CellPress

Neuron
Perspective

Confidence as Bayesian Probability: From Neural Origins to Behavior

Florent Meyniel,[1,*] Mariano Sigman,[2] and Zachary F. Mainen[3]
[1]Cognitive Neuroimaging Unit, CEA DSV/I2BM, INSERM, Université Paris-Sud, Université Paris-Saclay, NeuroSpin center, F-91191, Gif sur Yvette Cedex, France
[2]Departamento de Física, FCEN, UBA and IFIBA; Universidad Torcuato Di Tella, Buenos Aires, 1428 CABA, Argentina
[3]Champalimaud Neuroscience Programme, Champalimaud Centre for the Unknown, 1400-038, Lisbon, Portugal
*Correspondence: florent.meyniel@gmail.com
http://dx.doi.org/10.1016/j.neuron.2015.09.039

Research on confidence spreads across several sub-fields of psychology and neuroscience. Here, we explore how a definition of confidence as Bayesian probability can unify these viewpoints. This computational view entails that there are distinct forms in which confidence is represented and used in the brain, including distributional confidence, pertaining to neural representations of probability distributions, and summary confidence, pertaining to scalar summaries of those distributions. Summary confidence is, normatively, derived or "read out" from distributional confidence. Neural implementations of readout will trade off optimality versus flexibility of routing across brain systems, allowing confidence to serve diverse cognitive functions.

The sense of confidence has been defined as "a belief about the validity of our own thoughts, knowledge or performance that relies on a subjective feeling" (Grimaldi et al., 2015). This psychological definition would not seem out of place in the late 19th century, when psychologists began to ask human subjects about their confidence to unravel the determinants of this feeling (Peirce and Jastrow, 1884). Relatively recently, comparative psychology opened the study of confidence to non-human animals (for a review, see Smith et al., 2003) and neuroscience began to probe the electrophysiological underpinnings of confidence in monkeys and rodents (Hampton, 2001; Kepecs et al., 2008; Kiani and Shadlen, 2009). The translation of confidence from psychology to neuroscience has revealed underlying instabilities within the conceptual foundations of the still nascent area of confidence studies. Psychological definitions, such as that above, rely on concepts like "belief," "feelings," and "thought" that from a neuroscientific perspective pose unanswered translational challenges in themselves. Neuroscience definitions tend toward the notion that brains represent and process information using probabilistic codes at the level of populations of cells; their relationship to the psychological definition has been unclear. We hold that the study of confidence would benefit from a more unified framework that can provide more solid bridges between psychology and neuroscience and between research in humans and in other animals. Toward that end, in this review, we propose a view of subjective confidence that emphasizes its diverse functions and wide applicability to many different forms of neural representation and behavior. This view identifies both commonalities and unique features across these forms and identifies the importance of understanding the transformations among them. In particular, we identify a *distributional* form of confidence that pertains to probabilistic representations and a *summary* form that pertains to scalar representations derived from those distributions. We argue that recognizing this distinction and understanding the relationship between these two forms will help to reconcile several apparent controversies and to clarify the agenda for future work in the field.

Formal Definitions and Outline of the Proposal Review

A general understanding of the notion of confidence is that it fundamentally quantifies a degree of belief, or synonymously, a degree of reliability, trustworthiness, certitude, or plausibility. This common notion coincides closely with a formal one: that of Bayesian probability. Although a probability is sometimes considered to describe the likelihood of occurrence of random events in the world, from the viewpoint of an observer, whether such likelihoods constitute objective facts or reflect subjective knowledge is indistinguishable. Thus, probabilities simply *are* degrees of belief from the Bayesian viewpoint (Jaynes, 2003). Recognizing that much remains to be unpacked, we adopt the notion of Bayesian probability as the *formal definition* of subjective confidence.

From this modest premise, our seemingly lofty aim is to bridge the gap between psychology on the one hand and neuroscience on the other. The foundation for our approach is first to recognize that, semantically, confidence is a property (degree, probability, etc.) that describes or modifies a referent (belief, response, memory, future event, etc.). Therefore it is impossible to refer precisely to confidence without specifying the object to which it pertains. In common usage the referent is often not made explicit and this is likely to contribute to conceptual confusion. We propose that *the same general formal notion of confidence as Bayesian probability can be applied to widely different structures and processes*. These include populations of neurons, neural functions, behavioral outputs, persons, etc. Depending on the nature of its referent there are specific and significant consequences for the computational or conceptual definition and treatment of each particular use of confidence (see Box 1: "Current Status of the Field"). Fleshing out this point is the thread that ties together much of this review.

Box 1. Current Status of the Field

- Multiple domains. The sense of confidence characterizes the reliability of internal representations in a variety of cognitive domains, at least: perception, decision accuracy, reward probability, general knowledge, and memorization.
- Multiple manifestations. It can be probed experimentally through several behavioral measures, explicit (verbal reports, ratings, etc.) and implicit (choices, reaction times, etc.).
- Multiple species. The implicit behavioral measures of confidence demonstrate that the sense of confidence is not specifically humans, but shared with other mammals like monkeys and rodents.
- Multiple functions. The estimation of confidence can modulate learning, information seeking and decision-making.
- Multiple processing steps. Confidence is estimated at different stages of information processing: it may characterize sensory inputs, a decision variable, a prediction, a decision process, a post-decision evaluation.
- Different kinds of accuracies. The accuracy of confidence can be assessed as an absolute estimate (whether it can be mapped onto an objective variable) or as a relative estimate (whether trial-by-trial variations make sense).

A key claim of this review is that the notion of ''uncertainty'' used in research on Bayesian neural computation (Fiser et al., 2010; Ma and Jazayeri, 2014; Pouget et al., 2013) and the notion of ''confidence'' used in metacognitive research are two different manifestations of the same concept of Bayesian probability. First, we note that ''uncertainty'' and ''confidence'' are merely the inverse (or reciprocal) of one another, so the choice of emphasis is not an important difference. Instead, the critical difference is that ''confidence'' in the metacognitive field is a single number, such as a numerical rating, whereas ''uncertainty'' in the Bayesian computation field is a property of an array of numbers, such as a distribution of firing rates across neurons. What we will suggest is that the conceptual relationship between these two forms of confidence (uncertainty) is very much the same as the relationship between ''summary statistics'' (mean, standard deviation, etc.) and the data they describe. Summary statistics are scalars and data are sets of distributions of numbers. We will therefore borrow this terminology and refer to *summary confidence* and *distributional confidence*. While in principle summary confidence might share only a nominal relationship to distributional confidence, we argue that from a normative point of view, summary confidence is derived within the brain from distributional confidence, just as a statistician calculates the standard deviation of a distribution. We term this process *confidence readout*.

From this conceptual parcellation it becomes clear that reconciling neuroscientific and psychological approaches will hinge on understanding the relationship between distributional and summary forms of confidence. Our strategy is as follows: first, in Confidence and the Neural Representation of Uncertainty: Distributions and Summaries we review briefly the Bayesian coding field and important elements of this normative view that we embrace. Next, in From Data to Summary: Reading out Summary Confidence from Distributions, we consider the problem of readout of a summary from a computational perspective. We suggest that understanding how summary confidence is derived from distributional confidence is of great importance for confidence research going forward. We then turn to look at some of the diversity of uses of confidence in Uses of Summary Confidence and Behavioral Manifestations, pointing out that explicit reporting of confidence only scratches the surface of the important uses of confidence in adaptive behavior, which include critical functions such as setting learning rates and setting evidence thresholds. In A Brain-Scale, Hierarchical Neural Architecture for Confidence we review attempts to map confidence to neuronal substrates across different brain areas, emphasizing the implications of the fact that neural circuits use both distributional and summary representations of confidence. Finally, in The Rough Edges, we discuss the relationship between Bayesian optimality seen in sensorimotor behaviors and suboptimality seen in confidence reporting and other ''high level'' behaviors, arguing that understanding how confidence summaries are formed in the brain will help to illuminate the latter.

Confidence and the Neural Representation of Uncertainty: Distributions and Summaries

A central example of probabilistic computation is the problem of combining different sources of information. Normatively, this problem requires a solution in which each source is weighted by its inverse uncertainty, or confidence (Jaynes, 2003; Knill and Pouget, 2004; Ma et al., 2006; Pearl, 1997). This general uncertainty-weighting problem is illustrated in Figure 1. This problem occurs in cue combination, such as when inferring the orientation of a bar given both visual and haptic sensory inputs. At a behavioral level, human subjects are indeed close to optimal when performing multi-sensory cue combination (Ernst and Banks, 2002) and in sensorimotor integration (Körding and Wolpert, 2004; Todorov, 2004; Wolpert and Ghahramani, 2000). This raises the natural question of how such probabilistic computations take place in the brain.

Several prominent theories in computational neuroscience posit that computations and information processing in brain circuits are essentially probabilistic, or Bayesian. These theories are strongly normative because computing on probability distributions is considered to be the optimal solution.

A prominent computational theory of how brains implement normative solutions is known as probabilistic population coding. This theory suggests that neurons encode parameters of probability distributions (Knill and Pouget, 2004). Thus, tuning curves are interpreted as likelihood detectors: a neuron tuned to a particular orientation signals the likelihood that the stimulus has this orientation, and a population of neurons tuned to different orientations represents the full probability distribution of the orientation of the stimulus (see Figure 2A), thus forming a probabilistic population code (Deneve et al., 1999; Ma et al., 2006). Another theory, known as Bayesian sampling theory, is similar in spirit to probabilistic population coding but different in details. Sampling theory proposes that neurons encode

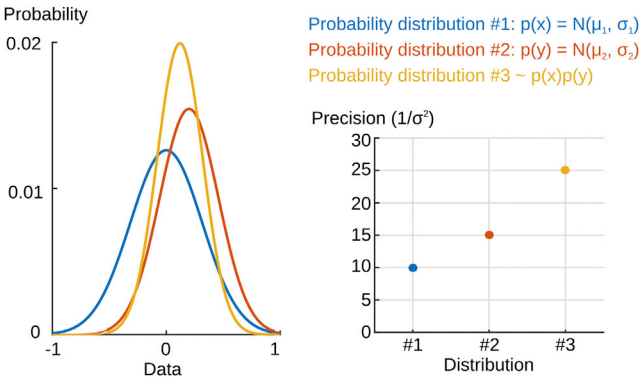

Figure 1. Confidence in a Combination of Inputs

The left plot shows the optimal combination (in yellow) of two input probability distributions (blue and red). Confidence can be read out as the precision of the distributions (their inverse variance). Note that confidence-weighting of information entails that the output distribution (yellow) is closer to the more precise (red) distribution. This optimal combination corresponds to different situations in practice. In the perceptual domain, the input data may be the orientation of a bar provided by visual and tactile information; and the output data the multimodal integration. In the learning domain, the input data may be prior information and current likelihood conveyed by sensory data, and the output data the posterior estimate. The right plot shows that the precision of the combined distribution is higher than that of the input distributions. When distributions are Gaussians as here, the combination of precision is exactly additive.

directly the inferred probabilistic variables (Fiser et al., 2010; Hoyer and Hyvärinen, 2003; Lee and Mumford, 2003). The activity of a neuron at a particular moment is thus interpreted as a sample from the inferred variable, such as the orientation of a stimulus.

We refer the reader to several reviews for more details and discussion about the relative merit of each theory (Fiser et al., 2010; Pouget et al., 2013; Ma and Jazayeri, 2014). For our purposes it is worth highlighting a few key points. First, because these theories posit that activity in neural populations represents (approximately) entire probability distributions, these representations inherently convey confidence information. We call this implicit representation of confidence *distributional confidence*. Second, these theories have not yet been empirically validated. Because we know that behavior can in some cases take into account uncertainty (Ernst and Banks, 2002; Körding and Wolpert, 2004; Ma and Jazayeri, 2014; Maloney and Zhang, 2010), we know that some kind of probabilistic representation must exist, but it need not be a full probability distribution (e.g., Rich et al., 2015). One alternative to the idea of neural codes based on probability distributions are codes in which summary statistics, such as the mean and variance, are represented and computed independently. While arguably less parsimonious, there is some evidence to support representations along these lines (e.g., O'Reilly et al., 2013). As we will argue, we believe it is likely that both such codes (as well as others) co-exist in the brain, in particular if one considers different stages of information processing. We present in From Data to Summary: Reading out Summary Confidence from Distributions the notion of readout: a process that extracts summary statistics from distributional representations.

A third point is that these theories have so far been explored and tested mostly in the domain of perceptual processing (Bejjanki et al., 2011; Berkes et al., 2011; Deneve et al., 1999; Fiser et al., 2010; Kim and Basso, 2010; Ma et al., 2006). It remains an open question to what extent probabilistic computation holds beyond low-level sensory and motor representations: e.g., the belief that "it may rain tomorrow," a reward expectation, etc. Forming probability distributions by simulating internal models could serve as the basis for a distributional neural representation of confidence in a variety of problems. There do exist a number of models for higher-level computations, for instance involving sampling schemes with integration of samples internally generated, e.g., for evaluating general-knowledge statements (Gigerenzer et al., 1991; Juslin et al., 2007; Koriat, 2012), for learning and goal-directed decisions (Hinton and Dayan, 1996; Legenstein and Maass, 2014; Solway and Botvinick, 2012), and even for probabilistic abstract reasoning (Chater et al., 2006; Denison et al., 2013; Vul et al., 2009).

It is clear that much work remains to understand the precise representations used by the brain and that this work will no doubt refine or possibly even upend our notion of confidence insofar as it is embedded within these neural representations.

From Data to Summary: Reading out Summary Confidence from Distributions

Cue combination (Ernst and Banks, 2002; Ma et al., 2006) and motor control (Körding and Wolpert, 2004; Todorov, 2004; Wolpert and Ghahramani, 2000) are examples of behaviors whose optimization requires the use of confidence (inverse uncertainty) and in which computations at the level of probability distributions could elegantly account for both choice and confidence implicitly. For instance, cue combination can be implemented optimally using probabilistic population codes simply by summing the activity of populations of neurons: the very format of probabilistic neural representations could therefore allow an automatic and optimal weighting by uncertainty (Ma et al., 2006). However, there are behaviors in which confidence must be expressed independently of the choice itself. One of the simplest examples is a two-alternative choice decision. Here, a subject is required to select one of two incompatible binary alternatives. This "forced choice" by design eliminates information about confidence that might have existed in the original information on which the decision was based (see Figure 2A). Yet, insofar as the original information was a probabilistic neural representation, the original distributional confidence information should also be available for independent readout. For example, readout into a summary confidence value would allow confidence to be reported verbally on a rating scale. Essentially, what needs to be done is to extract a single number, a scalar, as a summary statistic of an entire distribution. This process, which we term "confidence readout," could be considered closely analogous to the process of extracting a choice from a distribution; choice is just a summary of a different statistic.

To see how this works in a more formal manner, in what follows, we first consider the case of binary choice and we then consider the case of continuous choice or estimation. In Uses of Summary Confidence and Behavioral Manifestations, we will

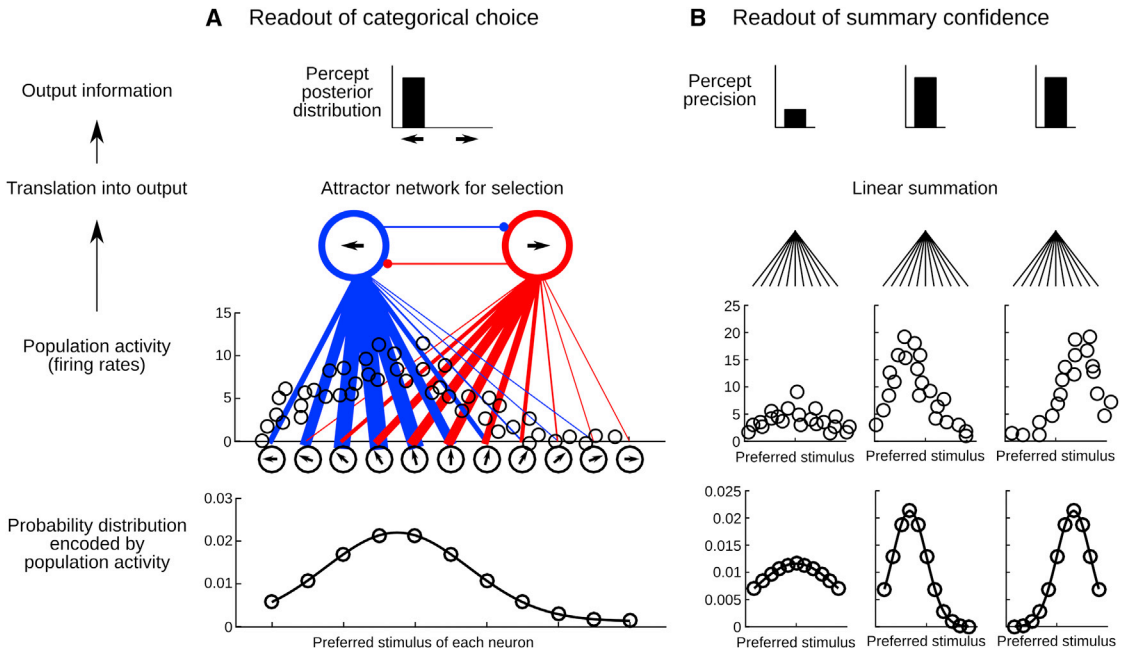

Figure 2. Readout of Choice and Confidence with Probabilistic Neural Codes

We compare two circuits that read out a choice (A) and a confidence level (B) from similar input information: a direction of motion encoded with a probabilistic neural code. In both cases, the bottom graphs depict the probability distributions that can be decoded by a Bayesian observer from the population activity illustrated above. Confidence at the level of the input is represented implicitly as the precision of the distribution.

(A) This circuit reads out a choice: it implements a categorization. It collapses the input distribution into a binary value: direction to the left or to the right. Categorization can be implemented by an attractor network, in which two pools of neurons mutually inhibit each other, and receive excitatory input, as in Wang (2002). The synaptic weights of these excitatory connections are fixed and reflect the feature encoded, e.g., neurons tuned to −90° connect strongly to the blue pool of neurons. The width of the blue and red lines denotes the excitatory drive, which is a function of the fixed synaptic weight and the stimulus-dependent firing rate. The precision of the input distribution contributes to the categorization and its robustness against noise, but this information is lost in the output: the distributional confidence information remains "encapsulated" in this circuit.

(B) This circuit reads out the confidence in the orientation. This corresponds to Example 3 in the text, and it is different from the confidence in a left/right categorization (Example 1 in the text). Here the circuit translates the input distribution into a level of activity that reflects its precision. With probabilistic population codes, this computation can be implemented by linear summation of activities (Ma et al., 2006). Three example distributions are shown to stress that, unlike the "categorization process" illustrated in (A), the output here does not depend on whether the mean value is closer to the left or the right direction, it reflects only precision.

go on to examine in more detail other kinds of functions in which summary readout of confidence may be critical.

Example 1: Readout of Confidence in a Binary Forced-Choice Decision

Consider the decision of whether a stimulus is tilted clockwise ($a+$) or counter-clockwise ($a-$) and the confidence in this decision. We first give a formal treatment of this problem and we then provide a potential implementation.

Given the evidence received, $\mathbf{r}$, (the bold font indicates a vector), there is formally a probability distribution $p(\boldsymbol{a}|\boldsymbol{r})$ that describes the posterior probability of possible angles $\boldsymbol{a}$ given $\boldsymbol{r}$. The optimal choice between $a+$ and $a-$ is to pick the option with the highest integrated probability over all the clockwise or counter-clockwise angles: $p(a+|\boldsymbol{r})$ or $p(a-|\boldsymbol{r})$. Confidence in this decision is then formally just the value of this maximal probability. In this case, the choice and the confidence are read out directly from $p(\boldsymbol{a}|\boldsymbol{r})$. Alternatively, one can compute an intermediate decision variable, d, which should take the form of the log probability ratio (LPR): $d = log(p(a+|\boldsymbol{r})/p(a-|\boldsymbol{r}))$. The choice is then determined by the sign of d (choose $a+$ if $d > 0$, otherwise choose $a-$) and confidence in this decision is the absolute value of d. This formalism is standard in Bayesian decision theory and signal detection theory. More examples and further discussion can be found for instance in Galvin et al. (2003) and Kepecs and Mainen (2012). This example shows (1) that a formal notion of confidence can be quantified in a principled manner, (2) that different, equivalent algorithms can be designed, and (3) that choice and confidence can be read out from the same information. In this example, the distribution $p(\boldsymbol{a}|\boldsymbol{r})$ carries confidence information in "distributional" form (to what extent the mass of the distribution is more on the $a+$ or $a-$ side) and the highest probability among $p(a+|\boldsymbol{r})$ and $p(a-|\boldsymbol{r})$ is a summary that provides a scalar value to express confidence in the decision.

Example 2: Integration of Evidence in Time

One aspect that was omitted in this signal detection formalism is that in real life, the evidence $\mathbf{r}$ often has a temporal dimension. Therefore, momentary evidence must be integrated over time. The computation of the LPR can be updated for each sample of evidence received across time, a procedure known as the sequential probability ratio test, to quantify, at any given moment, what is the best option to choose and what is the associated summary confidence level (Wald, 1945; Wald and

Wolfowitz, 1948). Relying on such a decision variable is the basis of the drift diffusion model (DDM) and related "accumulation-to-bound" models. These models have been extensively used in mathematical psychology (Pleskac and Busemeyer, 2010; Ratcliff, 1988; Smith and Vickers, 1988; Vickers et al., 1985). Importantly, the decision variables posited by these accumulation models have a candidate neural substrate exhibited in the ramping neural activity observed in parietal cortex and other brain regions (reviewed in Gold and Shadlen, 2007). Accumulation-to-bound models can account not only for choice and reactions times, but also for decision confidence (Fetsch et al., 2014; Kepecs et al., 2008; Vickers et al., 1985). The theory of probabilistic population codes can provide a normative algorithm for integration of evidence over time that may be optimal for action selection under a large range of conditions (Beck et al., 2008; Drugowitsch and Pouget, 2012). In these models, a summary confidence level (akin to the LRP) can be computed using linear integration of neural activity (Beck et al., 2008; Drugowitsch and Pouget, 2012).

Example 3: Readout of Confidence as Precision

Consider now the estimation of confidence in a quantitative variable, the orientation of the stimulus. If one thinks of the probability distribution over possible angles, high confidence in the orientation should correspond to a distribution concentrated onto one particular angle (see Figure 2B). This formally corresponds to the precision of the distribution, its inverse variance, which is a natural quantification of confidence in a continuous variable (Meyniel et al., 2015; Yeung and Summerfield, 2012). The distributional confidence information here is contained in the full shape of the distribution, and the summary confidence level by the precision. The precision expresses a specific "loss function," that is, how much cost one pays for being off in one's estimate by a given amount (Maloney and Zhang, 2010). For a loss function based on squared error, the inverse variance is *all* one needs to summarize about the distribution. However, one can easily imagine more complex loss functions in which precision would not be a good summary statistic. For example, errors in one direction may be worse than in the other direction. In this case, the precision may be an approximation, but it does not convey all the confidence information contained in the distribution. Interestingly, with a probabilistic population code and under some biologically plausible assumptions, the precision of a representation such as the orientation of a stimulus is simply proportional to the sum of activities across neurons in the probabilistic population code (Ma et al., 2006). The readout mechanism here is thus as simple as a linear summation (see Figure 2B). However, the proportionality factor in this mechanism, which relates to the number of neurons and the properties of their tuning curves, raises the problem of the calibration of the summary confidence, an issue to which we will return in The Rough Edges.

Relationship to Previous Models of Confidence

To sum up, we have given different names (summary confidence, distributional confidence) to aspects of confidence that we think are worth keeping distinct. We have described how, for simple examples, summary confidence can be derived normatively from the distributional confidence information conveyed by probabilistic neural representations. We will go into more complexity later, with less direct routes and deviations from optimality (see A Brain-Scale, Hierarchical Neural Architecture for Confidence and also The Rough Edges). For the moment, the implications of this basic conceptualization can be related to the classic literature on confidence. We suggest that some confusion in the field of confidence studies is due to the conflation of distributional and summary forms.

We propose that in decision-making, choice and confidence can be read out from the same neural representation (Kepecs and Mainen, 2012; Kepecs et al., 2008). This view resembles the "shared encoding" hypothesis reported by Grimaldi et al. (2015) or "first-order model" (Timmermans et al., 2012) in which the same stream of information accounts for choice and confidence. However, these models are usually thought to entail that the same circuitry underpins choice and confidence (Grimaldi et al., 2015). We suggest the opposite: the mechanisms that read out a choice and a summary confidence from the same representation must be partly different, simply because they result in different things. Such "parallel processing" of choice and confidence is the landmark of "dual route models" (Timmermans et al., 2012), but our framework rejects a pure parallelism by assuming a common initial representation. Our view could therefore seem closer to "hierarchical models" (Fleming and Dolan, 2012; Fleming et al., 2012; Timmermans et al., 2012). However, such models make a distinction between a first-order level (choice) and a second-order level (confidence) processing. This distinction is a landmark in the metacognition literature. In our view, there is no need for such a terminology: readout of choice and confidence are simply different without one being subordinate to the other.

We can see one case in which such a distinction of "orders" makes sense in our view. It is the distinction made between type 1 and type 2 confidence, reviewed in Galvin et al. (2003). We used the example of confidence in whether a stimulus was oriented clockwise or counter-clockwise. This would correspond to first-order (type 1) confidence. This is different from selecting one option, and *then* evaluating the confidence that this selection is correct, which would correspond to second-order (type 2) confidence. Type 2 confidence and the choice are not necessarily read out from the same representation. For instance, additional information about the orientation of the stimulus may be processed between the choice and the type 2 confidence report. Such a two-stage processing has been proposed (Pleskac and Busemeyer, 2010; Resulaj et al., 2009).

Uses of Summary Confidence and Behavioral Manifestations

We have so far discussed how confidence may be represented and how it may be read out from a distribution into a single summary value. We have already seen how distributional confidence (uncertainty) can be used without summary to optimize sensorimotor behaviors. Now we turn to look at some of the uses of summary confidence. We consider first the use of summary confidence in decision-making and then two other examples, learning and sampling.

Decision Optimization

A key example of the use of summary confidence is when subjects report it on a quantitative scale to communicate the reliability of an entity (a choice, a memory, an opinion) to others.

This usage of confidence requires a summary form because it must be reduced to a single value, a scalar. Explicit verbal ratings were the initial thread of research in psychology (Peirce and Jastrow, 1884) and they continue to be a focal point for psychological studies (for reviews, see Galvin et al., 2003; Pleskac and Busemeyer, 2010). Reports of confidence can be useful in collective decision (Bahrami et al., 2010; Bang et al., 2014). Indeed, an optimal collective decision can benefit from the uncertainty weighting of individual decisions and this is an important area for research (Pérez-Escudero and de Polavieja, 2011). For purely individual decisions, expressing summary confidence with a verbal report seems irrelevant, but experimental designs can translate this information into specific behaviors, such as post-decision wagers. Thus, rather than reporting confidence on a scale from high to low, subjects must decide an amount of money (again, a scalar) to invest in a decision with a degree of uncertainty (Persaud et al., 2007). Using a similar logic, animals can also be induced to "wager" on the outcomes of their decisions (reviewed by Kepecs and Mainen, 2012; Smith et al., 2003). Insofar as humans and animals optimize the payoffs of risky decisions, this will induce the expression of confidence in the wagers. This area is an active and important area of convergence of human and animal studies. Let us consider in more detail one of several paradigms, the so-called "opt-out" task.

Considering the example of orientation discrimination, we saw above that a summary confidence level in the orientation discrimination could be derived as the probability that the stimulus has one orientation rather than the other ($a+$ or $a-$). Suppose that the participant is given the opportunity to either provide an answer ($a+$ or $a-$) and gain a large reward if it is correct and no reward in case of error, or to opt-out (decline to provide an answer) and get a small reward for sure. This is a classic value-based decision-making problem (Glimcher et al., 2009; Rangel et al., 2008). Maximizing the expected reward in this problem requires multiplying the reward magnitude and the probability of reward (p_R). If the subject opts out, $p_R = 1$. Otherwise, the subject's estimate of p_R should correspond to the summary confidence in the orientation discrimination. Therefore, having observed a decision to opt-out or not, one can infer whether the subject's summary confidence was above or below the ratio of the value of the sure reward to the risky one (Hampton, 2001; Kiani and Shadlen, 2009; Middlebrooks and Sommer, 2012).

The opt-out example illustrates how choices based on uncertain information can serve to measure a subject's summary confidence. Insofar as people and animals seek to optimize their gains, these choices require that summary confidence is derived as accurately as possible from the subject's internal representation of that information. That is, optimal wagering decisions require optimal readout of summary confidence from distributional confidence, provided that such information is available to inform the outcome. Importantly, because wagering-based measures do not require verbal report, they are well-suited to measuring summary confidence in non-human animals. Also importantly, the opt-out task is only one of a larger class of wagering-like paradigms that can take advantage of ecologically relevant scenarios. In waiting-time paradigms, reward is delivered for correct decisions, but only after a delay and the subject has the opportunity to initiate a new trial instead of waiting. Waiting after an error thus has an opportunity cost, so willingness to wait should, normatively, depend on the estimated accuracy (Kepecs et al., 2008; Lak et al., 2014). For more examples and details on paradigms, see Kepecs and Mainen (2012) and for a review of comparative studies, see Smith et al. (2003). It is important to note that a subject's gain/loss function will also in general dependent on other factors and biases, such as loss or risk aversion, which will interact with it in ways that can make confidence difficult to disentangle from other factors (Fleming and Dolan, 2010).

The above description suggests that behaviors such as opt-out, wagering, waiting-time, etc. are indirect measures of the subject's summary confidence. In our framework, the relevant distributional confidence information could be read out into a summary confidence level that could then be translated into a specific report. However, we also note that the existence of a summary confidence level as an intermediate variable is not necessary as such: the readout of the distributional confidence information could be directly mapped onto a specific behavior.

Optimization of Learning

The role of confidence in learning is often overlooked in the confidence literature, but it is well established in computational learning theory, where it is typically referred to as "uncertainty." Decisions build on prior knowledge and learning. Confidence (or uncertainty) plays a key role in acquiring knowledge and updating it according to new data. The Bayesian view provides a normative account for the updating process, indicating how, based on uncertainty, prior knowledge and new observations should be combined during learning. Optimal algorithms, such as the Kalman filter, developed by engineers in the 1960s form a foundation for modern accounts of learning in cognitive science and neuroscience (Bach and Dolan, 2012; Bland and Schaefer, 2012; Daunizeau et al., 2010; Mathys et al., 2011; Nassar et al., 2010; Payzan-LeNestour and Bossaerts, 2011; Preuschoff and Bossaerts, 2007). Essentially, the more confident we are in a new observation (e.g., because the stimulus is clear), the more this observation should impact our prior knowledge. Conversely, the more confident we are in our prior knowledge (e.g., because of extensive and successful prior experience) the less a new discrepant observation should affect it.

Therefore, in learning, confidence should play the role of a weighting factor to balance incoming and prior information in updating one's current knowledge. There is evidence that human subjects indeed adapt their learning rates according to both prior knowledge (Behrens et al., 2007; McGuire et al., 2014; Nassar et al., 2010; Payzan-LeNestour and Bossaerts, 2011; Yu and Dayan, 2005) and the likelihood of observed data (O'Reilly et al., 2013). One can envision a role for either distributional confidence representations or summary (scalar) representations in this process. On the one hand, if priors and evidence are both represented in properly formed probability distributions (Berkes et al., 2011), then confidence could be assessed in the same manner as it would be for the representation of a sensory posterior (Bejjanki et al., 2011). On the other hand, confidence about current or prior knowledge could also be summarized and represented as a single value that could be used to scale other probability distributions or set a learning rate. The neural

implementation of adjustable learning rates could involve the major ascending neuromodulatory systems (dopamine, serotonin, noradrenaline, and acetylcholine) as they may convey confidence information (Doya, 2002; Yu and Dayan, 2005). Given the small number of neurons and their widespread projection patterns, these systems are more likely to represent scalar values (summary confidence) than distributions (distributional confidence). This view implies that a readout of the summary confidence broadcast by these systems might occur either in the related brainstem nuclei or in their input structures.

The role of confidence in learning raises interesting challenges. How to estimate (algorithmically and mechanistically) confidence in a new observation that suffers from sensory uncertainty is essentially the same problem as extracting confidence for a decision. However, estimating confidence from prior knowledge, so that it can be used to adjust learning rates, has received less attention. Computationally, one way to estimate the error rate of a model is by computing deviations between predictions derived from this knowledge and actual observations (Courville et al., 2004; Preuschoff and Bossaerts, 2007). Such a process is in some sense agnostic of the form in which this prior knowledge is represented before being turned into an error signal. In fact, confidence need not be represented explicitly in any manner other than the error signals themselves.

It has also been proposed that confidence may contribute not only to the optimization of learning algorithms, but also to the selection between different, competing algorithms, such as model-based and model-free learning strategies (Daw et al., 2005).

Optimization of Information-Seeking

Decisions require not only selecting between alternatives but also require knowing when to stop weighing evidence and to act on what is known. When we integrate information sequentially, more samples of information can enable more accurate decisions but at the cost of longer deliberation. A real-world example of this is a student deciding how long to study for an exam. The more information is acquired, the higher the likelihood of giving correct answers in the test and, hence, the greater the probability of the ensuring benefits of good grades. But studying is taxing and has the opportunity cost of not engaging in more interesting activities like socializing. To optimize this problem the student should take advantage of confidence in knowing the information being studied.

We have seen in From Data to Summary: Reading out Summary Confidence from Distributions how confidence can be obtained as an output from decision variables in sequential sampling models. We can also consider in the same models that confidence could also be critical as an *input* to set the level of the stopping bound. Formally, the samples determine the likelihood of each alternative, and the ratio of likelihoods (or its log, as the LPR above) quantifies the expressed confidence level that one alternative is better than the other. In the 1940s, Wald showed that, in decisions based on sequential samples, one can compute the LPR iteratively and commit to a choice when the test first meets a pre-defined confidence criterion (Wald, 1945). Most importantly, Wald showed that this strategy minimizes the number of samples integrated, while controlling the expected error rate (Wald and Wolfowitz, 1948). This decision rule is the principle of drift-diffusion models, among which several variants have been proposed (Kiani et al., 2014; Pleskac and Busemeyer, 2010; Ratcliff, 1988; Smith and Vickers, 1988; Vickers et al., 1985). Little is known, in terms of neural mechanisms, about what determines the bound or stopping rule in a decision process. Scaling a threshold parameter could be implemented by an appropriate confidence summary signal and this too might be conveyed by a neuromodulatory signal.

A Brain-Scale, Hierarchical Neural Architecture for Confidence

Feed-Forward Processing in Brain-Scale Circuits

Our basic view posits that widespread confidence information conveyed by probabilistic neural representations can sometimes be used directly in its implicit distributional form and may in other cases be read out into a summary (scalar) confidence signal. Readout is probably most evident when subjects are interrogated verbally, but it might underlie other behavioral expressions of confidence. The examples of setting learning rates or setting a decision threshold, which might be carried out by scalar signals carried by neuromodulators, illustrate an intrinsic advantage favoring the readout of confidence into a scalar confidence level: a scalar signal requires much less wiring to transmit than a full distribution and may facilitate flexible routing of this kind of information. Thus, although a summary statistic must necessarily be an approximation of a full probability distribution, representing confidence using a scalar could offer benefits of efficiency that outweigh that liability, particularly for functions involving more global computations.

This two-part model implies a kind of "bottom up" processing of information—from implicit distributional forms of confidence to simpler and more explicit, summarized forms. For the case of decision-making this entails two predictions. First, manipulating the source of the original probabilistic representation should affect the readouts of both choices and confidence levels. Stronger levels of evidence lead, as expected, to higher levels of confidence in humans, monkeys, and rodents across a variety of tasks, such as visual or olfactory discrimination (Barthelmé and Mamassian, 2010, 2009; Kepecs et al., 2008; Kiani and Shadlen, 2009; King and Dehaene, 2014; Peirce and Jastrow, 1884). Further evidence is provided by manipulation of neural representations in early sensory areas. Application of transcranial magnetic stimulation (TMS) to the visual cortex while human participants performed a visual task induced changes in both choice and confidence (Rahnev et al., 2012). In monkeys, small currents injected in motion-sensitive regions MT/MST during a motion discrimination task mimicked an increment in perceived motion, which shifted both choices and confidence levels; higher levels of currents mimicked an increment in perceptual noise that degraded the accuracy of both choices and confidence (Fetsch et al., 2014).

Readout from a neural representation can be considered akin to a routing process; it may serve to "untangle" different kinds of information that are implicit in a neural representation. For example, the ventral visual stream is thought to progressively untangle different features of a visual scene through non-linear decoding (DiCarlo and Cox, 2007). Once choice and confidence are read out from a neural representation, different downstream areas may process them independently. Therefore, the second

implication of our proposal is that manipulating a step that is downstream of the source representation might selectively and independently affect the readout of summary confidence or choice. In particular, it should sometimes be possible to manipulate summary confidence while leaving the decision accuracy unaffected.

Indeed, it has been shown that inactivation of the pulvinar nucleus of the thalamus, in a visual opt-out task in monkeys (Komura et al., 2013), and inactivation of the orbito-frontal cortex (OFC), in an olfactory waiting-time task in rats (Kepecs et al., 2008; Lak et al., 2014), selectively disrupted confidence while leaving decision accuracy unaffected. In these studies, it was further shown that activity in these regions correlated with the expressed confidence. However, it is not clear whether inactivation disrupted an area specifically concerned with the representation of summary confidence, or whether instead it disrupted the regulation of a behavior (opt-out, waiting-time) more closely associated with the expression of that confidence. Tests in which multiple reporting modalities can be queried would be helpful to tease apart these alternatives.

Specialized Brain-Scale Circuits for Confidence?

According to our basic view, we would expect that many brain circuits may ultimately prove to be implicated in confidence, but in different ways. Consider the case of a perceptual task that includes a waiting-time component (e.g., Lak et al. (2014); see Uses of Summary Confidence and Behavioral Manifestations). Sensory regions relevant for the decision are strongly involved in computing/representing the distributional confidence information. Additional circuits may perform the readout to extract the summary confidence level, such as perhaps the OFC. And this summary confidence level could then be used to regulate the behavior (how long to wait), which likely involves frontal control and motor circuits (e.g., Murakami et al., 2014). Taken together, this covers a large swath of cortical circuitry.

A more specific and challenging question is whether there are specific regions or circuits critical for confidence readout or summary confidence representation. It has been proposed that the anterior prefrontal cortex (corresponding to the OFC, the fronto-polar cortex, and dorso-medial PFC) could be involved in this process (Barttfeld et al., 2013; Fleming et al., 2010; Kepecs et al., 2008; Lak et al., 2014; Middlebrooks and Sommer, 2012; Yokoyama et al., 2010). De Martino and colleagues (De Martino et al., 2013) showed that in value-based decisions in humans, the ventromedial prefrontal cortex (vmPFC) conveys mixed information about value and confidence, hence providing an implicit code of confidence. Instead, the rostrolateral prefrontal cortex (rlPFC) correlated with confidence (independently of value) and hence may be, in this task, a more likely candidate to encode the result of the readout process. Moreover, the strength of the connectivity between vmPFC and rlPFC predicted the precision of confidence judgments of individuals in this task. This is a macroscopic marker that suggests that variability in the readout (here indexed by the strength between these two cortical regions) can account for inter-individual differences in the estimation of confidence.

The quest for a general brain circuit performing confidence readout is complicated in practice because it is difficult to disentangle between circuits that perform a readout of a summary confidence level from the circuits serving as input for the readout or those that use this readout. We point to two brain circuits that may suffer this issue. The first is the anterior prefrontal cortex reported above. This region is often associated with confidence in tasks that involve evaluating one's own performance and the detection of errors (Yeung and Summerfield, 2012), which is likely to be related to circuits involved in executive and cognitive control (Fernandez-Duque et al., 2000). The second is the ventral striatum, which has been related to confidence (d'Acremont et al., 2013; Hebart et al., 2014; O'Reilly et al., 2013; Preuschoff et al., 2006; Vilares et al., 2012), but mostly insofar as predicting a reward or success is involved. A more nuanced view would suggest that different circuits (which remain to be pinned down) will be flexibly involved depending on the way the confidence information is routed based on behavioral needs.

Interestingly, the readout of confidence can be selectively impaired in specific domains. Fleming and colleagues reported such a case: patients with brain lesions in the anterior PFC had preserved performance in the memory and perceptual domains and degraded confidence judgments specifically in the perceptual domain (Fleming et al., 2014). The fact that choice performance was preserved rules out the possibility that perception or memory, as a whole, were impaired, and points to the readout of confidence itself. This example suggests that one region alone does not suffice to read out confidence: at a minimum, it should involve a circuitry to collect specific inputs from different cognitive domains.

Beyond Forward Linear Processing of Information

The view adopted so far might suggest feed-forward linear processing of information. However, considering both the highly recurrent nature of brain circuitry and the fact that sources of uncertainty are often interdependent, we suspect that the real case will be a hierarchical and loopy architecture, with branches and feedback (Bach and Dolan, 2012). As a first example, following the example discussed in From Data to Summary: Reading out Summary Confidence from Distributions (Galvin et al., 2003), consider the distinction between confidence in an orientation discrimination (type 1) and the confidence in the answer made about this orientation discrimination (type 2). The contrast between two experimental studies illustrates the need, for post-decision (type 2) confidence, to represent both the sensory information and the answer made. The injection of noise by applying TMS to the visual cortex could be expected to perturb the confidence report in a visual task, as was shown in Rahnev et al. (2012). But curiously, TMS applied to the dorsal premotor cortex also disrupted the accuracy of the confidence report while leaving the accuracy of the decision unaffected (Fleming et al., 2015). This pair of experiments could be interpreted as implying that premotor TMS introduced uncertainty about which motor responses was made, and thus about whether the visual decision was accurate.

A second example relates to learning. Generally, each time new evidence is received, a prior and a current likelihood must be combined to infer a posterior. This implies a recursive process: at each iteration, the prior must be adjusted based on the observed accuracy of the prediction that arose from this prior and the momentary evidence (likelihood). The need for a

feedback loop again breaks the simple linear forward information processing view.

Encapsulation of Confidence

Not all the probabilistic representations will undergo summarization and broadcasting. Distributional confidence information could be described as "encapsulated" when it is implicitly conveyed by a probabilistic neural representation but remains confined within a particular, specialized circuit. In computer programming, from where the term is drawn, encapsulation is a way of *deliberately* shielding information; here, we consider only that it is *de facto* not accessible. There are several non-mutually exclusive reasons why distributional confidence information may not be accessible outside from a given circuit or brain region. The most trivial is that that area is poorly connected to other regions. Another reason is that the circuit is connected, but the connections are not amenable to a readout of confidence. A simple example of this is a circuit that performs response selection (see Figure 2; Soltani and Wang, 2010; Wang, 2002). This circuit may accurately compute the probability that a stimulus is present but it may collapse this probability to yield a binary variable signaling the presence or absence of the stimulus. In an action selection circuit such as this one, details about the computation, including the confidence information, remain inaccessible.

Dense connections of the sort that might be necessary for readout are typically found within specialized, well-tuned systems for perception, motor control or learning. But connections between systems are usually sparser, a connectivity profile known as "rich-club" (van den Heuvel et al., 2012; Zylberberg et al., 2010). This could suggest that not all sub-systems can read out confidence from all the other sub-systems, and that only a limited fraction of the distributional confidence information is read out and routed between systems. Or in other words, that confidence information may usually remain encapsulated.

A Global Workspace for Confidence?

When confidence is not encapsulated but read out, the summary confidence level can be used for multiple purposes (see Uses of Summary Confidence and Behavioral Manifestations) across a variety of cognitive tasks. This suggests a flexible routing of summary confidence levels between different domains in a "global neuronal workspace," a set of interconnected high-level cortical regions that underpins the flexible sharing and routing of information globally in the brain (for a review, see Dehaene et al., 2014).

A hallmark of global workspace processing is that only a limited amount of information is selected and amplified (Dehaene, 2014; Sergent et al., 2005; Zylberberg et al., 2010). Due to its limited capacity, global workspace processing may, unlike sensorimotor transformations, be incompatible with full probabilistic representation and inference. Global processing may nevertheless be able to access summary forms of confidence, as, for instance, the level of accumulated evidence can be monitored (Dehaene et al., 2014).

Limited capacity may be the cost of the flexibility afforded by processing in a global workspace. Some experimental designs make such a flexible routing particularly salient. One example is when confidence in performance must be used not only within a particular task type, but also compared between two different task types (de Gardelle and Mamassian, 2014). Use of simpler scalar confidence representations could allow the flexibility necessary to perform this comparison even for arbitrary pairs of tasks. Another example is the comparison of confidence from two brain systems (e.g., model-based and model-free learning systems) to decide which strategy to follow (Daw et al., 2005). In principle, this could also be performed by comparing summary confidence levels without the use of full distributional information.

Shea and colleagues (Shea et al., 2014) have argued that confidence reports made by non-human animals (see Uses of Summary Confidence and Behavioral Manifestations) do not necessarily require a global workspace. This is an important point, also valid in humans. Cognitive processes are full of examples in which some information is available to a certain extent and impacts behavior, but without being reportable (Atas et al., 2014; Rose et al., 2005; Schlaghecken et al., 2000; van Zuijen et al., 2006). A relevant example for confidence may be subliminal reinforcement learning: subjects can become confident in the reward delivery following specific cues, which is demonstrated by their choices and the neural prediction errors observed in case of violations, but they remain unaware of it (Pessiglione et al., 2008).

The Rough Edges

The aim of this section is to help to solve a conundrum. If neural circuits function inherently probabilistically, why is confidence sometimes estimated in a way that is inconsistent with probability theory, reflecting biases and reliable inconsistencies (Kahneman, 2013)? Many behaviors are close to optimal (Ma and Jazayeri, 2014; Maloney and Zhang, 2010; Pouget et al., 2013) but decades of experimentation in "real-life" decision problems have also shown that humans commonly assign confidence suboptimally, relying on sub-samples of the data, focusing on tokens (representative exemplars), ignoring the variance of the distribution, and over-weighting evidence confirming commitments and choices that have been made (Griffin and Tversky, 1992; Kahneman, 2013).

One simple possibility is that errors of confidence are found in real life because real life involves complex high dimensional problems for which there may not be accurate distributional representations available in neural circuits. However, distortions of confidence can be observed even in simple decision tasks (Graziano and Sigman, 2009; Jarvstad et al., 2013; Rahnev et al., 2012; Wu et al., 2009; Zylberberg et al., 2012, 2014). Distortions of confidence were mentioned already in the seminal work of Peirce and Jastrow on small differences in tactile perception (Peirce and Jastrow, 1884). Here we argue that suboptimality arises from approximations inherent in probabilistic representations, particularly in the readout of summary confidence from probabilistic representations, and that this framework can potentially help to explain the specific deviations from optimality that arise as a consequence of specific features of these approximations.

Calibration

Formally, a confidence level in X is calibrated if it reflects directly a normative quantity, such as the probability of X (Baranski and Petrusic, 1994; Kepecs and Mainen, 2012; Koriat, 2012). Uncalibrated, or miscalibrated estimates are pervasive issues for

confidence. Even in language, individuals use different expressions to describe identical probabilistic situations (Wallsten and Budescu, 1995). Verbal expression of probabilities is highly idiosyncratic, but characterizing individual idiosyncrasies helps to standardize linguistic reports of uncertainty between individuals (Karelitz and Budescu, 2004). This study showed that probabilities are represented precisely but translate differently between individuals' verbal readouts. The expression of confidence may also differ culturally: some groups of people express continuous notions of probabilities whereas other groups are more categorical (all-or-none) (Phillips and Wright, 1977). Well-calibrated confidence levels have obvious advantages, e.g., for sharing confidence between individuals (Bang et al., 2014). Calibrated summary confidence (be they linguistic or not) is also useful at the individual level, for instance to use confidence across different tasks (de Gardelle and Mamassian, 2014), or more generally, whenever confidence is used to adjust a behavior.

The readout of summary confidence in neural systems is subjected to calibration issues. In From Data to Summary: Reading out Summary Confidence from Distributions, we described a mechanism to compute the precision of a neural representation (the orientation of a stimulus), that relied on a scaling factor (Ma et al., 2006). The possibility that precision estimates from two different representations may be scaled differently impedes a direct comparison. Normalization would provide an absolute reference, but may not be trivial to compute. To establish a mapping between a scalar quantity and a norm (a probability, a precision), scaling factors and transfer functions may have to be adjusted. As with other mappings, this process may require learning and feedback, so that with substantial training distortions of confidence may be reduced to calibrate the readout process. Indeed, at least at the behavioral level, there is evidence that a better calibration of confidence reports can be achieved by relying on appropriate feedback (Baranski and Petrusic, 1994; Hart et al., 2015). At the neuronal level, the implementation of such a tuning of the readout is entirely an open issue and, similarly, the class of problems for which a precise readout of summary confidence levels can be achieved remains largely unknown. However, the fact that readout parameters ought to be learned indicates that a neural circuit for confidence cannot rely on purely feed-forward processing, but most likely also involves feedback mechanisms to calibrate and adjust the parameters.

Heuristics Revisited

Our framework posits that confidence and choice result from different readouts of the same neuronal circuits. Again this needs to be reconciled with a very different view that emerges from the field of behavioral economics, which has proposed that confidence estimates rely on short-cuts and "heuristics" (Kahneman and Tversky, 1982; Tversky and Kahneman, 1974). The term "heuristic" is often quite vague and may refer to very different computations. Some are simply approximations. Below we discuss how approximations in the readout process may result, in our framework, in expressions of confidence that are typically observed in human subjects. A second type of heuristic involves relying on observables beyond the relevant distributional neural representations of confidence. For instance in a visual task, one could index confidence by statistics of the input that reflect the difficulty of the decision, e.g., whether the stimulus is crowded, masked, etc., rather than on the more complex neural representation on which the choice is based.

It is possible to quantify the "richness" and sophistication of the information from which confidence is read out, by testing the extent to which a subset of the available information can account for the observed confidence. In the previous example, confidence about performance reported by subjects matched their actual performance more closely than what could be predicted from particular visual properties of the stimuli (crowdedness and masking), suggesting that subjects based their choice *and* confidence report on more complex information (Barthelmé and Mamassian, 2010). Similarly, in a probabilistic learning task, subjective confidence in the learned estimates followed the optimal Bayesian confidence levels in a tighter parallel than a whole list of "cues" taken together, suggesting that the representation from which confidence is read out was particularly rich and most likely probabilistic (Meyniel et al., 2015).

The two examples above (Barthelmé and Mamassian, 2010; Meyniel et al., 2015) are cases stressing that choice readout and confidence readout can be based on the same information. But it is a corollary of our thesis that the confidence readout may be based only on a subset of the information, compared to the choice readout. This is one restatement of the notion of "heuristic" in terms of the model that we propose in which different levels of description of probability distributions are assumed. Our framework also posits the existence of "encapsulated" confidence information: the relevant source of information from which to derive a summary confidence may simply not be accessible. In addition, we introduced the idea that readout mechanisms of confidence could "learn" how to produce reasonable summary confidence levels. It is conceivable in our framework that learned readout strategies have converged to very indirect correlates of confidence levels, especially when the distributional confidence is encapsulated. This is for instance what people do when they infer the subjective confidence of another person, simply through the observation of their actions (Patel et al., 2012).

Another way of reinterpreting "heuristic" with our probabilistic framework is to consider the case in which the correct computation is applied in the wrong situation. One example is provided by the drift diffusion model (DDM). In the DDM, under some assumptions, the time to reach a pre-defined threshold of accumulated evidence is a valid summary statistic for confidence in the decision (Kiani and Shadlen, 2009; Kiani et al., 2014). One of the assumptions is that the process that delivers the noisy evidence is stable across time. Reading out confidence from decision time, which is often called a heuristic, is therefore a valid readout mechanism, but only in some circumstances; using it when the input signal is not stable therefore produces sub-optimal summary confidence levels (Kiani et al., 2014).

Another example of correct computation in the wrong situation is the confirmatory bias or "halo" effect, according to which we seek and favor evidence that confirms our current hypothesis. This effect is seen even in very simple perceptual decisions such as luminance or motion judgments (Zylberberg et al., 2012). A Bayesian analysis can capture the asymmetry that evidence favoring the current hypothesis may be weighted differently than evidence against. In Bayesian terms, the confirmatory

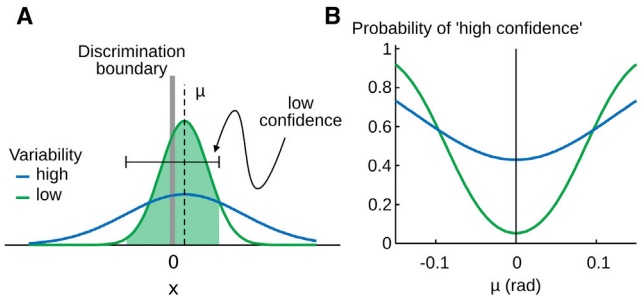

Figure 3. Deviation from optimality in the framework of signal detection theory

(A) The green curve illustrates how to account for decision (here, orientation discrimination) and confidence with signal detection theory. Sensory evidence is represented internally by a variable, x. The true orientation of the stimulus, μ, is translated into the internal variable x with some noise, as denoted by the green distribution. Say that the subject has to categorize this orientation with respect to 0. Despite the fact that true orientation (μ) is positive, it is not unlikely that a negative value is erroneously sampled, since a large portion of the green distribution crosses the boundary. The probability of error would be reduced for higher values of the true orientation (μ more positive). As a consequence, the sign of smaller values of x are more likely to differ from the actual true sign of μ because of perceptual noise: erroneous categorization is more likely. The width of the low confidence zone should depend on the estimated noise in the perceptual system (width of the green distribution).

(B) The curves show the probability for a given true orientation μ to be associated with sensory evidence (denoted x in panel A) in the high confidence zone. The blue curve illustrates the counter-intuitive result that when the signal is low (μ ~0) and the confidence zone remains the same, then an increase in perceptual noise should lead to over-confidence.

Box 2. Future Directions

- Neural codes for probabilities and uncertainty. At least two families of probabilistic neural representations were proposed: probabilistic neural codes and sampling-based codes. Do different codes correspond to different uses of confidence?
- Mechanism for readout. We described confidence as an emergent property of computations based on probabilistic neural representations. Confidence can however also be read out and summarized: what are the neural mechanisms that single out and extract the confidence information from probabilistic neural representations?
- How many systems? Distinct neural correlates were reported for distinct kinds of uncertainty. However these correlates can correspond to the readout of confidence levels as such or to computations that are entailed by this confidence level (decide to wait, collect more information, etc.). It is still unclear which systems truly read out summary confidence.
- The role of global processing. Confidence is sometimes reported explicitly, suggesting that it is processed in a global workspace. To what extent can confidence be read out without entering a global workspace and how may global processes interact with local readout processes?

bias corresponds to biasing the current sample by the posterior probability estimated so far. Biasing momentary samples by prior expectations facilitates perceptual decisions and is relevant when the input signals are structured in time, as they often are (Summerfield and de Lange, 2014). However, when one should learn equally and independently from each samples, then it is sub-optimal to bias momentary evidence in this way.

Approximated Probabilistic Computations

Deviations from optimality observed in behavior are often used as reasons to reject a probabilistic view of brain functioning. The examples above show that decision making, even in simple perceptual problems, can follow a probabilistic logic and still be suboptimal because of specific approximations (Griffiths et al., 2012; Ma and Jazayeri, 2014). Acerbi and colleagues showed in a sensory motor task that prior distributions with different shapes (Gaussian, non-symmetric, bimodal, etc.) could be used with only minor errors that, crucially, were independent from the shape of the distribution (Acerbi et al., 2014). Their results suggest that deviations from optimality observed in the behavior are not due to a fundamental inability to represent and combine probability distributions, but might instead be due to random noise in this process. A similar argument is made by Costello and Watts, who suggest that biases in probability judgment may arise from a fundamental adherence to probability theory, but corrupted by random approximations (Costello and Watts, 2014).

An algorithmic level of description also reveals that the Bayesian framework accounts for both optimal behaviors and systematic deviations. For instance, processing a sequence of inputs can be modeled algorithmically by particle filters, a popular approximation of Bayesian inference. These models have the advantage of relying only on a few free parameters (e.g., the number of samples used, the stability of samples between iterations, etc.), which delineate some regimes that are close to optimal inference, and other regimes that produce common biases such as primacy and recency effects (Abbott and Griffiths, 2011).

One last example shows that approximations in uncertainty (variance) estimates can be amplified in confidence levels. For instance, in signal detection theory, confidence relates formally to the ratio between the mean evidence ("signal") and its reliability ("variance"). Note that since the variance term is in the denominator, because of scaling effects, slight errors in the estimation of the variance may lead to large mis-estimations of confidence. A full Bayesian analysis confirms this intuition and experiments show that slight mis-perception in the variance of an internal representation may lead to marked overconfidence in trials with unreliable evidence (Zylberberg et al., 2014; see Figure 3).

In summary, what we perceive as heuristics in confidence judgments may result from different sources: (1) genuine approximations of a read-out process, including issues of calibration; (2) applying stereotyped read-out procedures that make certain assumptions that do not hold in a given context, a form of approximation referred as relaxation; and (3) using variables that covary with the relevant neural confidence information in cases in which this information is not accessible for explicit reports.

Challenges

We acknowledge that there are substantial experimental challenges for the identification of distributional and summary confidence signals in the brain (see Box 2 "Future directions"). Indeed, decoding distributional confidence information will ultimately require one to understand the nature of the relevant probabilistic neural representations. This may be particularly difficult

for representations that are akin to probability distributions, involving numerous neurons. Simultaneous recordings of the relevant neurons, together with "labels" of what each neuron encodes are necessary, which is difficult if these relevant neurons are intermingled and scattered in a large population, or if the information they encode is not a simple property amenable to selective manipulation.

It may seem comparatively easy to find neurons whose activity co-varies with summary confidence. However, as we have reviewed, such co-variation is not strong evidence that this activity results directly from a readout of confidence, since it could also reflect processes that correlate with confidence levels, either being downstream as part of the reporting mechanism or in confidence-regulated functions such as learning, information seeking, etc. We saw in A Brain-Scale, Hierarchical Neural Architecture for Confidence that it is difficult to tease apart the structures that are involved in reading out confidence from the structures that use summary confidence levels.

Another important challenge for our theory is that it is inspired by what we know about probabilistic neural codes, which is still largely restricted to sensory areas. We make the proposal that the separation between distributional confidence information and its readout into a summary could be general. However, the way it works will depend on the specifics of neural codes, which could differ in non-sensory domains. Interestingly, a number of classic models of cognitive and neural processing can be recast in the framework of probabilistic coding (Solway and Botvinick, 2012), providing candidate hypotheses for neural representations in these domains.

Conclusions

In this perspective we stressed that defining confidence as Bayesian probability clarifies the notion of confidence and invites one to consider a wide range of functions and implementations for confidence-based computations in the brain. In that sense, the concept of confidence may be pervasively relevant in neuroscience and broader than previously envisaged. We proposed a distinction between two fundamental levels: distributional confidence information, conveyed by probabilistic neural representations, and the summary confidence values that can be read out from these distributions. We highlighted different functional characteristics and kinds of confidence-based computations and caution that the study of confidence should therefore not be separated from its functions and the levels at which it operates.

ACKNOWLEDGMENTS

This work received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 604102 (Human Brain Project). It was also supported in part by a grant from the Simons Foundation (325057) to Z.F.M.

REFERENCES

Abbott, J.T., and Griffiths, T.L. (2011). Exploring the influence of particle filter parameters on order effects in causal learning. In Proceedings of the 33rd Annual Conference of the Cognitive Science Society.

Acerbi, L., Vijayakumar, S., and Wolpert, D.M. (2014). On the origins of suboptimality in human probabilistic inference. PLoS Comput. Biol. *10*, e1003661.

Atas, A., Faivre, N., Timmermans, B., Cleeremans, A., and Kouider, S. (2014). Nonconscious learning from crowded sequences. Psychol. Sci. *25*, 113–119.

Bach, D.R., and Dolan, R.J. (2012). Knowing how much you don't know: a neural organization of uncertainty estimates. Nat. Rev. Neurosci. *13*, 572–586.

Bahrami, B., Olsen, K., Latham, P.E., Roepstorff, A., Rees, G., and Frith, C.D. (2010). Optimally interacting minds. Science *329*, 1081–1085.

Bang, D., Fusaroli, R., Tylén, K., Olsen, K., Latham, P.E., Lau, J.Y.F., Roepstorff, A., Rees, G., Frith, C.D., and Bahrami, B. (2014). Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. Conscious. Cogn. *26*, 13–23.

Baranski, J.V., and Petrusic, W.M. (1994). The calibration and resolution of confidence in perceptual judgments. Percept. Psychophys. *55*, 412–428.

Barthelmé, S., and Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system. PLoS Comput. Biol. *5*, e1000504.

Barthelmé, S., and Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. Proc. Natl. Acad. Sci. USA *107*, 20834–20839.

Barttfeld, P., Wicker, B., McAleer, P., Belin, P., Cojan, Y., Graziano, M., Leiguarda, R., and Sigman, M. (2013). Distinct patterns of functional brain connectivity correlate with objective performance and subjective beliefs. Proc. Natl. Acad. Sci. USA *110*, 11577–11582.

Beck, J.M., Ma, W.J., Kiani, R., Hanks, T., Churchland, A.K., Roitman, J., Shadlen, M.N., Latham, P.E., and Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. Neuron *60*, 1142–1152.

Behrens, T.E.J., Woolrich, M.W., Walton, M.E., and Rushworth, M.F.S. (2007). Learning the value of information in an uncertain world. Nat. Neurosci. *10*, 1214–1221.

Bejjanki, V.R., Beck, J.M., Lu, Z.-L., and Pouget, A. (2011). Perceptual learning as improved probabilistic inference in early sensory areas. Nat. Neurosci. *14*, 642–648.

Berkes, P., Orbán, G., Lengyel, M., and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science *331*, 83–87.

Bland, A.R., and Schaefer, A. (2012). Different varieties of uncertainty in human decision-making. Front. Neurosci. *6*, 85.

Chater, N., Tenenbaum, J.B., and Yuille, A. (2006). Probabilistic models of cognition: conceptual foundations. Trends Cogn. Sci. *10*, 287–291.

Costello, F., and Watts, P. (2014). Surprisingly rational: probability theory plus noise explains biases in judgment. Psychol. Rev. *121*, 463–480.

Courville, A.C., Gordon, G.J., Touretzky, D.S., and Daw, N.D. (2004). Model uncertainty in classical conditioning. Adv. Neural Inf. Process. Syst. *16*, 977–984.

d'Acremont, M., Schultz, W., and Bossaerts, P. (2013). The human brain encodes event frequencies while forming subjective beliefs. J. Neurosci. *33*, 10887–10897.

Daunizeau, J., den Ouden, H.E.M., Pessiglione, M., Kiebel, S.J., Stephan, K.E., and Friston, K.J. (2010). Observing the observer (I): meta-bayesian models of learning and decision-making. PLoS ONE *5*, e15554.

Daw, N.D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. *8*, 1704–1711.

de Gardelle, V., and Mamassian, P. (2014). Does confidence use a common currency across two visual tasks? Psychol. Sci. *25*, 1286–1288.

De Martino, B., Fleming, S.M., Garrett, N., and Dolan, R.J. (2013). Confidence in value-based choice. Nat. Neurosci. *16*, 105–110.

Dehaene, S. (2014). Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts (New York: Viking).

Dehaene, S., Charles, L., King, J.-R., and Marti, S. (2014). Toward a computational theory of conscious processing. Curr. Opin. Neurobiol. *25*, 76–84.

Deneve, S., Latham, P.E., and Pouget, A. (1999). Reading population codes: a neural implementation of ideal observers. Nat. Neurosci. *2*, 740–745.

Denison, S., Bonawitz, E., Gopnik, A., and Griffiths, T.L. (2013). Rational variability in children's causal inferences: the Sampling Hypothesis. Cognition *126*, 285–300.

DiCarlo, J.J., and Cox, D.D. (2007). Untangling invariant object recognition. Trends Cogn. Sci. *11*, 333–341.

Doya, K. (2002). Metalearning and neuromodulation. Neural Netw. *15*, 495–506.

Drugowitsch, J., and Pouget, A. (2012). Probabilistic vs. non-probabilistic approaches to the neurobiology of perceptual decision-making. Curr. Opin. Neurobiol. *22*, 963–969.

Ernst, M.O., and Banks, M.S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature *415*, 429–433.

Fernandez-Duque, D., Baird, J.A., and Posner, M.I. (2000). Executive attention and metacognitive regulation. Conscious. Cogn. *9*, 288–307.

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

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

Fleming, S.M., and Dolan, R.J. (2010). Effects of loss aversion on post-decision wagering: implications for measures of awareness. Conscious. Cogn. *19*, 352–363.

Fleming, S.M., and Dolan, R.J. (2012). The neural basis of metacognitive ability. Philos. Trans. R. Soc. Lond. B Biol. Sci. *367*, 1338–1349.

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

Fleming, S.M., Dolan, R.J., and Frith, C.D. (2012). Metacognition: computation, biology and function. Philos. Trans. R. Soc. Lond. B Biol. Sci. *367*, 1280–1286.

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

Fleming, S.M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., and Lau, H. (2015). Action-specific disruption of perceptual confidence. Psychol. Sci. *26*, 89–98.

Galvin, S.J., Podd, J.V., Drga, V., and Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. Psychon. Bull. Rev. *10*, 843–876.

Gigerenzer, G., Hoffrage, U., and Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. Psychol. Rev. *98*, 506–528.

Glimcher, P.W., Camerer, C., Fehr, E., and Poldrack, R.A. (2009). Introduction: A brief history of neuroeconomics. In Neuroeconomics Decision Making and the Brain, P.W. Glimcher, C.F. Camerer, E. Fehr, and R.A. Poldrack, eds. (Elsevier).

Gold, J.I., and Shadlen, M.N. (2007). The neural basis of decision making. Annu. Rev. Neurosci. *30*, 535–574.

Graziano, M., and Sigman, M. (2009). The spatial and temporal construction of confidence in the visual scene. PLoS ONE *4*, e4909.

Griffin, D., and Tversky, A. (1992). The weighing of evidence and the determinants of confidence. Cognit. Psychol. *24*, 411–435.

Griffiths, T.L., Chater, N., Norris, D., and Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): comment on Bowers and Davis (2012). Psychol. Bull. *138*, 415–422.

Grimaldi, P., Lau, H., and Basso, M.A. (2015). There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making. Neurosci. Biobehav. Rev. *55*, 88–97.

Hampton, R.R. (2001). Rhesus monkeys know when they remember. Proc. Natl. Acad. Sci. USA *98*, 5359–5362.

Hart, W., Tullett, A.M., Shreves, W.B., and Fetterman, Z. (2015). Fueling doubt and openness: experiencing the unconscious, constructed nature of perception induces uncertainty and openness to change. Cognition *137*, 1–8.

Hebart, M.N., Schriever, Y., Donner, T.H., and Haynes, J.-D. (2014). The Relationship between Perceptual Decision Variables and confidence in the human brain. Cereb. Cortex. Published online August 11, 2014. http://dx.doi.org/10.1093/cercor/bhu181.

Hinton, G.E., and Dayan, P. (1996). Varieties of Helmholtz machine. Neural Netw. *9*, 1385–1403.

Hoyer, P.O., and Hyvärinen, A. (2003). Interpreting neural response variability as Monte Carlo sampling of the posterior. Adv. Neural Inf. Process. Syst. *15*, 277–284.

Jarvstad, A., Hahn, U., Rushton, S.K., and Warren, P.A. (2013). Perceptuo-motor, cognitive, and description-based decision-making seem equally good. Proc. Natl. Acad. Sci. USA *110*, 16271–16276.

Jaynes, E.T. (2003). Probability Theory: The Logic of Science (Cambridge University Press).

Juslin, P., Winman, A., and Hansson, P. (2007). The naïve intuitive statistician: a naïve sampling model of intuitive confidence intervals. Psychol. Rev. *114*, 678–703.

Kahneman, D. (2013). Thinking, Fast and Slow (New York: Farrar, Straus and Giroux).

Kahneman, D., and Tversky, A. (1982). Variants of uncertainty. Cognition *11*, 143–157.

Karelitz, T.M., and Budescu, D.V. (2004). You say "probable" and I say "likely": improving interpersonal communication with verbal probability phrases. J. Exp. Psychol. Appl. *10*, 25–41.

Kepecs, A., and Mainen, Z.F. (2012). A computational framework for the study of confidence in humans and animals. Philos. Trans. R. Soc. Lond. B Biol. Sci. *367*, 1322–1337.

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

Kiani, R., and Shadlen, M.N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. Science *324*, 759–764.

Kiani, R., Corthell, L., and Shadlen, M.N. (2014). Choice certainty is informed by both evidence and decision time. Neuron *84*, 1329–1342.

Kim, B., and Basso, M.A. (2010). A probabilistic strategy for understanding action selection. J. Neurosci. *30*, 2340–2355.

King, J.-R., and Dehaene, S. (2014). A model of subjective report and objective discrimination as categorical decisions in a vast representational space. Philos. Trans. R. Soc. Lond. B Biol. Sci. *369*, 20130204.

Knill, D.C., and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci. *27*, 712–719.

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

Körding, K.P., and Wolpert, D.M. (2004). Bayesian integration in sensorimotor learning. Nature *427*, 244–247.

Koriat, A. (2012). The self-consistency model of subjective confidence. Psychol. Rev. *119*, 80–113.

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

Lee, T.S., and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. J. Opt. Soc. Am. A Opt. Image Sci. Vis. *20*, 1434–1448.

Legenstein, R., and Maass, W. (2014). Ensembles of spiking neurons with noise support optimal probabilistic inference in a dynamically changing environment. PLoS Comput. Biol. *10*, e1003859.

Ma, W.J., and Jazayeri, M. (2014). Neural coding of uncertainty and probability. Annu. Rev. Neurosci. *37*, 205–220.

Ma, W.J., Beck, J.M., Latham, P.E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nat. Neurosci. *9*, 1432–1438.

Maloney, L.T., and Zhang, H. (2010). Decision-theoretic models of visual perception and action. Vision Res. *50*, 2362–2374.

Mathys, C., Daunizeau, J., Friston, K.J., and Stephan, K.E. (2011). A bayesian foundation for individual learning under uncertainty. Front. Hum. Neurosci. *5*, 39.

McGuire, J.T., Nassar, M.R., Gold, J.I., and Kable, J.W. (2014). Functionally dissociable influences on learning rate in a dynamic environment. Neuron *84*, 870–881.

Meyniel, F., Schlunegger, D., and Dehaene, S. (2015). The sense of confidence during probabilistic learning: a normative account. PLoS Comput. Biol. *11*, e1004305.

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

Murakami, M., Vicente, M.I., Costa, G.M., and Mainen, Z.F. (2014). Neural antecedents of self-initiated actions in secondary motor cortex. Nat. Neurosci. *17*, 1574–1582.

Nassar, M.R., Wilson, R.C., Heasly, B., and Gold, J.I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. J. Neurosci. *30*, 12366–12378.

O'Reilly, J.X., Jbabdi, S., Rushworth, M.F.S., and Behrens, T.E.J. (2013). Brain systems for probabilistic and dynamic prediction: computational specificity and integration. PLoS Biol. *11*, e1001662.

Patel, D., Fleming, S.M., and Kilner, J.M. (2012). Inferring subjective states through the observation of actions. Proc. Biol. Sci. *279*, 4853–4860.

Payzan-LeNestour, E., and Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. PLoS Comput. Biol. *7*, e1001048.

Pearl, J. (1997). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (Morgan Kaufmann).

Peirce, C.S., and Jastrow, J. (1884). On small differences in sensation. Mem. Natl. Acad. Sci. *3*, 75–83.

Pérez-Escudero, A., and de Polavieja, G.G. (2011). Collective animal behavior from Bayesian estimation and probability matching. PLoS Comput. Biol. *7*, e1002282.

Persaud, N., McLeod, P., and Cowey, A. (2007). Post-decision wagering objectively measures awareness. Nat. Neurosci. *10*, 257–261.

Pessiglione, M., Petrovic, P., Daunizeau, J., Palminteri, S., Dolan, R.J., and Frith, C.D. (2008). Subliminal instrumental conditioning demonstrated in the human brain. Neuron *59*, 561–567.

Phillips, L.D., and Wright, C.N. (1977). Cultural Differences in Viewing Uncertainty and Assessing Probabilities. In Decision Making and Change in Human Affairs, H. Jungermann and G.D. Zeeuw, eds. (Springer Netherlands), pp. 507–519.

Pleskac, T.J., and Busemeyer, J.R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. Psychol. Rev. *117*, 864–901.

Pouget, A., Beck, J.M., Ma, W.J., and Latham, P.E. (2013). Probabilistic brains: knowns and unknowns. Nat. Neurosci. *16*, 1170–1178.

Preuschoff, K., and Bossaerts, P. (2007). Adding prediction risk to the theory of reward learning. Ann. N Y Acad. Sci. *1104*, 135–146.

Preuschoff, K., Bossaerts, P., and Quartz, S.R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. Neuron *51*, 381–390.

Rahnev, D.A., Maniscalco, B., Luber, B., Lau, H., and Lisanby, S.H. (2012). Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. J. Neurophysiol. *107*, 1556–1563.

Rangel, A., Camerer, C., and Montague, P.R. (2008). A framework for studying the neurobiology of value-based decision making. Nat. Rev. Neurosci. *9*, 545–556.

Ratcliff, R. (1988). Continuous versus discrete information processing modeling accumulation of partial information. Psychol. Rev. *95*, 238–255.

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

Rich, D., Cazettes, F., Wang, Y., Peña, J.L., and Fischer, B.J. (2015). Neural representation of probabilities for Bayesian inference. J. Comput. Neurosci. *38*, 315–323.

Rose, M., Haider, H., and Büchel, C. (2005). Unconscious detection of implicit expectancies. J. Cogn. Neurosci. *17*, 918–927.

Schlaghecken, F., Stürmer, B., and Eimer, M. (2000). Chunking processes in the learning of event sequences: electrophysiological indicators. Mem. Cognit. *28*, 821–831.

Sergent, C., Baillet, S., and Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. Nat. Neurosci. *8*, 1391–1400.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., and Frith, C.D. (2014). Supra-personal cognitive control and metacognition. Trends Cogn. Sci. *18*, 186–193.

Smith, P.L., and Vickers, D. (1988). The accumulator model of two-choice discrimination. J. Math. Psychol. *32*, 135–168.

Smith, J.D., Shields, W.E., and Washburn, D.A. (2003). The comparative psychology of uncertainty monitoring and metacognition. Behav. Brain Sci. *26*, 317–339, discussion 340–373.

Soltani, A., and Wang, X.-J. (2010). Synaptic computation underlying probabilistic inference. Nat. Neurosci. *13*, 112–119.

Solway, A., and Botvinick, M.M. (2012). Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. Psychol. Rev. *119*, 120–154.

Summerfield, C., and de Lange, F.P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. Nat. Rev. Neurosci. *15*, 745–756.

Timmermans, B., Schilbach, L., Pasquali, A., and Cleeremans, A. (2012). Higher order thoughts in action: consciousness as an unconscious re-description process. Philos. Trans. R. Soc. Lond. B Biol. Sci. *367*, 1412–1423.

Todorov, E. (2004). Optimality principles in sensorimotor control. Nat. Neurosci. *7*, 907–915.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. Science *185*, 1124–1131.

van den Heuvel, M.P., Kahn, R.S., Goñi, J., and Sporns, O. (2012). High-cost, high-capacity backbone for global brain communication. Proc. Natl. Acad. Sci. USA *109*, 11372–11377.

van Zuijen, T.L., Simoens, V.L., Paavilainen, P., Näätänen, R., and Tervaniemi, M. (2006). Implicit, intuitive, and explicit knowledge of abstract regularities in a sound sequence: an event-related brain potential study. J. Cogn. Neurosci. *18*, 1292–1303.

Vickers, D., Burt, J., Smith, P., and Brown, M. (1985). Experimental paradigms emphasising state or process limitations: I effects on speed-accuracy tradeoffs. Acta Psychol. (Amst.) *59*, 129–161.

Vilares, I., Howard, J.D., Fernandes, H.L., Gottfried, J.A., and Kording, K.P. (2012). Differential representations of prior and likelihood uncertainty in the human brain. Curr. Biol. *22*, 1641–1648.

Vul, E., Goodman, N.D., Griffiths, T.L., and Tenenbaum, J.B. (2009). One and done? Optimal decisions from very few samples. In Proceedings of the 31st Annual Conference of the Cognitive Science Society, pp. 66–72.

Wald, A. (1945). Sequential tests of statistical hypotheses. Ann. Math. Stat. *16*, 117–186.

Wald, A., and Wolfowitz, J. (1948). Optimum character of the Sequential Probability Ratio Test. Ann. Math. Stat. *19*, 326–339.

Wallsten, T.S., and Budescu, D.V. (1995). A review of human linguistic probability processing: general principles and empirical evidence. Knowl. Eng. Rev. *10*, 43–62.

Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. Neuron *36*, 955–968.

Wolpert, D.M., and Ghahramani, Z. (2000). Computational principles of movement neuroscience. Nat. Neurosci. *3* (*Suppl*), 1212–1217.

Wu, S.-W., Delgado, M.R., and Maloney, L.T. (2009). Economic decision-making compared with an equivalent motor task. Proc. Natl. Acad. Sci. USA *106*, 6088–6093.

Yeung, N., and Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. Philos. Trans. R. Soc. Lond. B Biol. Sci. *367*, 1310–1321.

Yokoyama, O., Miura, N., Watanabe, J., Takemoto, A., Uchida, S., Sugiura, M., Horie, K., Sato, S., Kawashima, R., and Nakamura, K. (2010). Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in short-term recognition memory performance. Neurosci. Res. *68*, 199–206.

Yu, A.J., and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. Neuron *46*, 681–692.

Zylberberg, A., Fernández Slezak, D., Roelfsema, P.R., Dehaene, S., and Sigman, M. (2010). The brain's router: a cortical network model of serial processing in the primate brain. PLoS Comput. Biol. *6*, e1000765.

Zylberberg, A., Barttfeld, P., and Sigman, M. (2012). The construction of confidence in a perceptual decision. Front. Integr. Nuerosci. *6*, 79.

Zylberberg, A., Roelfsema, P.R., and Sigman, M. (2014). Variance misperception explains illusions of confidence in simple perceptual decisions. Conscious. Cogn. *27*, 246–253.