Cell Press

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