through alterations in the precision and coupling of internal states, such as the decision and confidence variables. However it is also possible that decision-making and metacognition have different inferential goals, and may be differentially sensitive to different types of information. Introducing these normative constraints into models of metacognition is an important goal for future work. For instance, it would be of interest to explore whether differential sensitivity to evidence for or against a choice (Koizumi et al., 2015; Maniscalco et al., 2016; Zylberberg et al., 2012), and differential effects of attention on performance and confidence (Rahnev et al., 2011; Solovey et al., 2015) could be accommodated from a second-order perspective. The current framework may also provide a benchmark from which to assess other apparent suboptimalities in confidence that are normative when appropriate computational considerations are taken into account (e.g. the effects of actions on subsequent confidence ratings). Finally, we have shown that mismatches between the subject's beliefs (hyperparameters) about different sources of uncertainty and the true parameters can lead to systematic over- and underconfidence (Adams et al., 2013; Drugowitsch et al., 2014), and thus potentially account for variability across individuals in metacognitive bias. How such hyperparameters are learnt over time is an important topic for future investigation.

Influence of choices on confidence judgments

A counterintuitive feature of second-order computation is that actions influence subsequent confidence ratings, all else being equal. This influence arises because actions contribute information about possible world states, leading a rational observer to incorporate his own actions as additional data when computing confidence (cf. Bem, 1967). This feature of the model has several empirical implications. A practical implication is that it pays to be cautious when comparing data from studies in which confidence is elicited with or without a preceding action. Several behavioural paradigms have been developed for eliciting decision confidence in both humans and non-human animals (Kepecs & Mainen, 2012). In retrospective judgment paradigms, an action intervenes between the stimulus and the confidence rating whereas in opt-out and simultaneous-report paradigms, confidence is elicited in parallel to or instead of the decision itself. Measures of confidence from these paradigms are often taken to be equivalent. However the current model predicts subtle differences in the role played by actions in retrospective judgment designs where the subject's own responses may contribute additional evidence to the computation of confidence. While perhaps counterintuitive, this is rational under the model architecture: to the extent that the confidence and decision variable have partially distinct information, the subject may gain additional information about the world state by "observing" her own actions.

A second-order framework makes concrete predictions about the effect of choices on confidence ratings – namely a decrease in overall confidence and an increase in sensitivity. In addition to the results of Siedlecka et al. (2016) that we document in Figure 6, other recent findings are consistent with these predictions. Manipulating the order of identification responses and subjective awareness ratings (including confidence and visibility scales) revealed increases in metacognitive sensitivity when identification responses preceded the rating (Wierzchoń et al., 2014). Zehetleitner & Rausch (2013) similarly compared first-order subjective ratings of a stimulus with second-order confidence in a previous decision, and found that the latter was associated with greater metacognitive sensitivity. Finally, Kvam and colleagues compared a choice with a no-choice (arbitrary mouse click) condition in a random-dot motion discrimination task (Kvam et al., 2015). They found that confidence judgments were less extreme and more accurate in the choice compared to the no-choice condition (see also Ronis & Yates, 1987; Sniezek et al., 1990 for similar findings); however in this case effects of choice were modeled as interfering with a second stage of evidence accumulation, as sensory evidence continued to be available after the decision was made. Finally, in a recent study we tested for the influence of action-specific information on confidence in a near-threshold visual discrimination task by applying single-pulse TMS to the premotor cortex (Fleming et al., 2015). When stimulation was incongruent with the subjects' actions, confidence judgments on correct trials were decreased, whereas congruent stimulation led to increased confidence. Performance remained unchanged. This pattern is potentially consistent with a contribution of action information to second-order computation.

The role of action in a second-order framework also reveals subtleties in the relationship between confidence and visibility judgments. In consciousness studies, confidence ratings are often considered proxies for perceptual awareness (Pierce & Jastrow, 1885). For instance, King & Dehaene (2012) suggest that within a signal detection framework, visibility is equivalent to assessing confidence in a detection response, and their model is able to account for several classical characteristics of conscious and unconscious perception. However to the extent that subjects are applying second-order computation to assess their confidence in their response, we might observe that subjects leverage the information content of the response itself to inform their confidence ratings. For instance, blindsight patients with lesions to visual cortex may nevertheless develop a "hunch" that their response was correct, without acknowledging the existence of a corresponding visual conscious experience (Persaud et al., 2011). As described above, these effects may also lead to changes in visibility ratings following responses in psychophysics experiments in healthy observers (Wierzchon et al., 2014). More broadly, these considerations suggest that one should be careful in inferring perceptual awareness from confidence ratings about the observer's response, and alternative approaches for determining perceptual awareness may be preferred, such as forced-choice discrimination of stimulus visibility (Peters & Lau, 2015).

We note that there are certain cases in which one would *not* expect an influence of action on metacognitive judgments. For instance, if the confidence variable has access to the same information as the decision variable, then there is nothing more to learn from the identity of the action. This is the case in the post-decisional model shown in Figure 1B – the confidence variable is determined by the sum of pre- and post-decision evidence (equivalent to accumulating log-odds correct; Kiani & Shadlen, 2009), and the action provides no further information beyond that provided by the pre-decision evidence (formally, d is conditionally independent of a given X_{conf}). However, even in these cases of sequential evidence accumulation, effects of action may be obtained in practice. For instance, if the influence of pre-decision evidence decays over time, this would weaken the cross-talk between the decision and confidence variables, and actions would again carry weight when inferring the world state. In other words, if I make a perceptual decision based on some sensory evidence, but then go on to forget this evidence at a later point in time, I am left with only my decision when inferring what the world state might have been. Interestingly empirical data are potentially consistent with this prediction. Jazayeri and Movshon (2007) found that estimates of the direction of a random dot motion stimulus were biased in the direction of a previous binary choice. Such effects may be consistent with rational inference on possible world states in the face of imperfect integration or the inevitable decay of sensory evidence over time (Stocker & Simoncelli, 2008).

More broadly, the influence of one's own actions on self-evaluation dovetails with the proposal that preferences and beliefs are constructed rather than revealed by judgments and decisions (Lichtenstein & Slovic, 2006). Post-choice preference change occurs when subjects increase their estimate of the value of an object after choosing it, while simultaneously decreasing the values of rejected items (Brehm, 1956; Sharot, De Martino, & Dolan, 2009). Although this phenomenon is famously theorized to result from subjects' attempts to reduce cognitive dissonance, it can also be viewed in terms of rational inference in a model analogous to ours. Akin to perceptual categories, choice values are not perfectly known to the subject, but are probabilistic (De Martino et al., 2013; McFadden, 1980). To the extent that a subject's reports reflect posterior beliefs about the value of the items, it becomes rational to incorporate one's own actions if one has limited access to the decision variable underpinning choice, thereby leading to boosts in valuation after an object is chosen.

Neural implementation of metacognition

The models considered here suggest an organizing framework for nascent findings on the neural basis of confidence and self-evaluation. In particular, it predicts that correlates of confidence will be found across multiple putative internal states, including both those directly supporting actions and those supporting confidence ratings. Empirical studies in humans and non-human primates show that neural precursors of a decision are modulated by the eventual degree of confidence of the subject (Gherman & Philiastides, 2015; Kiani & Shadlen, 2009; Komura et al., 2013; Zizlsperger et al., 2014), and microstimulation of neurons encoding sensory evidence leads to biases in both choices and confidence ratings (Fetsch, Kiani, Newsome, & Shadlen, 2014). However, while confidence may covary with the activity of putative decision variables (Beck et al., 2008; Fiser, Berkes, Orbán, & Lengyel, 2010), the current framework predicts that metacognitive reports of confidence will critically depend on additional correlated states. Indeed, the mere fact that one brain area may “read-out” the decision variable from upstream neural populations may lead to a natural separation between decision and confidence variables. A study by Komura and colleagues is consistent with this proposal. In a motion discrimination task, the firing rate of pulvinar neurons correlated with the likelihood the monkey would choose an opt-out response. Inactivation of these neurons with muscimol led to an increase in opt-out responses without affecting first-order decision performance, as if the monkey lost confidence in its decision (Komura et al., 2013). This is potentially consistent with a confidence variable being encoded in cortico-thalamic loops (Kanai, Komura, Shipp, & Friston, 2015), and similar findings have been obtained through OFC inactivation in rodents (Lak et al., 2014).

A related line of work has identified a central role for the human prefrontal cortex (PFC) in metacognition (see Fleming & Dolan, 2012 for a review). Damage to the PFC leads to deficits in self-evaluation and impairments on a variety of tasks taxing metacognition (Pannu & Kaszniak, 2005; Schmitz & Johnson, 2007; Schnyer et al., 2004). Crucially these deficits may manifest in the absence of any changes in first-order performance: for instance, applying repetitive transcranial magnetic stimulation to the dorsolateral PFC in humans alters confidence but not performance in a visual discrimination task (Rounis et al., 2010), and patients with lesions to anterior sectors of the PFC show a reduced correspondence between confidence and accuracy (reduced type 2 ROC area) on a perceptual task despite performance remaining unaffected (Fleming et al., 2014). In addition, studies using functional imaging in humans and single-unit recording in non-human primates and rodents have identified correlates of confidence in prefrontal cortex and interconnected subcortical regions (De Martino et al., 2013; Fleming et al., 2012; Hebart, Schriever, Donner, & Haynes, 2014; Hilgenstock, Weiss, & Witte, 2014; Kepecs et al., 2008; Lak et al., 2014; Middlebrooks & Sommer, 2012). In relation to the current framework, these findings may be consistent with prefrontal involvement in representing a confidence variable and/or hyperparameters about sources of decision uncertainty (Lau, 2007), and/or in representing the output of a confidence computation for subsequent report (Fleming & Dolan, 2012).

Second-order computation requires integration of state information (e.g. X_{conf}) with knowledge about the selected action. Importantly this convergence should be flexible and domain-general[5]. Consider a task where auditory stimuli are arbitrarily mapped to eye movements, and visual stimuli to hand movements. To compute confidence in the model in Figure 1C one would need to combine information about each sensory modality with corollary discharge (or proprioceptive feedback) from the relevant motor system. One solution to this problem would be to maintain global representations of sensory evidence in a response-independent frame of reference (Heekeren, Marrett, Ruff, Bandettini, & Ungerleider, 2006; Ho, Brown, & Serences, 2009; O'Connell, Dockree, & Kelly, 2012; Tosoni, Galati, Romani, & Corbetta, 2008). The frontopolar cortex (FPC; Brodmann area 10) in primates is one potential convergence zone for integrating state and action information in the service of second-order computation. The FPC receives multimodal inputs from higher-order sensory and motor regions in the parietal, frontal and temporal lobes (Burman, Reser, Yu, & Rosa, 2011; Neubert, Mars, Thomas, Sallet, & Rushworth, 2014; Ramnani & Owen, 2004), and convergent evidence supports its role in human metacognition (Baird, Smallwood, Gorgolewski, & Margulies, 2013; De Martino et al., 2013; Del Cul et al., 2009; Fleming et al., 2010; 2012; 2014; Hilgenstock et al., 2014; McCurdy et al., 2013; Miele, Wager, Mitchell, & Metcalfe, 2011; Yokoyama et al., 2010). There is a paucity of single-neuron recording studies from the FPC. However, the one exception finds that FPC neurons code the response chosen by the monkey at the time of feedback in a decision task, but do so differentially depending on whether the response was correct or erroneous. Critically these signatures emerge before external feedback is given, potentially consistent with an evaluation of whether the action taken was appropriate (Tsujimoto, Genovesio, & Wise, 2010; 2011). Another candidate neural nexus for state-action integration is the dorsomedial prefrontal cortex (dmPFC; encompassing the paracingulate cortex and pre-supplementary motor area). Studies of

[5] Similarly Timmermans et al. (2012) point out that metacognition “requires that the first-order representations that are responsible for performance be accessed in a manner that is independent from their expression in behaviour” (p. 1416)

error detection observe increased activity in dmPFC when errors are made on simple choice reaction-time tasks in the absence of external feedback (Carter et al., 1998; Dehaene et al., 1994; Gehring et al., 1993), and the dmPFC is in turn interconnected with insula and FPC, suggesting a possible circuit for metacognitive evaluation (Baird et al., 2013; Hilgenstock et al., 2014).

Finally, the model of metacognition we outline here has much in common with schemes for recursive inference in social cognition (Goodman & Baker, 2009; Shafto, Goodman, & Frank, 2012). Confidence is formed through second-order evaluation of a coupled but distinct decision system, computationally equivalent to inferring the performance of another actor. While here we have focused on the implications of this framework for self-directed metacognition, to the extent that self- and other-evaluation rely on common mechanisms, brain networks previously linked to theory of mind (ToM) may also play a role in metacognition (Carruthers, 2009). Previous studies have identified similarities in neural activity for self- and other-judgments (Decety, 2003; C. D. Frith & Frith, 1999; Jenkins et al., 2008; Mitchell, Banaji, & Macrae, 2006) albeit with a focus on personal-level judgments about beliefs, attitudes or personality characteristics. It will be of interest to determine whether these ToM networks are additionally recruited when inferring subpersonal states such as one's confidence in percepts or memories.

Relationship between metacognitive monitoring and control

Computing confidence in a decision is a type of metacognitive monitoring, and may be distinct from processes supporting metacognitive control (Nelson & Narens, 1990). However, accurately inferring one's confidence in a task is important for the future control of behavior. For instance, a child studying for an exam will perform better if they have an accurate impression of how much there is still to learn (Veenman et al., 2004). In the absence of external feedback, such estimates may be furnished by second-order computation, which outputs a subjective probability of success. This probability provides a useful indicator of whether a previous decision should be corrected (Resulaj et al., 2009), whether a subsequent step in a chain of decisions should be initiated (Dehaene &

Sigman, 2012), whether to make the task easier by offloading intentions into the environment (Gilbert, 2015), or more generally when it is advantageous to deliberate (Keramati et al., 2011) or engage cognitive control (Boureau et al., 2015; Shenhav, Botvinick, & Cohen, 2013). Here we focus on the generation of confidence in a single task, but one could envisage replicating this architecture to maintain internal estimates of long-run confidence over a number of tasks (Donoso, Collins, & Koechlin, 2014). We would therefore predict a close relationship between metacognitive estimates of confidence and the strategic control of decision-making.

Metacognition and clinical insight

A common factor in a range of neurological and psychiatric disorders is a loss of insight (David et al., 2012) – the ability to recognize and describe one's own behaviour, cognition and mental states. For instance, a patient with addiction may not recognize a need for treatment due to impaired insight into his or her addictive behaviours (Goldstein et al., 2009), consistent with impairments of metacognitive sensitivity in this population (Moeller et al., 2016). Deficits in metacognitive sensitivity have also been documented in pathological gambling (Brevers et al., 2014) and brain injury (Ham et al., 2013; Fleming et al., 2014), and have been suggested to underpin a variety of impairments in schizophrenia, ADHD and anosagnosia (Klein et al., 2013). Second-order computation provides a possible framework within which to understand such deficits. For instance, loss of insight may correspond to a pathologically increased coupling between internal states, reducing the ability for error detection (Fig. 4D), a reduction in the precision of the confidence variable (Fig. 8A), aberrant beliefs (hyperparameters) about different sources of uncertainty (Fig. 9), or any combination of these factors. Actions would occur but the subject would have little knowledge of *why* they occurred, or whether they were appropriate for the current situation. Restoring insight in such cases may therefore be aided by a better understanding of the computational and neural basis of metacognition.

Limitations and future directions

We have focused on modeling a two-choice perceptual discrimination for computational simplicity. However, the key feature of the model is qualitative – second-order states are harnessed to infer confidence in first-order decisions. This holds promise for generalizing the framework to other domains, such as memory- or value-based choices. In addition, we have not considered the role of learning or prior beliefs about the task structure in constructing self-evaluations. For instance, expectations about possible world states ($P(d)$) should influence the computation of confidence (Sherman, Seth, Barrett, & Kanai, 2015). We have also not touched upon how subjects learn the model of the task in the first place (corresponding to reduction in uncertainty at the rule or strategy level, Bach & Dolan, 2012; Donoso et al., 2014) or learn beliefs (hyperparameters) about self-ability, but these are likely to be important for understanding the dynamics of self-evaluation over longer timescales. Moreover such learning is likely to be influenced by our interactions with other individuals, allowing coordination of confidence at the group level (Bahrami et al., 2012; Shea et al., 2014).

In many laboratory decision-making tasks (and in the simulations carried out here), actions are binary, such as a button press or eye movement. In practice however even simple actions are constructed by specifying the kinematics and forces needed to produce a particular motor output. Indeed, individuals have been shown to take action kinematics into consideration when judging the confidence of another individual (Patel et al., 2012), and the specifics of action planning impacts upon error-related brain activity (Bernstein, Scheffers, & Coles, 1995; Torrecillos, Albouy, Brochier, & Malfait, 2014). An interesting avenue for future investigation is the extent to which this richness of action specification is incorporated into decision confidence, and how this information is routed to metacognitive computations.

Finally, as touched upon above, our model is situated at the computational level (Marr, 1982), and remains agnostic about algorithmic or mechanistic implementation. Future efforts could harness our framework to guide construction of finer-grained Bayesian models incorporating temporal dynamics or candidate neural network implementations

(Fiser et al., 2010; Insabato et al., 2010; Ma & Jazayeri, 2014; Pasquali et al., 2010; Rao, 2004).

Conclusions

The model outlined in this paper casts self-evaluation as a second-order inference on the efficacy of one's own behaviour. Such a model has the potential to provide common ground for comparing data from different paradigms such as confidence and error detection, and provides a normative framework for understanding a range of dissociations between metacognition and performance. In addition, it predicts a novel role for actions in contributing to estimates of decision confidence. We have outlined the implications of second-order computation for behavioural control and for candidate neurobiological implementations of metacognition. We hope this will provide a conceptual and theoretical framework for studies of metacognitive computation, and motivate a number of empirical hypotheses to be tested in future research.

APPENDIX A

Derivation of second-order confidence

The second-order model posits that the decision and confidence variables are draws from a multivariate Gaussian with covariance matrix Σ:

$$\begin{bmatrix} X_{act} \\ X_{conf} \end{bmatrix} \sim N(\boldsymbol{d}, \Sigma)$$

$$\Sigma = \begin{bmatrix} \sigma_{act}^2 & \rho\sigma_{act}\sigma_{conf} \\ \rho\sigma_{act}\sigma_{conf} & \sigma_{conf}^2 \end{bmatrix}$$

To compute confidence, *z*, the observer infers (for the purpose of marginalizing) the state of the decision variable driving choice (X_{act}) from the confidence variable (X_{conf}).

$$z = P\left(a = d \middle| X_{conf}, a, \Sigma\right) = \begin{cases} P\left(d = 1 \middle| X_{conf}, a, \Sigma\right) \text{ if } a = 1 \\ 1 - P\left(d = 1 \middle| X_{conf}, a, \Sigma\right) \text{ if } a = -1 \end{cases}$$

In the following we unpack computation of $P\left(d \middle| X_{conf}, a, \Sigma\right)$, suppressing covariance parameters Σ for clarity. As in the first-order model, confidence depends on the posterior over *d* computed using Bayes rule:

$$P\left(d \middle| X_{conf}, a\right) = \frac{P(d|X_{conf})P\left(a \middle| X_{conf}, d\right)}{\sum_d P\left(d \middle| X_{conf}\right)P\left(a \middle| X_{conf}, d\right)}$$

Starting with the second term, $P\left(a \middle| X_{conf}, d\right)$:

$$P\left(a \middle| X_{conf}, d\right) = \int P(a|X_{act})P(X_{act}|X_{conf}, d)\, dX_{act}$$

$$= \int_0^\infty P\left(X_{act} \middle| X_{conf}, d\right) dX_{act} \text{ if } a = 1$$

$$= \int_{-\infty}^0 P\left(X_{act} \middle| X_{conf}, d\right)\, dX_{act} \text{ if } a = -1$$

where the latter two expressions are due to the threshold response rule, $a = 1$ whenever $X_{act} > 0$. This expression is a cumulative density function of the conditional density of a multivariate Gaussian, which itself is a univariate Gaussian with the following mean and standard deviation:

$P(X_{act}|X_{conf}, d) \sim N(\mu_{X_{act}|X_{conf}}, \sigma_{X_{act}|X_{conf}})$

where $\mu_{X_{act}|X_{conf}} = d + \frac{\sigma_{act}}{\sigma_{conf}}\rho(X_{conf} - d)$

$\sigma_{X_{act}|X_{conf}} = \sqrt{(1 - \rho^2)\sigma_{act}^2}$

The first term is the normalized likelihood of X_{conf} given *d*:

$$P(d|X_{conf}) = \frac{P(X_{conf}|d)}{\sum_d P(X_{conf}|d)}$$

i.e. Bayes' rule with the uniform prior $P(d)$ canceled, where

$P(X_{conf}|d) = \phi(X_{conf}, d, \sigma_{conf})$

and $\phi()$ is the standard Gaussian density function:

$\phi(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

APPENDIX B

Simulation details

Internal representations supporting decision confidence

To produce the plots in Figure 3 we simulated 10,000 trials at each of 7 levels of stimulus strength $\theta = [0\ 0.032\ 0.064\ 0.128\ 0.256\ 0.512\ 1.0]$. For the first-order and post-decisional models $\sigma = 1$. For the second-order model, parameter settings were $\sigma_{act} = 1$,

$\sigma_{conf} = 0.5$ and $\rho = 0.4$. Confidence was sorted according to whether the model's response was correct or incorrect. For all models we also binned confidence into tertiles of the unsigned confidence variable $|X_{conf}|$.

Internal representations supporting error monitoring

For each cell of the second-order model parameter grid in Figure 4D we simulated 10,000 trials and recorded the proportion of errors that were detected (errors with confidence levels of less than 0.5). σ_{act} (and therefore the the objective error rate) was kept constant.

Influence of actions on confidence

To examine the effects of actions on subsequent ratings we simulated two conditions, "rate-choose" and "choose-rate" for both the first- and second-order models. Confidence in the rate-choose condition was defined as the posterior probability of a future decision being correct (the max over possible actions; Kvam et al., 2015):

$$\text{confidence} = \max\left[\, P\left(d = 1 \middle| X_{conf}\right) P\left(d = -1 \middle| X_{conf}\right)\right]$$

To create Figures 6B and C we simulated 10,000 trials at each of 7 levels of stimulus strength $\theta = [0\ 0.032\ 0.064\ 0.128\ 0.256\ 0.512\ 1.0]$ with $\sigma_{act} = 1$, $\sigma_{conf} = 1$ and $\rho = 0.6$. To determine the choice-dependence of bias and metacognitive sensitivity on second-order model parameters we simulated 10,000 trials at a single level of stimulus strength $\theta = 1$ while varying ρ and σ_{conf}. σ_{act} was fixed at 1, ensuring constant performance. ρ varied across 10 levels equally spaced between 0.1 and 0.9 while keeping σ_{conf} fixed at 1; σ_{conf}varied across 10 levels equally spaced between 0.5 and 1.5 while keeping ρ fixed at 0.6. Bias was calculated as the mean confidence level collapsing across correct and error trials; metacognitive sensitivity was calculated as the difference between mean confidence on correct and incorrect trials.

Modeling dissociations between performance and confidence

Type 2 ROCs were plotted by sweeping confidence criteria across 20 evenly spaced steps from 0 to 1 and calculating type 2 hit rates (the proportion of high confidence trials when the subject is correct) and false alarm rates (the proportion of high confidence trials when the subject is incorrect) (see Fleming & Lau, 2014; Galvin et al., 2003 for further details). 10,000 trials were simulated at each parameter setting.

To construct Figures 8C and D, 100 datasets were simulated each containing 1000 trials. σ_{conf} and σ_{act} were each generated from independent uniform random draws on the interval [1.5 2.5]. For both simulated and empirical datasets, meta-d' was fit using maximum likelihood methods instantiated in the code provided by Maniscalco & Lau (www.columbia.edu/~bsm2105/type2sdt/).

The datasets contributing to Figure 8E have been published in full elsewhere (Fleming et al., 2010; 2012; E. C. Palmer et al., 2014; L. G. Weil et al., 2013). Briefly, each study administered a perceptual decision task with trial-by-trial confidence ratings elicited post-decision on an arbitrary numerical scale ranging from 1 to 6. The number of trials available for analysis ranged from 250 to 500 per subject. In all studies, task difficulty was controlled by a one-up two-down staircase that targeted a performance level of approximately 71% correct. Three of the four studies employed a 2-interval forced choice detection task in which subjects were asked to report which interval contained a pop-out Gabor patch (Fleming et al., 2010; E. C. Palmer et al., 2014; L. G. Weil et al., 2013); one study employed a face/house discrimination task (Fleming et al., 2012).

REFERENCES

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, *4*, 47.

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making. *PLoS Computational Biology*, *11*(10), e1004519.

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*(10), 1338–1344.

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. *Nature Reviews Neuroscience*, *13*(8), 572–586.

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). Together, slowly but surely: The role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 3–8.

Baird, B., Cieslak, M., Smallwood, J., Grafton, S. T., & Schooler, J. W. (2015). Regional white matter variation associated with domain-specific metacognitive accuracy. *Journal of Cognitive Neuroscience*, *27*(3), 440–452.

Baird, B., Smallwood, J., Gorgolewski, K. J., & Margulies, D. S. (2013). Medial and Lateral Networks in Anterior Prefrontal Cortex Support Metacognitive Ability for Memory and Perception. *Journal of Neuroscience*, *33*(42), 16657–16665.

Baranski, J. V., & Petrusic, W. M. (2001). Testing architectures of the decision-confidence relation. *Canadian Journal of Experimental Psychology*, *55*(3), 195–206.

Baranski, J., & Petrusic, W. (1998). Probing the Locus of Confidence Judgments: Experiments on the Time to Determine Confidence* 1. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 929–945.

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*(4), 535–552.

Barttfeld, P., Wicker, B., McAleer, P., Belin, P., Cojan, Y., Graziano, M., et al. (2013). Distinct patterns of functional brain connectivity correlate with objective performance and subjective beliefs. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(28), 11577–11582.

Barthelmé, S., & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. *Proceedings of the National Academy of Sciences*, *107*(48), 20834–20839.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., et al. (2008). Probabilistic population codes for Bayesian decision making. *Neuron*, *60*(6), 1142–1152.

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*(3), 183–200.

Beran, M. J., Brandl, J., Perner, J., & Proust, J. (2012). Foundations of Metacognition. Oxford University Press.

Bernstein, P. S., Scheffers, M. K., & Coles, M. G. H. (1995). “Where did I go wrong?” A psychophysiological analysis of error detection. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1312–1322.

Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *Journal of Neuroscience*, *35*(8), 3478–3484.

Bona, S., & Silvanto, J. (2014). Accuracy and confidence of visual short-term memory do not go hand-in-hand: behavioral and neural dissociations. *PLoS ONE*, *9*(3), e90808.

Boureau, Y.-L., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding How To Decide: Self-Control and Meta-Decision Making. *Trends in Cognitive Sciences*, *19*(11), 700–710.

Brehm, J. W. (1956). Postdecision changes in the desirability of alternatives. *The Journal of Abnormal and Social Psychology*, *52*(3), 384–389.

Brevers, D., Cleeremans, A., Bechara, A., Greisen, M., Kornreich, C., Verbanck, P., & Noël, X. (2014). Impaired metacognitive capacities in individuals with problem gambling. *Journal of Gambling Studies*, *30*(1), 141–152.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, *12*(12), 4745–4765.

Bronfman, Z. Z., Brezis, N., Moran, R., Tsetsos, K., Donner, T., & Usher, M. (2015). Decisions reduce sensitivity to subsequent information. *Proceedings of the Royal Society of London Series B-Biological Sciences*, *282*(1810), 20150228.

Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, *307*(5712), 1118–1121.

Burman, K. J., Reser, D. H., Yu, H.-H., & Rosa, M. G. P. (2011). Cortical input to the frontal pole of the marmoset monkey. *Cerebral Cortex (New York, N.Y.: 1991)*, *21*(8), 1712–1737.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences, 32*(02), 121-138.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, *280*(5364), 747–749.

Cartwright, D., & Festinger, L. (1943). A quantitative theory of decision. *Psychological Review*, *50*(6), 595–621.

Charles, L., King, J.-R., & Dehaene, S. (2014). Decoding the dynamics of action, intention, and error detection for conscious and subliminal stimuli. *Journal of Neuroscience*, *34*(4), 1158–1170.

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *NeuroImage*, *73*, 80–94.

Clarke, F., Birdsall, T., & Tanner, W. (1959). Two types of ROC curves and definition of parameters. *Journal of the Acoustical Society of America*, *31*, 629–630.

Coles, M. G. H., Scheffers, M. K., & Holroyd, C. B. (2001). Why is there an ERN/Ne on correct trials? Response representations, stimulus-related components, and the theory of error-processing. *Biological Psychology*, *56*(3), 173–189.

David, A. S., Bedford, N., Wiffen, B., & Gilleen, J. (2012). Failures of metacognition and lack of insight in neuropsychiatric disorders. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1379–1390.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110. http://doi.org/10.1038/nn.3279

Decety, J. (2003). Shared representations between self and other: a social cognitive neuroscience view. *Trends in Cognitive Sciences*, *7*(12), 527–533.

Dehaene, S., & Sigman, M. (2012). From a single decision to a multi-step algorithm. *Current Opinion in Neurobiology*, *22*(6), 937–945.

Dehaene, S., Posner, M. I., & Tucker, D. M. (1994). Localization of a neural system for error detection and compensation. *Psychological Science*, *5*(5), 303–305.

Del Cul, A., Dehaene, S., Reyes, P., Bravo, E., & Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain*, *132*(9), 2531.

Donoso, M., Collins, A. G. E., & Koechlin, E. (2014). Human cognition. Foundations of human reasoning in the prefrontal cortex. *Science*, *344*(6191), 1481–1486.

Drugowitsch, J., Moreno-Bote, R., & Pouget, A. (2014). Relation between Belief and Performance in Perceptual Decision Making. *PLoS ONE*, *9*(5), e96511.

Faisal, A. A., & Wolpert, D. M. (2009). Near optimal combination of sensory and motor uncertainty in time during a naturalistic perception-action task. *Journal of Neurophysiology*, *101*(4), 1901–1912.

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance*, *26*(1), 32–53.

Fetsch, C. R., Kiani, R., & Shadlen, M. N. (2015). Predicting the Accuracy of a Decision: A Neural Mechanism of Confidence. *Cold Spring Harbor Symposia on Quantitative Biology*, *79*, 024893–197.

Fetsch, C. R., Kiani, R., Newsome, W. T., & Shadlen, M. N. (2014). Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron*, *83*(4), 797–804.

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, *14*(3), 119–130.

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1338–1349.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443.

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal Contributions to Metacognition in Perceptual Decision Making. *Journal of Neuroscience*, *32*(18), 6117–6125.

Fleming, S. M., Maloney, L. T., & Daw, N. D. (2013). The irrationality of categorical perception. *Journal of Neuroscience*, *33*(49), 19060–19070.

Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. C. (2015). Action-specific disruption of perceptual confidence. *Psychological Science*, *26*(1), 89–98.

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, *137*(Pt 10), 2811–2822.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–1543.

Frith, C. D., & Frith, U. (1999). Interacting minds--a biological basis. *Science*, *286*(5445), 1692–1695. http://doi.org/10.1126/science.286.5445.1692

Fuss, I. G., & Navarro, D. J. (2013). Open parallel cooperative and competitive decision processes: a potential provenance for quantum probability decision models. *Topics in Cognitive Science*, *5*(4), 818–843.

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. *Psychonomic Bulletin & Review*, *10*(4), 843–876.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*(6), 385.

Gherman, S., & Philiastides, M. G. (2015). Neural representations of confidence emerge from the process of decision formation during perceptual choices. *NeuroImage*, *106*, 134–143.

Gilbert, S. J. (2015). Strategic use of reminders: Influence of both domain-general and task-specific metacognitive confidence, independent of objective memory ability. *Consciousness and Cognition*, *33*, 245–260.

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, *36*(2), 299–308.

Goldstein, R. Z., Craig, A. D. B., Bechara, A., Garavan, H., Childress, A. R., Paulus, M. P., & Volkow, N. D. (2009). The neurocircuitry of impaired insight in drug addiction. *Trends in Cognitive Sciences*, *13*(9), 372–380.

Goodman, N. D., & Baker, C. L. (2009). Cause and intent: Social reasoning in causal learning. Presented at the Proceedings of the 31st annual conference of the Cognitive Science Society.

Graziano, M., & Sigman, M. (2009). The Spatial and Temporal Construction of Confidence in the Visual Scene. *PLoS ONE*, *4*(3), e4909.

Green, D., & Swets, J. (1966). Signal detection theory and psychophysics. New York: Wiley.

Ham, T. E., Bonnelle, V., Hellyer, P., Jilka, S., Robertson, I. H., Leech, R., & Sharp, D. J. (2014). The neural basis of impaired self-awareness after traumatic brain injury. *Brain*, *137*(Pt 2), 586–597.

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences*, *1*(2), 78–82.

Heatherton, T. F. (2011). Neuroscience of self and self-regulation. *Annual Review of Psychology*, *62*, 363–390.

Hebart, M. N., Schriever, Y., Donner, T. H., & Haynes, J.-D. (2016). The Relationship between Perceptual Decision Variables and Confidence in the Human Brain. *Cerebral Cortex*, *26*(1), 118–130.

Heekeren, H., Marrett, S., Ruff, D., Bandettini, P., & Ungerleider, L. (2006). Involvement of human left dorsolateral prefrontal cortex in perceptual decision making is independent of response modality. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(26), 10023–10028.

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, *18*(3), 186–201.

Hilgenstock, R., Weiss, T., & Witte, O. W. (2014). You'd better think twice: post-decision perceptual confidence. *NeuroImage*, *99*, 323–331.

Ho, T. C., Brown, S., & Serences, J. T. (2009). Domain General Mechanisms of Perceptual Decision Making in Human Cortex. *Journal of Neuroscience*, *29*(27), 8675–8687.

Holroyd, C. B., & Coles, M. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–708.

Holroyd, C. B., Yeung, N., Coles, M. G. H., & Cohen, J. D. (2005). A mechanism for error detection in speeded response time tasks. *Journal of Experimental Psychology. General*, *134*(2), 163–191.

Insabato, A., Pannunzi, M., Rolls, E. T., & Deco, G. (2010). Confidence-Related Decision Making. *Journal of Neurophysiology*, *104*(1), 539–547.

James, W. (1950). The Principles of Psychology, Vol. 1. Dover Publications.

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review*, *119*(1), 186.

Jazayeri, M., & Movshon, J. A. (2007). A new perceptual illusion reveals mechanisms of sensory decoding. *Nature*, *446*(7138), 912–915.

Jenkins, A. C., Jenkins, A. C., Macrae, C. N., Macrae, C. N., Mitchell, J. P., & Mitchell, J. P. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proceedings of the National Academy of Sciences*, *105*(11), 4507–4512.

Kanai, R., Komura, Y., Shipp, S., & Friston, K. (2015). Cerebral hierarchies: predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*(1668), 20140169–20140169.

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1322–1337.

Kepecs, A., Uchida, N., Zariwala, & Mainen. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, *455*(7210), 227–231.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, *7*(5), e1002055.

Kiani, R., & Shadlen, M. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron*, *84*(6), 1329–1342.

King, J.-R., & Dehaene, S. (2014). A model of subjective report and objective discrimination as categorical decisions in a vast representational space., *369*(1641), 20130204.

Klein, T. A., Ullsperger, M., & Danielmeier, C. (2013). Error awareness and the insula: links to neurological and psychiatric diseases. *Frontiers in Human Neuroscience*, *7*, 14.

Ko, Y., & Lau, H. C. (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*(1594), 1401–1411.

Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T., & Miyamoto, A. (2013). Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nature Neuroscience*, *16*(6), 749–755.

Koizumi, A., Maniscalco, B., & Lau, H. C. (2015). Does perceptual confidence facilitate cognitive control? *Attention, Perception & Psychophysics*, *77*(4), 1295–1306.

Kvam, P. D., Pleskac, T. J., Yu, S., & Busemeyer, J. R. (2015). Interference effects of choice on confidence: Quantum characteristics of evidence accumulation. *Proceedings of the National Academy of Sciences*, *112*(34), 10645–10650.

Lak, A., Costa, G. M., Romberg, E., Koulakov, A. A., Mainen, Z. F., & Kepecs, A. (2014). Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron*, *84*(1), 190–201.

Lau, H. C. (2007). A higher order Bayesian decision theory of consciousness. *Progress in Brain Research*, *168*, 35–48.

Lau, H. C., & Passingham, R. E. (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*(49), 18763–18768.

Lau, H. C., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, 1–9.

Lichtenstein, S., & Slovic, P. (2006). The Construction of Preference. Cambridge University Press.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, *40*(1), 77–105.

Ma, W. J., & Jazayeri, M. (2014). Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, *37*(1), 205–220.

Macmillan, N., & Creelman, C. (2005). Detection theory: a user's guide. New York: Lawrence Erlbaum.

Maniscalco, B., & Lau, H. C. (2016). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness*, *2016*(1), niw002.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430.

Maniscalco, B., & Lau, H. (2014). Signal Detection Theory Analysis of Type 1 and Type 2 Data: Meta-d", Response-Specific Meta-d," and the Unequal Variance SDT Model. In S. M. Fleming & C. D. Frith (Eds.), *The Cognitive Neuroscience of Metacognition*. Springer.

Marr, D. (1982). Vision. New York: WH Freeman.

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., de Lange, F. P., & Lau, H. C. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *Journal of Neuroscience*, *33*(5), 1897–1906.

McFadden, D. (1980). Econometric models for probabilistic choice among products. *Journal of Business*, S13–S29.

Merkle, E. C., & Van Zandt, T. (2006). An application of the poisson race model to confidence calibration. *Journal of Experimental Psychology. General*, *135*(3), 391–408.

Metcalfe, J. (1996). Metacognition: Knowing About Knowing. MIT Press.

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian Probability:

From Neural Origins to Behavior. *Neuron*, *88*(1), 78–92.

Meyniel, F., Schlunegger, D., & Dehaene, S. (2015). The Sense of Confidence during Probabilistic Learning: A Normative Account. *PLoS Computational Biology*, *11*(6), e1004305.

Middlebrooks, P. G., & Sommer, M. A. (2012). Neuronal correlates of metacognition in primate frontal cortex. *Neuron*, *75*(3), 517–530.

Miele, D. B., Wager, T. D., Mitchell, J. P., & Metcalfe, J. (2011). Dissociating neural correlates of action monitoring and metacognition of agency. *Journal of Cognitive Neuroscience*, *23*(11), 3620–3636.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2006). The Link between Social Cognition and Self-referential Thought in the Medial Prefrontal Cortex. *Dx.Doi.org*, *17*(8), 1306–1315.

Moeller, S. J., Fleming, S. M., Gan, G., Zilverstand, A., Malaker, P., d Oleire Uquillas, F., et al. (2016). Metacognitive impairment in active cocaine use disorder is associated with individual differences in brain structure. *European Neuropsychopharmacology*, *26*(4), 653–662.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147.

Moreno-Bote, R. (2010). Decision confidence and uncertainty in diffusion models with partially correlated neuronal integrators. *Neural Computation*, *22*(7), 1786–1811.

Navajas, J., Bahrami, B. & Latham, P. E. (2016). Post-decisional accounts of biases in confidence. *Current Opinion in Behavioral Sciences, 11,* 55-60.

Navarro-Cebrian, A., Knight, R. T., & Kayser, A. S. (2013). Error-monitoring and post-error compensations: dissociation between perceptual failures and motor errors with and without awareness. *Journal of Neuroscience*, *33*(30), 12375–12383.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation: Advances in Research and Theory*, *26*, 125–173.

Neubert, F.-X., Mars, R. B., Thomas, A. G., Sallet, J., & Rushworth, M. F. S. (2014). Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex. *Neuron*, *81*(3), 700–713.

Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A*, *231*, 289–337.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behaviour. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation Advances in research and theory* (pp. 1–18).

O'Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, *15*(12), 1729–1735.

Palmer, E. C., David, A. S., & Fleming, S. M. (2014). Effects of age on metacognitive efficiency. *Consciousness and Cognition*, *28*, 151–160.

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*(5), 1–1.

Pannu, J., & Kaszniak, A. (2005). Metamemory experiments in neurological populations: A review. *Neuropsychology Review*, *15*(3), 105–130.

Pasquali, A., Timmermans, B., & Cleeremans, A. (2010). Know thyself: metacognitive networks and measures of consciousness. *Cognition*, *117*(2), 182–190.

Patel, D., Fleming, S. M., & Kilner, J. M. (2012). Inferring subjective states through the observation of actions. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1748), 4853–4860.

Peirce, C. S., & Jastrow, J. (1885). On small differences in sensation. *Memoirs of the National Acadmey of Sciences*, *3*, 73–83.

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature Neuroscience*, *10*(2), 257–261.

Persaud, N., Davidson, M., Maniscalco, B., Mobbs, D., Passingham, R. E., Cowey, A., & Lau, H. C. (2011). Awareness-related activity in prefrontal and parietal cortices in blindsight reflects more than superior visual performance. *NeuroImage*, *58*(2), 605–611.

Peters, M. A. K., & Lau, H. C. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *eLife*, *4*, e09651.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374.

Rabbitt, P. M. A. (1966). Error Correction Time without External Error Signals. *Nature*, *212*(5060), 438. http://doi.org/10.1038/212438a0

Rabbitt, P., & Rodgers, B. (1977). What does a man do after he makes an error? An analysis of response programming. *The Quarterly Journal of Experimental Psychology*, *29*(4), 727–743.

Rabbitt, P., & Vyas, S. (1981). Processing a display even after you make a response to it. How perceptual errors can be corrected. *The Quarterly Journal of Experimental ….*

Rahnev, D., Maniscalco, B., Graves, T., Huang, E., de Lange, F. P., & Lau, H. C. (2011). Attention induces conservative subjective biases in visual perception. *Nature Neuroscience*, *14*(12), 1513–1515.

Ramnani, N., & Owen, A. M. (2004). Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience*, *5*(3), 184–194.

Rao, R. P. (2004). Hierarchical Bayesian inference in networks of spiking neurons. *Advances in Neural Information Processing Systems*.

Ratcliff, R. (1978). A Theory of Memory Retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*(1), 59–83.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263–266.

Ro, T., Shelton, D., Lee, O. L., & Chang, E. (2004). Extrageniculate mediation of unconscious vision in transcranial magnetic stimulation-induced blindsight. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(26), 9933–9935.

Roitman, J., & Shadlen, M. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience*, *22*(21), 9475.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, *40*(2), 193–218.

Rounis, E., Maniscalco, B., Rothwell, J., Passingham, R., & Lau, H. C. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience*, *1*(3), 165–175.

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron*, *90*(3), 499–506.

Scheffers, M. K., & Coles, M. G. H. (2000). Performance monitoring in a confusing world: Error-related brain activity, judgments of response accuracy, and types of errors. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(1), 141–151.

Schmid, M. C., Mrowka, S. W., Turchi, J., Saunders, R. C., Wilke, M., Peters, A. J., et al. (2010). Blindsight depends on the lateral geniculate nucleus. *Nature*, *466*(7304), 373–377.

Schmitz, T. W., & Johnson, S. C. (2007). Relevance to self: A brief review and framework of neural systems underlying appraisal. *Neuroscience and Biobehavioral Reviews*, *31*(4), 585–596.

Schnyer, D. M., Verfaellie, M., Alexander, M. P., LaFleche, G., Nicholls, L., & Kaszniak, A. W. (2004). A role for right medial prefontal cortex in accurate feeling-of-knowing judgements: evidence from patients with lesions to frontal cortex. *Neuropsychologia*, *42*(7), 957–966.

Scott, R. B., Dienes, Z., Barrett, A. B., Bor, D., & Seth, A. K. (2014). Blind Insight Metacognitive Discrimination Despite Chance Task Performance. *Psychological Science*, *25*(12), 2199–2208.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning From Others The Consequences of Psychological Reasoning for Human Learning. *Perspectives on Psychological Science*, *7*(4), 341–351. http://doi.org/10.1177/1745691612448481

Sharot, T., De Martino, B., & Dolan, R. J. (2009). How Choice Reveals and Shapes Expected Hedonic Outcome. *Journal of Neuroscience*, *29*(12), 3760–3765.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Sherman, M. T., Seth, A. K., Barrett, A. B., & Kanai, R. (2015). Prior expectations facilitate metacognition for perceptual decision. *Consciousness and Cognition*, *35*, 53–65.

Shimamura, A. P. (2000). Toward a Cognitive Neuroscience of Metacognition. *Consciousness and Cognition*, *9*(2), 313–323.

Siedlecka, M., Paulewicz, B., & Wierzchoń, M. (2016). But I Was So Sure! Metacognitive Judgments Are Less Accurate Given Prospectively than Retrospectively. *Frontiers in Psychology*, *7*(240), 218.

Sniezek, J. A., Paese, P. W., & Switzer, F. S. (1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes*, *46*(2), 264–282.

Solovey, G., Graney, G. G., & Lau, H. C. (2015). A decisional account of subjective inflation of visual perception at the periphery. *Attention, Perception & Psychophysics*, *77*(1), 258–271.

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition*, *20*(4), 1787–1792.

Steinhauser, M., & Yeung, N. (2010). Decision processes in human performance monitoring. *Journal of Neuroscience*, *30*(46), 15643–15653.

Stocker, A., & Simoncelli, E. P. (2008). A Bayesian model of conditioned perception. In *Advances in neural information processing systems* (pp. 1409-1416).

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard–easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes, 67*(2), 201–221.

Timmermans, B., Schilbach, L., Pasquali, A., & Cleeremans, A. (2012). Higher order thoughts in action: consciousness as an unconscious re-description process. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1412–1423.

Torrecillos, F., Albouy, P., Brochier, T., & Malfait, N. (2014). Does the processing of sensory and reward-prediction errors involve common neural resources? Evidence from a frontocentral negative potential modulated by movement execution errors. *Journal of Neuroscience*, *34*(14), 4845–4856.

Tosoni, A., Galati, G., Romani, G. L., & Corbetta, M. (2008). Sensory-motor mechanisms in human parietal cortex underlie arbitrary visual decisions. *Nature Neuroscience*, *11*(12), 1446–1453.

Treisman, M., & Faulkner, A. (1984). The setting and maintenance of criteria representing levels of confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(1), 119–139.

Tsujimoto, S., Genovesio, A., & Wise, S. P. (2010). Evaluating self-generated decisions in frontal pole cortex of monkeys. *Nature Neuroscience*, *13*(1), 120–126.

Tsujimoto, S., Genovesio, A., & Wise, S. P. (2011). Frontal pole cortex: encoding ends at the end of the endbrain. *Trends in Cognitive Sciences*, *15*(4), 169–176.

van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife*, *5*, e12192.

Veenman, M., Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction*, *14*(1), 89–109.

Vickers, D. (1979). Decision processes in visual perception. New York: Academic Press.

Vlassova, A., Donkin, C., & Pearson, J. (2014). Unconscious information changes decision accuracy but not confidence. *Proceedings of the National Academy of Sciences of the United States of America*, 201403619.

Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., et al. (2013). The development of metacognitive ability in adolescence. *Consciousness and Cognition*, *22*(1), 264–271.

Weiskrantz, L. (1998). Consciousness and commentaries. *International Journal of*

Psychology, *33*(3), 227–233.

Weiskrantz, L., Warrington, E. K., Sanders, M. D., & Marshall, J. (1974). Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain*, *97*(4), 709–728.

Wierzchoń, M., Paulewicz, B., Asanowicz, D., Timmermans, B., & Cleeremans, A. (2014). Different subjective awareness measures demonstrate the influence of visual identification on perceptual awareness ratings. *Consciousness and Cognition*, *27*, 109–120.

Wilimzig, C., Tsuchiya, N., Fahle, M., Einhäuser, W., & Koch, C. (2008). Spatial attention increases performance but not subjective confidence in a discrimination task. *Journal of Vision*, *8*(5), 7.1–10.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1310–1321.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The Neural Basis of Error Detection: Conflict Monitoring and the Error-Related Negativity. *Psychological Review*, *111*(4), 931–959.

Yokoyama, O., Miura, N., Watanabe, J., Takemoto, A., Uchida, S., Sugiura, M., et al. (2010). Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in short-term recognition memory performance. *Neuroscience Research*, *68*(3), 199–206.

Yu, S., Pleskac, T. J., & Zeigenfuse, M. D. (2015). Dynamics of postdecisional processing of confidence. *Journal of Experimental Psychology. General*, *144*(2), 489–510.

Zandbelt, B., Purcell, B. A., Palmeri, T. J., Logan, G. D., & Schall, J. D. (2014). Response times from ensembles of accumulators. *Proceedings of the National Academy of Sciences*, *111*(7), 2848–2853.

Zehetleitner, M., & Rausch, M. (2013). Being confident without seeing: What subjective measures of visual consciousness are about. *Attention, Perception & Psychophysics*, *75*(7), 1406–1426.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, *6*, 1.

Zizlsperger, L., Sauvigny, T., Händel, B., & Haarmeier, T. (2014). Cortical representations of confidence in a visual perceptual decision. *Nature Communications*, *5*, 3940.

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*.