The Neurobiology of Confidence: From Beliefs to Neurons

Torben Ott,[1,4] Paul Masset,[1,2,3,4] and Adam Kepecs[1]

[1]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA*

[2]*Watson School of Biological Sciences, Cold Spring Harbor, New York 11724, USA*

[3]*Department of Molecular and Cellular Biology & Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138, USA*

Correspondence: kepecs@cshl.edu

How confident are you? As humans, aware of our subjective sense of confidence, we can readily answer. Knowing your level of confidence helps to optimize both routine decisions such as whether to go back and check if the front door was locked and momentous ones like finding a partner for life. Yet the inherently subjective nature of confidence has limited investigations by neurobiologists. Here, we provide an overview of recent advances in this field and lay out a conceptual framework that lets us translate psychological questions about subjective confidence into the language of neuroscience. We show how statistical notions of confidence provide a bridge between our subjective sense of confidence and confidence-guided behaviors in nonhuman animals, thus enabling the study of the underlying neurobiology. We discuss confidence as a core cognitive process that enables organisms to optimize behavior such as learning or resource allocation and that serves as the basis of metacognitive reasoning. These approaches place confidence on a solid footing and pave the way for a mechanistic understanding of how the brain implements confidence-based algorithms to guide behavior.

In a world that is often ambiguous and unpredictable, we nonetheless have a remarkable ability to form a range of beliefs based on different degrees of confidence about future possibilities, decisions, or events. Will I, for instance, benefit from reading this review? Will I make money by investing in the stock market? From an economic perspective, knowing the degree of certainty in a belief is fundamental to attaining our goals. How much time you should invest reading a review will depend on your assessment of how much useful information you might gain. How much money you should invest in the stock market will depend on your assessment of the economy and whether particular investments will thrive. But it is not just humans: When navigating uncertain environments, all living organisms face these kinds of challenges. In predicting the availability of food resources, animals must use ambiguous information to decide which option will lead to a richer resource or if they should abandon a chosen path. Survival requires having accurate estimates of confidence about each option. In this sense, confidence is an essential faculty to guide optimal behavior.

Yet we experience the sense of confidence as deeply subjective, generated by a process of apparent self-reflection termed metacognition (Flavell 1979; Dunlosky and Metcalfe 2009). Given this, could animals experience a similar sense of certainty? Can they think about their thoughts and report their confidence? And even if they could, how would one establish this rigorously without an explicit verbal self-report? Over the past decade, numerous neuroscientific studies have addressed many of these issues and this exciting research area has been thoroughly reviewed (Metcalfe 2008; Rolls et al. 2010; Kepecs and Mainen 2012; Middlebrooks and Sommer 2012; Shadlen and Kiani 2013; Fleming and Dolan 2014; Fleming and Lau 2014; Grimaldi et al. 2015; Meyniel et al. 2015b; Pouget et al. 2016), a task we will not attempt here. Rather, we outline a research program focusing on how neuroscientists can turn these deep but apparently squishy questions about confidence into neuroscientific ones and eventually provide answers in terms of neural circuit mechanisms.

THE TWO FACES OF CONFIDENCE: BELIEFS AND STATISTICS

Confidence has been extensively studied in multiple disciplines, primarily psychology and statistics, and the distinct conceptual frameworks of the two fields reveal confidence's dual manifestation as a subjective belief and an objective prediction. In psychology, confidence is often studied as a subjective feeling associated with beliefs about the world (Dunlosky and Metcalfe 2009; Kepecs and Mensh 2015). Introspection seems to be the only way to access this sense of confidence, and it is best communicated through verbalized self-reports. However, confidence is also a statistical quantity that can be defined as the likelihood that a belief is correct (Kepecs and Mainen 2012; Hangya et al. 2016; Pouget et al. 2016). This objective face of confidence measuring the degree of a belief can be precisely quantified using standard mathematical and statistical techniques. Indeed, confidence calculations lie at

[4]These two authors contributed equally.

Published by Cold Spring Harbor Laboratory Press; doi:10.1101/sqb.2018.83.038794

the foundation of statistical decision theory, machine learning, and hypothesis testing (Berger 1985).

But what is the relationship between the subjective sense of confidence in beliefs we experience as humans and the objective statistical computation that is also referred to as confidence? And because we are primarily interested in understanding the subjective sense of confidence, could we entirely avoid questions related to objective, statistical notions of confidence and simply focus on self-reported confidence measures? In other words, could we not just link human verbal reports of confidence to measures of neural activity and entirely sidestep issues related to statistical computations? Alas, confidence reports are often idiosyncratic: They are influenced by many factors, vary across contexts, and are inconsistent across individuals. For instance, the specific degrees of beliefs reported, whether low, high, or in-between levels of confidence, differs widely across individuals (Ais et al. 2016). The same decision situations can also lead to different confidence reports in one individual depending on context. At worst, self-reports do not provide useful predictions at all (Dunlosky and Metcalfe 2009). In these situations, identifying neural correlates of self-reported confidence would be expected to reveal a range of neural processes related to each individual's private notion of confidence and context. Therefore, although it is incumbent on any research on confidence to start from the human sense of confidence, there is a danger in taking confidence reports at face value, leading to a circular definition that makes it difficult to identify the neural underpinnings of confidence. But the challenge of linking subjective phenomenal experience with neural activity can be overcome by grounding self-reported confidence in objective statistical computations.

A THEORY OF STATISTICAL DECISION CONFIDENCE

How should a statistical theory of decision confidence be formulated? We can start by taking a closer look at mathematical formulations of generic decision-making processes. In common decision-making models, we base decisions on evidence from the environment—for instance, sensory stimuli. But because this evidence is often ambiguous, we use subjective percepts based on these external stimuli to form beliefs about the world. With a statistical definition of confidence, we can precisely quantify the likelihood that a decision is correct given our subjective perception of the evidence used to make that decision ("decision confidence") (Kepecs and Mainen 2012; Hangya et al. 2016; Fleming and Daw 2017). This definition of confidence is akin to a statistician calculating the probability that a hypothesis is correct given the observed evidence—hence, it is identical to statistical confidence. Importantly, using statistical decision confidence to guide future behavior enables optimal behavior; that is, it provides adaptive advantage in an uncertain world (Kepecs and Mainen 2012; Meyniel et al. 2015b; Pouget et al. 2016). Thus, a statistical theory of confidence provides a normative account of behavior that describes human and animal behavior as the result of an adaptive process and has fundamental implications for how confidence could be computed in neural circuits that represent sensory, cognitive, and motor variables (Knill and Pouget 2004; Körding and Wolpert 2004; Ma and Jazayeri 2014; Vasconcelos et al. 2017).

To see how this type of confidence estimate can be useful, consider a situation where you are driving on a foggy night with a broken navigation system to a restaurant you are unfamiliar with. You just passed a street sign you could hardly see—do you have to turn right or go straight? You made a decision based on what you could make out on the sign—for instance, you believe that the sign showed the street you were looking for and make the turn. How confident you are will depend on how well you could perceive the sign ("subjective evidence")—the less you could see, the less confident you are. As you keep driving on the street and the restaurant does not appear, how long you go before turning around depends on your confidence that the decision to turn onto the street was correct. Indeed, to optimize your search time, you should set the time invested into each turn according to your degree of confidence informed by the statistics of your available evidence.

Statistical decision confidence can be formally defined as the probability estimate that the chosen hypothesis is correct, given the evidence available to a subject—that is, $P(\text{correct}|\text{subjective evidence, choice})$ (Hangya et al. 2016). Here the subjective evidence can be any source of evidence contributing to a decision: perceptual, memory, or otherwise. However, the subjective evidence is internal to the decision-maker and cannot be directly experimentally observed or manipulated. How then can we hope to determine the contribution of a statistical confidence computation to behavior and neural signals? By leveraging our statistical model of decision confidence, we can construct precise predictions of optimal confidence (i.e., what an ideal observer would do to maximize success) for a range of variables that we can control or observe: the external evidence in the environment and the observed choices and outcomes.

This model of decision confidence yields several testable predictions about the relationship between optimal statistical confidence, evidence, and choice (Fig. 1A; Hangya et al. 2016). First, the degree of confidence predicts the fraction of correct choices—that is, choice accuracy, as intuitively expected ("calibration curve"). Second, statistical confidence increases with evidence strength for correct choices, but counterintuitively; for incorrect choices, confidence decreases with increasing evidence strength ("vevaiometric curves," from Greek *vevaios*, certain). Finally, although evidence strength determines accuracy (as expressed by a psychometric function), confidence provides further information improving the prediction of accuracy for any given level of evidence ("conditioned psychometric curves").

Under a set of moderate assumptions, these three signatures of decision confidence provide a powerful qualitative tool to determine if behavioral confidence reports (e.g., verbal self-reports) are informed by decision confidence and to delineate potential distinct contributions

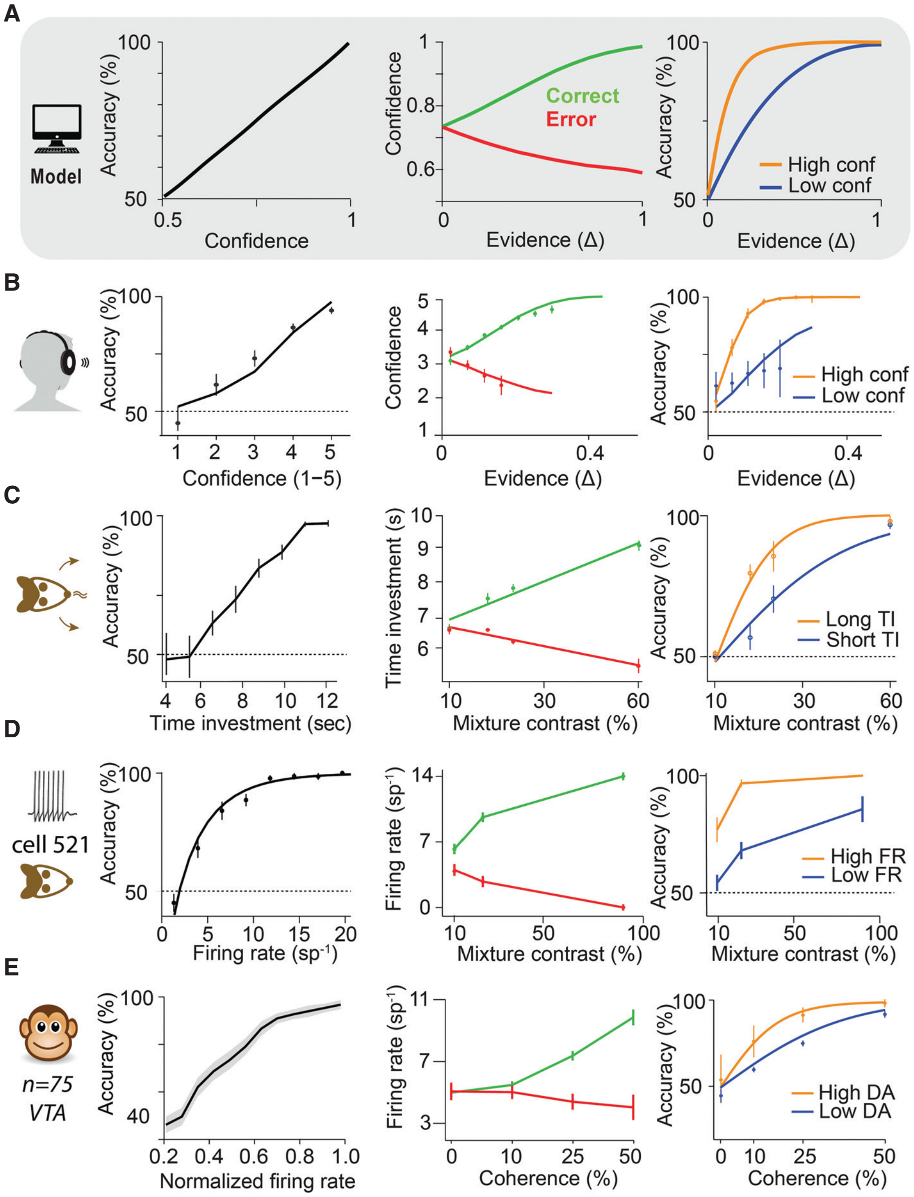

Figure 1. Behavioral and neural signatures predicted by a normative computational model of statistical decision confidence. (*A*) Model predictions of how confidence—the probability of being correct—relates to observable variables in a perceptual decision experiment in which subjects must make a binary choice based on available evidence. (*Left*) Calibration curve: Confidence predicts accuracy. (*Middle*) Vevaiometric curve: Confidence increases with increasing evidence strength in correct trials but decreases in error trials. (*Right*) Conditioned psychometric curve: Psychometric curve for high-confidence trials is steeper as compared to low-confidence trials (Hangya et al. 2016). (*B*) Verbal self-reports of confidence in humans (points) follow all key signatures of statistical decision confidence (lines) (Sanders et al. 2016). (*C*) Time investment (TI) behavior in rats (points) follows all key signatures of statistical decision confidence (Lak et al. 2014). (*D*) A single neuron in rat orbitofrontal cortex encodes statistical decision confidence (Hirokawa et al. 2017). (FR) Firing rate. (*E*) Dopamine (DA) neurons in primate ventral tegmental area (VTA) encode statistical decision confidence (Lak et al. 2017).

related to evidence or choice (Hangya et al. 2016). The assumptions amount to decision situations where there are right and wrong choices about a potential outcome and an explicit model of the internal, noisy representation of the evidence that determines the choice. The precise shape of the confidence signatures can depend on the specific decision model used and the structure of the environment (Kiani et al. 2014; King and Dehaene 2014; Adler and Ma 2018; Rausch and Zehetleitner 2018). Nevertheless, an ideal observer model provides a solid starting point for formal investigations of decision confidence.

This framework enables a comparison of optimal, model-predicted confidence estimates with observed confidence signals, whether these are human verbal confidence reports or neurons putatively signaling confidence. Importantly, this approach provides a way to quantify the degree to which confidence reports are informed by or depart from optimal statistical confidence computation.

FROM HUMANS TO ANIMALS: STATISTICAL CONFIDENCE INFORMS BEHAVIOR

How can we tell whether our sense of confidence is informed by an objective, statistical computation of confidence? By precisely controlling the external evidence and leveraging the normative theory of confidence, human self-reports of subjective confidence can be compared with statistical decision confidence. Under controlled conditions—by systematically varying choice difficulty and analyzing choice patterns—subjective confidence reports followed the statistical predictions in sensory and general-knowledge tasks (Fig. 1B), thus linking objective and subjective notions of confidence (Sanders et al. 2016). Human self-reports were remarkably close to optimal confidence levels predicted by the model, providing a framework to assess if and how other factors corrupt the relationship between statistical confidence and subjective reports of confidence.

How can we ask nonverbal species to report their confidence? The key idea is to incentivize the use of confidence by creating situations in which using confidence information benefits the subject. For instance, we can precisely control ambiguous evidence available to a subject and provide an opportunity for it make an investment in its decision, time, or effort, in order to earn a reward. In this approach, the "investment" provided is an implicit confidence report that can be analyzed using the normative theory of confidence in the same way one would use explicit self-reports. Our laboratory has developed a post–decision time investment task in which animals place bets on difficult decisions by how long they are willing to wait for an uncertain reward (Kepecs et al. 2008; Lak et al. 2014). In these tasks, rats decide between two choices based on noisy sensory information ("evidence") in order to obtain a reward (Kepecs et al. 2008; Hirokawa et al. 2017). Reward delivery is randomly delayed, and no feedback is provided when the subjects make an incorrect choice. Thus, rats earn a drop of water for correct choices but must invest time in waiting for its arrival. Alternatively, a rat could decide to start a new trial. A single trial can hence provide a continuous measure of time investment that can be quantitatively related to the amount of sensory information and the choice. We showed that time investment, reflecting the rats' willingness to wait for an uncertain reward, follows the three signatures of the normative theory of confidence (Kepecs and Mainen 2012; Lak et al. 2014). Rats use statistical confidence to decide how much time to invest in decisions, thus providing a nonverbal readout of their subjective confidence levels (Fig. 1C).

This approach to study confidence without verbal reports makes it possible to ask whether infants are capable of estimating their confidence or if this ability develops only later in life. This question has been a challenging problem for developmental psychology because the delayed ability of verbal expression precluded testing preverbal toddlers and infants. Using a similar behavioral paradigm and quantitative approach as in the rat studies, preverbal infants were shown to guide their behavior based on confidence. One-year-old babies persisted longer in their attempts to find hidden toys for correct choices than for error choices, which corresponded with predictions of the statistical confidence model (Goupil et al. 2016; Goupil and Kouider 2016). These studies show that the statistical framework can help us study decision confidence without having to rely on verbal reports. Extending this approach, we are now ready to ask how neural systems realize confidence computations that guide these behaviors.

HOW TO FIND A CONFIDENCE NEURON?

By operationalizing decision confidence as a neural computation, we can use the tools of neuroscience to search for neurons coding for confidence. Our current understanding of how neural sensory and motor circuits support perception and action fundamentally builds on conceptualizing the function of neural circuits as realizing specific computations. Neurons in cortical areas are often characterized by their response properties related to changes in the environment (i.e., their "tuning curve"). For instance, neurons in the primary visual cortex are characterized as edge detectors because they preferentially respond to moving bars at a specific orientation over another (Hubel and Wiesel 1962). We can understand the transformation of retinal information along the cortical hierarchy into more and more complex features like edges or objects as a series of computational operations such as filtering, amplification, or normalization (Heeger et al. 1996; Gollisch and Meister 2010; Carandini and Heeger 2012; Yamins and DiCarlo 2016). By identifying the tuning curves for a range of complex variables, from faces to places, neuroscience has begun to determine the computations relevant for numerous brain regions. Analogously, we can use the normative model of statistical confidence to search for "confidence-tuning curves." This lets us ask whether the neuronal responses are consistent with these confidence-tuning properties and establish whether they are informed by a confidence computation (Fig. 2, left). The confidence-tuning properties can be further characterized by evaluating their invariance to contextual changes such as decisions based on a range of sensory information from different sensory modalities. This approach has been used to find "confidence-tuned" neurons in rat orbitofrontal cortex (Kepecs et al. 2008; Hirokawa et al. 2017) and primate pulvinar (Komura et al. 2013). These studies revealed that confidence computations contributed to neural activity and that other computations such as reinforcement learning could be ruled out as producing this neural activity (Fig. 2, left; Kepecs et al. 2008).

In our quest to identify neurons that represent decision confidence, "confidence neurons," we need to additionally ask how this neural activity is related to confidence-guided behaviors. Determining if neural activity or any other signal is informed by a confidence computation is not sufficient to establish its role in behavior. For instance, sensory cortical neurons (Britten et al. 1993) or even physiological signals such as pupil dilation (Urai et al. 2017; Kawaguchi et al. 2018) can also be informed by a

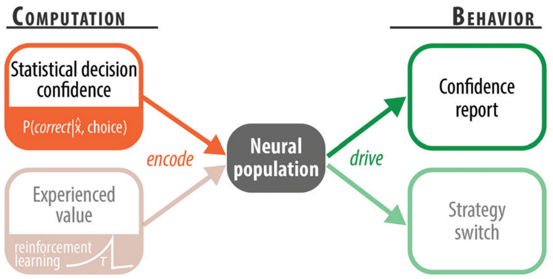

Figure 2. Neuronal representations viewed as the result of a computational process and as driving behavior. We can characterize a neuron or neural population by determining the computational processes informing its activity (*left* side). This amounts to constructing tuning curves for the sensory or cognitive variable in question under varying contexts. Different computational processes, including statistical decision confidence (the probability of being correct given subject evidence $\hat{x}$ and choice) or experienced value (computed by reinforcement learning) that are part of a computational model, can influence neural activity. Another consideration is how neural activity drives behavior, such as reports of confidence (e.g., verbal reports or time investments) or choice strategies (e.g., switching behavior) (*right* side). The neuron's function can be understood by considering the encoding process, the computation that describes its activity, and its relationship to behavior. Note that the same neurons could be informed by additional computational processes and drive other kinds of behaviors besides the ones under consideration.

confidence computation. To probe a neuron's role in confidence-guided behavior, we also need to establish that its neural activity supports confidence reporting behavior by correlating neural activity with time investment on a trial-to-trial basis and causally manipulating its activity in order to change behavior. Because the same confidence signal could be expected to drive a multitude of behaviors, including time investment or strategy switching (Fig. 2, right), we need to determine whether a particular neural signal predicts confidence-guided behavior such as explicit confidence reports, investment decisions, or other behaviors informed by confidence like changes of strategy (Meyniel et al. 2015a). For instance, neurons in the parietal cortex (Kiani and Shadlen 2009) and pulvinar (Komura et al. 2013) predict opt-out choices based on confidence. We then need to manipulate the activity of these specific neurons to directly probe their contribution to confidence-guided behavior. Finally, a confidence neuron is expected to respond irrespective of the source of information (e.g., invariant across sensory modalities contributing to a decision). Together, these criteria serve to identify "confidence neurons" by understanding their function as realizing a confidence computation and driving specific confidence-guided behaviors.

COGNITIVE CELL TYPES: PLACE, FACE, AND CONFIDENCE

The notion of a "confidence neuron" is akin to other success stories in systems neuroscience that identified key functions of specific neurons each instantiating different neural computation. For example, "face cells" in the primate temporal cortex selectively respond to faces or features of faces but are invariant to many other features like brightness or orientation (Tsao 2014; Freiwald et al. 2016; Yamins and DiCarlo 2016). "Place cells" and "grid cells" in rodent hippocampus and entorhinal cortex are active only at specific spatial locations because they compute spatial location based on sensory and mnemonic information (O'Keefe and Dostrovsky 1971; Moser et al. 2008). In the same sense, establishing that a particular neural tuning curve reflects a confidence computation (Kepecs et al. 2008) establishes that it represents decision confidence.

One might question whether these neurons should be called "confidence" neurons rather than "anxiety" or "arousal" or even "attention" neurons. Indeed, relating these squishy subjective experiences to neurons has been a major challenge for cognitive neuroscience (Anderson and Adolphs 2014). The solution offered by computational models is that they formalize cognitive processes as a series of well-defined computations with variables that serve as proxies for those that are unobservable (e.g., confidence, value, or attention) (Sugrue et al. 2005; Daw and Doya 2006; Corrado and Doya 2007; Doya 2008; Reynolds and Heeger 2009; Carrasco 2011). By systematically probing observable and controllable parameters—such as sensory variables—and evaluating decision patterns using computational models, we can infer unobservable or internal variables explaining behavior and relate these to the activity of neurons in the brain. Hence, our interpretations do not hinge on the subjective notion of confidence but rather on the computational process that best explains neural firing patterns. This way the label "confidence" refers to the computation required and the requisite class of computational models that can be tested and iteratively improved, scientifically grounding the process of identifying neural representations.

NEURAL MECHANISM OF CONFIDENCE

Where should we look for confidence neurons in the brain? We have said that estimating confidence is a fundamental computation providing important summary statistics that guide behavior. If confidence estimates are central to behavior, they should be present across many brain regions. Using the conceptual framework to identify neural signatures of confidence, we and others have recorded single neurons coding for at least some aspects of decision confidence in the orbitofrontal cortex (Kepecs et al. 2008; Hirokawa et al. 2017), parietal cortex (Kiani and Shadlen 2009; Rutishauser et al. 2018), pulvinar (Komura et al. 2013), and prefrontal cortex (Middlebrooks and Sommer 2012; Teichert et al. 2014).

Recent and ongoing work from our laboratory has identified the orbitofrontal cortex in rodents as a key hub for confidence computations (Kepecs et al. 2008; Lak et al. 2014; Hirokawa et al. 2017). This work builds on two pillars outlined above: a normative theory of decision confidence and a postdecision wagering task that allows us to obtain a nonverbal behavioral measure of decision confidence—the time invested in a decision. Single-neuron recordings in rat orbitofrontal cortex revealed that the activity of a group of neurons closely followed the predictions from the normative theory of confidence (Fig. 1D). The activity of these neurons directly after making a choice predicted if rats made a mistake, even with the same amount of sensory evidence, and the activity was graded with average accuracy before any feedback about choice-correctness was provided (Kepecs et al. 2008). Confidence signals also contribute to value representations because they are an estimate of the outcome probability. Therefore to distinguish whether the neural activity we observed specifically encodes confidence, we systematically varied reward size for different choices. Interestingly, distinct groups of neurons encode decision confidence (irrespective of reward size) and outcome value (confidence combined with reward size) (Hirokawa et al. 2017). Recently, we have also established that these neuronal signatures of confidence are invariant across sensory modalities and that single-neuron activity predicted confidence-guided time investment behavior on a trial-to-trial basis.

Pharmacologically silencing the orbitofrontal cortex produced a specific impairment of investment behavior: Choice behavior and average investment time were unaffected, but time investment was no longer informed by statistical confidence (i.e., all relations predicted by the normative theory of confidence broke down). These results suggest that orbitofrontal cortex plays a key role in estimating decision confidence, along with other areas—like the pulvinar (Komura et al. 2013) and frontal pole (Miyamoto et al. 2018)—that have been found to contribute to decision confidence in primates. In other brain regions, confidence estimates can contribute to other computations, such as prediction error signals in the midbrain dopamine neurons involved in reinforcement learning (Fig. 1E; Lak et al. 2017).

These results bring us close to identifying confidence neurons—neurons characterized by a confidence computation. It is important to note, however, that these neurons would be expected to drive a range of confidence-guided behaviors beyond time investment—"anxiety" or "curiosity," for instance, as both these concepts are related to specific uses of uncertainty information (Gottlieb et al. 2013).

OUTLOOK: QUANTIFYING CONFIDENCE FROM CIRCUITS TO METACOGNITION

The subjective nature of confidence presents a challenge for its scientific study. Here we have described an approach that leverages advances of computational theories to establish the relationship between subjective confidence, objective statistical mental computations, and neural activity. By grounding confidence as a normative computational process, it becomes possible to evaluate whether a subjectively experienced sense of confidence is related to an objectively computed estimate of confidence, providing a bridge between the two. This framework has been used to show that self-reported confidence in a range of decision tasks follows the normative statistical predictions (Sanders et al. 2016; Lebreton et al. 2018), and that preverbal infants have the ability to act on their confidence (Goupil and Kouider 2016), as do rats investing time into perceptual decisions (Lak et al. 2014). Indeed, the capacity to estimate and deploy confidence is an adaptive process that maximizes fitness in uncertain environments. Based on these considerations, we expect that computing confidence is a widespread ability across the animal kingdom and that, because of its broad computational role, it is realized by diverse neural systems.

Placing confidence on solid footing allows us to use this framework to identify its neural basis. The operationalization of decision confidence as a cognitive computation provides the means to identify the neural circuit mechanisms analogous to the approaches used by sensory and motor neuroscience. It has become clear that confidence is represented in a number of brain regions, consistent with theoretical ideas about the centrality of uncertainty in neural computations (Beck et al. 2008; Denève et al. 2017). Finally, inactivation studies have revealed that the orbitofrontal and frontopolar cortex have central roles in confidence reports (Lak et al. 2014; Miyamoto et al. 2018). These results provide a strong foundation for identifying neural circuit mechanisms underlying confidence estimation and deployment for different behaviors. Further progress will likely reveal how specific neuron types contribute to the algorithms underlying confidence computations.

The orbitofrontal cortex contains a centralized confidence representation that is thought be a prerequisite for metacognition—explicitly reasoning about one's own beliefs (Dehaene et al. 2017). Quantifying the "degree of metacognition" is an area of active research in psychology (Fleming and Lau 2014; Sherman et al. 2018). Previous suggestions that metacognition could be quantified via signal detection theory rely on strong assumptions and are often limited in their application (see the following back and forth discussion: Rounis et al. 2010; Bor et al. 2017, 2018; Ruby et al. 2018). Our normative theory of confidence offers the advantage of quantifying an optimal use of statistical confidence defined using a generative model. This approach allows us to dissociate choice sensitivity from metacognitive ability and offers a path toward investigating how metacognitive ability differs across individuals, thus seemingly offering a way to predict psychological traits or psychopathologies (Rouault et al. 2018). Indeed, how well-calibrated confidence estimates are (i.e., how well they predict accuracy) varies across subjects (Björkman et al. 1993; Olsson and Winman 1996; Moore and Healy 2008; Shea et al. 2014; Ais et al. 2016), and in some cases confidence can be low irrespective of objective probability (i.e., statistical confidence), a hallmark of underconfidence in anxiety disorders such as obsessive–compulsive disorder (Barahmand et al. 2014). This framework to measure confidence in humans can be also applied as a method in computational psychiatry to quantify how confidence computations are impacted in these disorders (Montague et al. 2012; Kepecs and Mensh 2015).

Our approach thus roots the subjective experience of confidence and metacognitive ability in a precisely defined computation, a statistical theory of confidence. This framework enables us to use the tools of a neuroscientist to address psychological questions. New experiments will now be able to identify neural circuit mechanisms for how the brain implements confidence-based algorithms to guide behavior. We anticipate that a mechanistic understanding of confidence and metacognition will bridge a psychological understanding of the mind and its disorders with its neurobiological basis in human and nonhuman animals.

REFERENCES

Ais J, Zylberberg A, Barttfeld P, Sigman M. 2016. Individual consistency in the accuracy and distribution of confidence judgments. *Cognition* **146:** 377–386. doi:10.1016/j.cognition.2015.10.006

Adler WT, Ma WJ. 2018. Limitations of proposed signatures of Bayesian confidence. *Neural Comput* **30:** 3327–3354.

Anderson DJ, Adolphs R. 2014. A framework for studying emotions across species. *Cell* **157:** 187–200.

Barahmand U, Tavakolian E, Alaei S. 2014. Association of metacognitive beliefs, obsessive beliefs and symptom severity with quality of life in obsessive–compulsive patients. *Arch Psychiatr Nurs* **28:** 345–351. doi:10.1016/j.apnu.2014.08.005

Beck JM, Ma WJ, Kiani R, Hanks T, Churchland AK, Roitman J, Shadlen MN, Latham PE, Pouget A. 2008. Probabilistic population codes for Bayesian decision making. *Neuron* **60:** 1142–1152. doi:10.1016/j.neuron.2008.09.021

Berger JO. 1985. *Statistical decision theory and Bayesian analysis*. Springer, New York.

Björkman M, Juslin P, Winman A. 1993. Reply to William R. Ferrell's paper "A model for realism of confidence judgments: implications for underconfidence in sensory discrimination". *Percept Psychophys* **57:** 255–259. doi:10.3758/BF03206512

Bor D, Schwartzman DJ, Barrett AB, Seth AK. 2017. Theta-burst transcranial magnetic stimulation to the prefrontal or parietal cortex does not impair metacognitive visual awareness. *PLoS ONE* **12:** e0171793. doi:10.1371/journal.pone.0171793

Bor D, Barrett AB, Schwartzman DJ, Seth AK. 2018. Response to Ruby et al.: on a 'failed' attempt to manipulate conscious perception with transcranial magnetic stimulation to prefrontal cortex. *Conscious Cogn* **65:** 334–341. doi:10.1016/j.concog.2018.07.011

Britten KH, Shadlen MN, Newsome WT, Movshon JA. 1993. Responses of neurons in macaque MT to stochastic motion signals. *Vis Neurosci* **10:** 1157–1169. doi:10.1017/S0952523800010269

Carandini M, Heeger DDJ. 2012. Normalization as a canonical neural computation. *Nat Rev Neurosci* **13:** 51–62. doi:10.1038/nrn3136

Carrasco M. 2011. Visual attention: the past 25 years. *Vision Res* **51:** 1484–1525. doi:10.1016/j.visres.2011.04.012

Corrado G, Doya K. 2007. Understanding neural coding through the model-based analysis of decision making. *J Neurosci* **27:** 8178–8180. doi:10.1523/JNEUROSCI.1590-07.2007

Daw ND, Doya K. 2006. The computational neurobiology of learning and reward. *Curr Opin Neurobiol* **16:** 199–204. doi:10.1016/j.conb.2006.03.006

Dehaene S, Lau H, Kouider S. 2017. What is consciousness, and could machines have it? *Science* **358:** 486–492. doi:10.1126/science.aan8871

Denève S, Alemi A, Bourdoukan R. 2017. The brain as an efficient and robust adaptive learner. *Neuron* **94:** 969–977. doi:10.1016/j.neuron.2017.05.016

Doya K. 2008. Modulators of decision making. *Nat Neurosci* **11:** 410–416. doi:10.1038/nn2077

Dunlosky J, Metcalfe J. 2009. *Metacognition*. Sage, Newbury Park, CA.

Flavell JH. 1979. Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. *Am Psychol* **34:** 906–911. doi:10.1037/0003-066X.34.10.906

Fleming SM, Daw ND. 2017. Self-evaluation of decision-making: a general Bayesian framework for metacognitive computation. *Psychol Rev* **124:** 91–114. doi:10.1037/rev0000045

Fleming SM, Dolan RJ. 2014. The neural basis of metacognitive ability. In *The cognitive science of metacognition* (ed. Fleming SM, Frith CD), pp. 245–265. Springer-Verlag, New York. doi:10.1007/978-3-642-45190-4_11

Fleming SM, Lau HC. 2014. How to measure metacognition. *Front Hum Neurosci* **8:** 443.

Freiwald W, Duchaine B, Yovel G. 2016. Face processing systems: from neurons to real-world social perception. *Annu Rev Neurosci* **39:** 325–346. doi:10.1146/annurev-neuro-070815-013934

Gollisch T, Meister M. 2010. Eye smarter than scientists believed: neural computations in circuits of the retina. *Neuron* **65:** 150–164. doi:10.1016/j.neuron.2009.12.009

Gottlieb J, Oudeyer PY, Lopes M, Baranes A. 2013. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends Cogn Sci* **17:** 585–593. doi:10.1016/j.tics.2013.09.001

Goupil L, Kouider S. 2016. Behavioral and neural indices of metacognitive sensitivity in preverbal infants. *Curr Biol* **26:** 3038–3045. doi:10.1016/j.cub.2016.09.004

Goupil L, Romand-Monnier M, Kouider S. 2016. Infants ask for help when they know they don't know. *Proc Natl Acad Sci* **113:** 3492–3496. doi:10.1073/pnas.1515129113

Grimaldi P, Lau H, Basso MA. 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. doi:10.1016/j.neubiorev.2015.04.006

Hangya B, Sanders JI, Kepecs A. 2016. A mathematical framework for statistical decision confidence. *Neural Comput* **28:** 1840–1858. doi:10.1162/NECO_a_00864

Heeger DJ, Simoncelli EP, Movshon JA. 1996. Computational models of cortical visual processing. *Proc Natl Acad Sci* **93:** 623–627. doi:10.1073/pnas.93.2.623

Hirokawa J, Vaughan A, Kepecs A. 2017. Categorical representations of decision-variables in orbitofrontal cortex. *bioRxiv* doi:10.1101/135707

Hubel DH, Wiesel TN. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* **160:** 106–154. doi:10.1113/jphysiol.1962.sp006837

Kawaguchi K, Clery S, Pourriahi P, Seillier L, Haefner XM, Nienborg XH. 2018. Differentiating between models of perceptual decision making using pupil size inferred confidence. *J Neurosci* **38:** 8874–8888. doi:10.1523/JNEUROSCI.0735-18.2018

Kepecs A, Mainen ZF. 2012. A computational framework for the study of confidence in humans and animals. *Philos Trans R Soc B Biol Sci* **367:** 1322–1337. doi:10.1098/rstb.2012.0037

Kepecs A, Mensh BD. 2015. Emotor control: computations underlying bodily resource allocation, emotions, and confidence. *Dialogues Clin Neurosci* **17:** 391–401.

Kepecs A, Uchida N, Zariwala HA, Mainen ZF. 2008. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455:** 227–231. doi:10.1038/nature07200

Kiani R, Shadlen MN. 2009. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324:** 759–764. doi:10.1126/science.1169405

Kiani R, Corthell L, Shadlen MN. 2014. Choice certainty is informed by both evidence and decision time. *Neuron* **84:** 1329–1342. doi:10.1016/j.neuron.2014.12.015

King J-R, Dehaene S. 2014. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn Sci* **18:** 203–210. doi:10.1016/j.tics.2014.01.002

Knill DC, Pouget A. 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci* **27:** 712–719. doi:10.1016/j.tins.2004.10.007

Komura Y, Nikkuni A, Hirashima N, Uetake T, Miyamoto A. 2013. Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nat Neurosci* **16:** 749–755. doi:10.1038/nn.3393

Körding KP, Wolpert DM. 2004. Bayesian integration in sensorimotor learning. *Nature* **427:** 244–247. doi:10.1038/nature02169

Lak A, Costa GM, Romberg E, Koulakov AA, Mainen ZF, Kepecs A. 2014. Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* **84:** 190–201. doi:10.1016/j.neuron.2014.08.039

Lak A, Nomoto K, Keramati M, Sakagami M, Kepecs A. 2017. Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Curr Biol* **27:** 821–832. doi:10.1016/j.cub.2017.02.026

Lebreton M, Langdon S, Slieker MJ, Nooitgedacht JS, Goudriaan AE, Denys D, van Holst RJ, Luigjes J, van Holst RJ. 2018. Two sides of the same coin: monetary incentives concurrently improve and bias confidence judgments. *Sci Adv* **4:** eaaq0668.

Ma WJ, Jazayeri M. 2014. Neural coding of uncertainty and probability. *Annu Rev Neurosci* **37:** 205–220. doi:10.1146/annurev-neuro-071013-014017

Metcalfe J. 2008. Evolution of metacognition. In *Handbook of metamemory and memory* (ed. Dunlosky J, Bjork RA), pp. 29–46. Psychology Press, New York.

Meyniel F, Schlunegger D, Dehaene S. 2015a. The sense of confidence during probabilistic learning: a normative account. *PLoS Comput Biol* **11:** 1004305. doi:10.1371/journal.pcbi.1004305

Meyniel F, Sigman M, Mainen ZF. 2015b. Confidence as Bayesian probability: from neural origins to behavior. *Neuron* **88:** 78–92. doi:10.1016/j.neuron.2015.09.039

Middlebrooks PG, Sommer MA. 2012. Neuronal correlates of metacognition in primate frontal cortex. *Neuron* **75:** 517–530. doi:10.1016/j.neuron.2012.05.028

Miyamoto K, Setsuie R, Osada T, Miyashita Y. 2018. Reversible silencing of the frontopolar cortex selectively impairs metacognitive judgment on non-experience in primates. *Neuron* **97:** 980–989.e6. doi:10.1016/j.neuron.2017.12.040

Montague PR, Dolan RJ, Friston KJ, Dayan P. 2012. Computational psychiatry. *Trends Cogn Sci* **16:** 72–80. doi:10.1016/j.tics.2011.11.018

Moore DA, Healy PJ. 2008. The trouble with overconfidence. *Psychol Rev* **115:** 502–517. doi:10.1037/0033-295X.115.2.502

Moser EI, Kropff E, Moser M-B. 2008. Place cells, grid cells, and the brain's spatial representation system. *Annu Rev Neurosci* **31:** 69–89. doi:10.1146/annurev.neuro.31.061307.090723

O'Keefe J, Dostrovsky J. 1971. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res* **34:** 171–175. doi:10.1016/0006-8993(71)90358-1

Olsson H, Winman A. 1996. Underconfidence in sensory discrimination: the interaction between experimental setting and response strategies. *Percept Psychophys* **58:** 374–382. doi:10.3758/BF03206813

Pouget A, Drugowitsch J, Kepecs A. 2016. Confidence and certainty: distinct probabilistic quantities for different goals. *Nat Neurosci* **19:** 366–374. doi:10.1038/nn.4240

Rausch M, Zehetleitner M. 2018. Multiple statistical signatures of human confidence. *bioRxiv* doi:10.1101/426635

Reynolds JH, Heeger DJ. 2009. The normalization model of attention. *Neuron* **61:** 168–185. doi:10.1016/j.neuron.2009.01.002

Rolls ET, Grabenhorst F, Deco G. 2010. Choice, difficulty, and confidence in the brain. *Neuroimage* **53:** 694–706. doi:10.1016/j.neuroimage.2010.06.073

Rouault M, Seow T, Gillan CM, Fleming SM. 2018. Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biol Psychiatry* **84:** 443–451. doi:10.1016/j.biopsych.2017.12.017

Rounis E, Maniscalco B, Rothwell JC, Passingham RE, Lau H. 2010. Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cogn Neurosci* **1:** 165–175. doi:10.1080/17588921003632529

Ruby E, Maniscalco B, Peters MAK. 2018. On a 'failed' attempt to manipulate visual metacognition with transcranial magnetic stimulation to prefrontal cortex. *Conscious Cogn* **62:** 34–41. doi:10.1016/j.concog.2018.04.009

Rutishauser U, Aflalo T, Rosario ER, Pouratian N, Andersen RA. 2018. Single-neuron representation of memory strength and recognition confidence in left human posterior parietal cortex. *Neuron* **97:** 209–220.e3. doi:10.1016/j.neuron.2017.11.029

Sanders JI, Hangya B, Kepecs A. 2016. Signatures of a statistical computation in the human sense of confidence. *Neuron* **90:** 499–506. doi:10.1016/j.neuron.2016.03.025

Shadlen MN, Kiani R. 2013. Decision making as a window on cognition. *Neuron* **80:** 791–806. doi:10.1016/j.neuron.2013.10.047

Shea N, Boldt A, Bang D, Yeung N, Heyes C, Frith CD. 2014. Supra-personal cognitive control and metacognition. *Trends Cogn Sci* **18:** 186–193. doi:10.1016/j.tics.2014.01.006

Sherman MT, Seth AK, Barrett A. 2018. Quantifying metacognitive thresholds using signal-detection theory. *bioRxiv* doi:10.1101/361543

Sugrue LP, Corrado GS, Newsome WT. 2005. Choosing the greater of two goods: neural currencies for valuation and decision making. *Nat Rev Neurosci* **6:** 363–375. doi:10.1038/nrn1666

Teichert T, Yu D, Ferrera VP. 2014. Performance monitoring in monkey frontal eye field. *J Neurosci* **34:** 1657–1671. doi:10.1523/JNEUROSCI.3694-13.2014

Tsao D. 2014. The macaque face patch system: a window into object representation. *Cold Spring Harb Symp Quant Biol* **79:** 109–114. doi:10.1101/sqb.2014.79.024950

Urai AE, Braun A, Donner TH. 2017. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nat Commun* **8:** 1–11. doi:10.1038/s41467-016-0009-6

Vasconcelos M, Fortes I, Kacelnik A. 2017. On the structure and role of optimality models in the study of behavior. In *APA handbook of comparative psychology: perception, learning, and cognition* (eds. Call J, et al.), pp. 287–307. American Psychological Association, Washington, DC.

Yamins DLK, DiCarlo JJ. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci* **19:** 356–365. doi:10.1038/nn.4244