Annual Review of Vision Science

Computational and Neuronal Basis of Visual Confidence

Robbe L.T. Goris,[1] Zhongzheng Fu,[2] and Christopher R. Fetsch[3]

[1]Center for Perceptual Systems, University of Texas at Austin, Austin, Texas, USA; email: robbe.goris@utexas.edu

[2]Department of Neurological Surgery, University of Texas Southwestern Medical Center, Dallas, Texas, USA

[3]Department of Neuroscience and Zanvyl Krieger Mind/Brain Institute, Johns Hopkins University, Baltimore, Maryland, USA

Annu. Rev. Vis. Sci. 2025. 11:385–410

The *Annual Review of Vision Science* is online at vision.annualreviews.org

https://doi.org/10.1146/annurev-vision-110323-120909

Keywords

neural coding, perceptual decision-making, visual metacognition

Abstract

The primate brain excels at transforming photons into knowledge. When light strikes the back of the eye, opsin molecules within rods and cones absorb photons, triggering a change in membrane potential. This energy transfer initiates a cascade of neural events that endows us with useful knowledge. This knowledge manifests as subjectively experienced perceptual interpretations and mostly pertains to the 3D structure of the visual environment and the affordances of the objects within the scene. However, some of this knowledge instead pertains to the quality of these interpretations and contributes to our sense of confidence in perceptual decisions. Because such confidence reflects knowledge about knowledge, psychologists consider this the domain of metacognition. Here, we examine what is known about the neuronal basis of perceptual decision confidence, with a focus on vision. We review the crucial computational processes and neural operations that underlie and constrain the transformation of photons into visual metacognition.

INTRODUCTION

Humans and other animals operate in a world that cannot be known perfectly. Raw visual inputs give rise to patterns of neural activity in the retina that by themselves are too coarse (Strasburger et al. 2011), noisy (Croner et al. 1993), and indirect (Knill & Richards 1996) to unambiguously reveal the state of the environment. It follows that perceptual interpretations of the world are the brain's best momentary guess, selected from a large set of candidate world states (von Helmholtz 1948, Knill & Richards 1996). A rich body of research has found that these interpretations arise from computational strategies that are sophisticated (Glaze et al. 2015, Purcell & Kiani 2016), flexible (Ernst & Banks 2002, Adams et al. 2004, Młynarski & Hermundstad 2018, Norton et al. 2019, Charlton et al. 2023), and principled (Weiss et al. 2002, Stocker & Simoncelli 2006, Kilpatrick et al. 2019, Hahn & Wei 2024). Despite this, errors are inevitable. There simply exists no scenario under which perception of a complex environment can consistently be error free. Perceptual misjudgments are sometimes innocuous. For example, fruit in a faraway tree may appear ripe from a distance, while closer inspection reveals that it is not. In this case, the cost associated with the behavior guided by the perceptual misinterpretation (i.e., checking out the tree to no avail) is limited. However, perceptual misjudgments can sometimes result in risky behavior that is extremely costly. A car in the distance may appear to be slowing down, but if this interpretation turns out to be wrong, the decision to cross the street could result in a collision. As these examples illustrate, to accomplish goals, the brain needs to do more than just identify the most plausible state of the environment. It also ought to keep track of the certainty of these interpretations (von Helmholtz 1948, Knill & Pouget 2004, Pouget et al. 2016). And it does. This process contributes to our sense of confidence in perceptual decisions. Although much of the field has adopted a consensus definition of confidence as the subjective probability that a decision is correct (Pouget et al. 2016), here we follow recent work that has shown that it is more accurate to associate confidence with subjective decision reliability (Koriat 2012, Li & Ma 2020, Caziot & Mamassian 2021, Boundy-Singer et al. 2023). In other words, confidence seems to reflect the belief that we would make the same choice again, if we had to make the decision a second time.

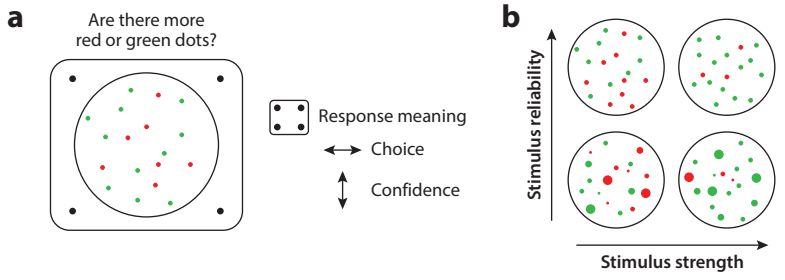
Humans can be acutely aware of the confidence they have in a perceptual decision or proposition. We experience noticeable doubt when we wonder whether we recognize a movie actor or whether a handwritten digit is a 3 or an 8. This awareness enables us to communicate the reliability of our judgments to others, which in turn can play a vital role in optimizing group decision-making (Bahrami et al. 2010). The ability to verbalize perceptual confidence is uniquely human. However, that does not imply that other animal species lack knowledge about the limits of perception. There is now ample evidence that nonhuman animals such as monkeys and rats exhibit confidence-mediated behavior (reviewed in Smith 2009, Kepecs & Mainen 2012). Fortunately, this capacity can be studied in binary perceptual decision tasks, a popular vehicle for probing the mechanisms underlying perception and cognition. The earliest animal experiments that adopted this approach used experimental paradigms in which the observable behavior implicitly indicated decision confidence (Smith et al. 1997, Hampton 2001, Foote & Crystal 2007, Kepecs et al. 2008, Kiani & Shadlen 2009). This approach differs from the paradigms that have long been used in humans and which require explicit confidence reports (Peirce & Jastrow 1884, Johnson 1939, Festinger 1943). But this gap is shrinking. Recent studies of perceptual decision confidence in macaques employed behavioral assays that approach something close to an explicit confidence report (Boundy-Singer et al. 2025, Vivar-Lazo & Fetsch 2025), albeit still motivated by reward (see the section titled Behavioral Measurements of Decision Confidence). This methodological evolution enabled direct quantitative comparison of the metacognitive capacities of humans and macaques (Boundy-Singer et al. 2025). It also provided a new test bed for connecting computational models of perceptual decision confidence to the neurobiological substrate

of confidence-mediated behavior (Boundy-Singer et al. 2025, Vivar-Lazo & Fetsch 2025; see also Middlebrooks & Sommer 2012). Establishing this connection has additionally been helped by the development of signal processing techniques capable of uncovering neural signatures of decision-making and confidence assignment from noninvasive human recordings (Balsdon et al. 2020, 2021; Geurts et al. 2022; Balsdon & Philiastides 2024). This tapestry of recent developments and the resulting discoveries are the focus of this review.

We provide a brief overview of popular experimental methods for measuring decision confidence, computational accounts of the mental operations underlying behavioral confidence reports, and current insights into the neural correlates of confidence computations. We highlight recent studies whose results suggest that confidence in perceptual decisions arises from a hierarchical transformation of sensory population activity that aims to estimate decision reliability and unfolds in parallel with decision formation. We end by speculating on the connections between perceptual confidence and related capacities that fall under the rubric of performance monitoring.

BEHAVIORAL MEASUREMENTS OF DECISION CONFIDENCE

Confidence measurements have a long history (reviewed in Kepecs & Mainen 2012, Mamassian 2016), show up in distinct parts of the scientific literature (Fleming 2024), and take on various forms. Here, we focus on confidence in binary perceptual decisions. Our discussion thus pertains to tasks in which a subject is presented with a sensory stimulus and must judge whether this stimulus belongs to Category A or Category B, for example, whether a patch of colored dots contains more red or green dots (**Figure 1*a***). What is common to all confidence paradigms is that one way or another the task involves an additional question about the subject's confidence in their perceptual decision. In the most straightforward paradigm, this question is asked directly by presenting the subjects with a response scale that discretizes confidence into two or more levels. In some experiments, subjects simultaneously communicate their perceptual decision and confidence report.



Implicit confidence report		Explicit confidence report	
Opt out task	Waiting time task	Direct report	Comparison report

Partial taxonomy of confidence probes in binary perceptual decision-making tasks

Figure 1

Measuring confidence in perceptual decisions. (*a*) Example psychophysical task. The observer judges whether the stimulus contains more red or green dots. They jointly communicate their perceptual decision and their decision confidence by selecting one of four choice targets. The horizontal dimension indicates the perceptual judgment, and the vertical dimension indicates the confidence report. (*b*) Task difficulty can be manipulated by changing either the stimulus strength (the red–green ratio) or the stimulus reliability (variability in dot size). (*c*) A partial taxonomy of experimental paradigms suitable to probe confidence in binary perceptual decisions.

When this is the case, there are at least four response options to choose from: Category A—confident, Category A—not confident, Category B—not confident, and Category B—confident (**Figure 1*a***). Across studies, there is considerable variation in the number of available confidence levels and in the behavioral procedure used to indicate the selected option. Regardless of the details of the confidence reporting method, if the available response options are only associated with a verbal label or numerical score (for example, 3—somewhat confident), their meaning remains vague.

What exactly does it mean to be confident in a decision? We are not asking this question in a philosophical sense, but as a practical matter: If a subject wanted to perform the task illustrated in **Figure 1*a*** perfectly, on which trials should they claim to be confident? Only when they feel 100% convinced that the perceptual judgment is correct? Or would 75% be sufficient? What if they feel certain of their perceptual experience but are not convinced that the associated choice is correct? Without further specification, these questions cannot be answered objectively. An ideal strategy can be identified only when there is an explicit goal to strive for (Tanner et al. 1960). In the absence of such a goal, the confidence reporting scale could mean different things to different subjects. For a confidence task to be well defined, the behavioral options need to be coupled to consequences in the domain of reward. Confidence incentivization typically takes the form of a variable reward scheme whereby confident responses are high risk and high reward, while not confident responses are low risk and low reward (Persaud et al. 2007). The reward-maximizing strategy requires estimating the likelihood of the primary decision being correct and applying a criterion to this estimate, for example, select the confident option when you believe that the likelihood of a correct decision exceeds 80%. Such incentivization is always present in animal studies but rarely in human experiments. We speculate that this is one reason why human confidence assignment strategies sometimes appear idiosyncratic (Navajas et al. 2017) and irrational (Peters et al. 2017, Bertana et al. 2021).

The sense of confidence helps us distinguish easy from difficult decisions. To illuminate the neural and computational processes underlying this ability, confidence experiments often involve manipulations of task difficulty, although some studies instead seek to maintain a constant level of difficulty (Fleming et al. 2010). In the former case, the stimulus will typically be varied along the task-relevant perceptual dimension. For our example task, this entails creating stimulus conditions that vary in the ratio of red and green dots (**Figure 1*b***). We refer to this as a manipulation of stimulus strength with respect to a task-imposed categorization boundary, whereby strong stimuli are easier to judge than weak stimuli. Task difficulty is determined not only by stimulus strength but also by orthogonal factors, such as stimulus eccentricity, size, duration, and contrast. Manipulating the stimulus along these various dimensions can impact task difficulty without changing the stimulus strength (**Figure 1*b***). Such effects are well understood as arising from a change in the reliability of the stimulus or at least in the associated perceptual estimate. We refer to this as manipulations of stimulus reliability, whereby reliable stimuli yield more precise perceptual estimates than unreliable stimuli. In the natural environment, multidimensional stimulus variability is omnipresent (Webb et al. 2023). While the sense of confidence is adapted to this complexity, it is rarely reflected in laboratory experiments. In summary, the difficulty of a perceptual decision is jointly determined by the perceptual dimension of interest and the orthogonal dimensions that govern the reliability of perceptual estimates. It follows that the sense of confidence must take both factors into account.

Decision confidence can be probed in various ways. One key distinction is whether the task invites explicit or implicit confidence reports (**Figure 1*c***). Explicit tasks require subjects to either directly rate their confidence in a single decision (Peirce & Jastrow 1884) or compare confidence in a pair of decisions (Barthelmé & Mamassian 2009, de Gardelle & Mamassian 2014). Implicit tasks

instead measure indirect behavioral indicators of perceptual decision confidence. For example, in waiting time tasks, the key behavioral measure is how long the subject is willing to wait for an uncertain reward (Kepecs et al. 2008). In opt out tasks, the key measure is the fraction of trials for which the subject prefers the safe small bet over the uncertain big reward (Persaud et al. 2007). Early animal studies of decision confidence used implicit tasks (Smith et al. 1997, Hampton 2001, Foote & Crystal 2007, Kepecs et al. 2008, Kiani & Shadlen 2009). Two recent studies went a step further and designed an incentivization scheme that invited monkeys to jointly report a perceptual choice and their confidence in this decision (Boundy-Singer et al. 2025, Vivar-Lazo & Fetsch 2025; see also Middlebrooks & Sommer 2012). Both studies found that, once fully trained, monkeys' behavioral responses closely resemble human direct confidence reports.

While these task paradigms differ substantially, they all seem to work. Subjects typically exhibit more confidence in correct than incorrect decisions under each of these paradigms (Peirce & Jastrow 1884, Kepecs et al. 2008, Barthelmé & Mamassian 2009, Kiani & Shadlen 2009). This is true both across and within experimental conditions. These observations have long intrigued psychologists. What are the mental processes and neural operations that endow us with this self-knowledge? How can the brain possibly know which perceptual interpretations of the environment are at risk of being flawed? And what distinguishes good from bad self-knowledge? Answers to these questions are provided by idealized mathematical descriptions of the confidence assignment processes at work.

PROCESS MODELS FOR DECISION CONFIDENCE

The search for the neuronal basis of visual confidence is greatly helped by quantifiable hypotheses that specify how a cascade of sensory transformations leads to observable behavior. Though the definition is debated, we can think of a process model as a theoretical framework that describes the causal chain of events comprising a single instance of a decision or other cognitive process. The focus is on the mechanistic or algorithmic level rather than specifying a normative ideal or statistical description of aggregate behavior. Two such frameworks have shaped much of the research in this domain: static signal detection theory (SDT) and the dynamic framework of sequential sampling (SS). Both frameworks have roots that go back almost a century (Tanner & Swets 1954, Stone 1960). We do not attempt to document this history but limit our discussion to recently proposed models that have emerged from these research traditions. Our goal is not to provide a complete overview of the model variants that populate the literature. We simply seek to illustrate how process models offer an essential bridge between observable behavior and its neurobiological basis. Note that the models we discuss are related to but distinct from the Bayesian confidence hypothesis, which proposes that the sense of confidence equates to a Bayesian decision-maker's belief in the posterior probability of a choice being correct (Meyniel et al. 2015, Sanders et al. 2016, Li & Ma 2020, Xue et al. 2024).

Signal Detection Theory Models

The stimuli used in perceptual confidence tasks (sinusoidal gratings, random dot motion, human faces, etc.) excite millions of sensory neurons with complex response properties. The neural processes underlying decision-making and confidence assignment might therefore be expected to be similarly high-dimensional and complex. Fortunately, this intuition is wrong. Although both processes involve large populations of neurons, the task-relevant component of this population activity typically resides in low-dimensional subspaces (Mante et al. 2013, Peixoto et al. 2021, Latimer & Freedman 2023, Charlton & Goris 2024, Boundy-Singer et al. 2025). It follows that we can develop useful intuitions about this activity by examining low-dimensional mathematical

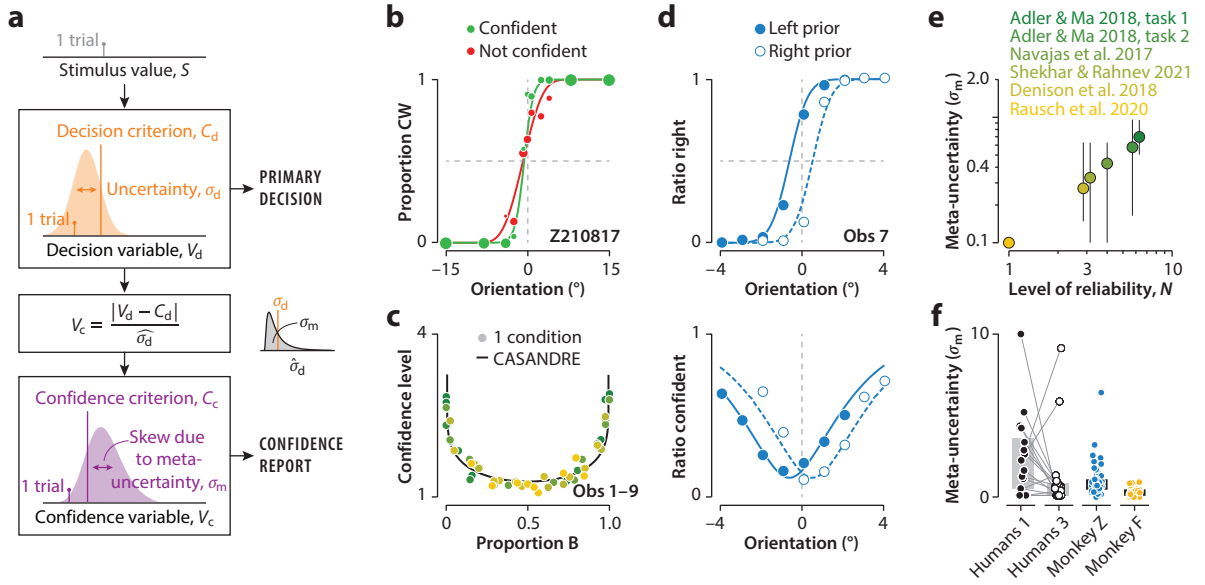


Figure 2

Process models based on the signal detection theory framework offer explicit, falsifiable hypotheses for the computations that underlie and constrain confidence in perceptual decisions. (*a*) Schematic of the hierarchical decision-making process underlying choice confidence data in one such model (CASANDRE). (*b*) Choice confidence data of a monkey performing an orientation discrimination task with direct binary confidence reports. Confident choices are shown in green, and not confident choices are shown in red. The lines show the fit of the CASANDRE model. (*c*) The confidence-consistency relationship of a human subject performing an orientation categorization task with direct confidence reports (four levels). Symbol color indicates stimulus contrast. The line shows the fit of the CASANDRE model. (*d*) Choice confidence data of a human subject performing an orientation discrimination task with comparison confidence reports under two different stimulus distributions (left prior versus right prior). (*Top*) Perceptual choices. (*Bottom*) Confidence reports. The lines show the fit of a decision reliability model. (*e*) Median level of meta-uncertainty plotted against the number of stimulus reliability levels for six confidence experiments. (*f*) Meta-uncertainty for a group of human subjects and two monkeys performing the same perceptual confidence task. For the humans, each symbol represents the metacognitive performance of one subject in one block of trials (block 1 is shown in black, and block 3 is shown in white). For the monkeys, each symbol represents metacognitive performance in one behavioral session. Abbreviations: c, confidence; CASANDRE, confidence as a noisy decision reliability estimate; d, decision; m, meta; Obs, observer. Data in panel *c* were originally reported in Adler & Ma (2018) and reanalyzed in Boundy-Singer et al. (2023). Panels *a*, *c*, and *e* adapted from Boundy-Singer et al. (2023); copyright 2022 The Author(s). Panels *b* and *f* adapted from Boundy-Singer et al. (2025) (CC BY-NC-ND 4.0). Panel *d* adapted from Caziot & Mamassian (2021) (CC BY 4.0).

models that offer idealized descriptions of these processes. As an extreme example, assume that a perceptual decision is based on a single neurally encoded number. We can think of this number as representing an observer's perceptual estimate of a task-relevant stimulus feature (for example, the ratio of red to green dots) and refer to it as the decision variable (DV). Further, assume that perception is subject to noise such that repeated presentations of the same stimulus give rise to variable perceptual estimates. In this scenario, a goal-oriented decision-making strategy consists of comparing the DV with a decision criterion (Tanner et al. 1960, Green & Swets 1966) (**Figure 2*a***, top). The simplicity of this extreme abstraction, formalized in SDT, elegantly reveals that choices always reflect a combination of a subject's sensitivity (i.e., the fidelity of their perceptual representations) and their response bias (i.e., the tendency to prefer one response over the other).

How can introspection reveal which decisions are likely to be correct and should be accompanied by a high degree of confidence, and which are not? Under the SDT framework, more

extreme DV values will occur more frequently for correct than incorrect decisions. It follows that a decision-maker can exploit this association to assign confidence to decisions. Specifically, the distance to the criterion offers a principled confidence variable (Treisman & Faulkner 1984, Kepecs et al. 2008, Komura et al. 2013), at least for tasks in which all conditions are subject to the same level of perceptual variability. However, in many experimental tasks and real-world situations, the dispersion of the DV will differ across conditions. To obtain a principled confidence variable in such settings, the distance to the criterion needs to be normalized by the DV's uncertainty (**Figure 2*a***, middle). This operation yields an estimate of the reliability of the decision. Several recently developed SDT-based models propose that an estimate of decision reliability guides confidence-mediated behavior (Caziot & Mamassian 2021, Shekhar & Rahnev 2021, Mamassian & de Gardelle 2022, Boundy-Singer et al. 2023) (**Figure 2*a***, bottom). As we discuss next, this hypothesis explains some intriguing aspects of perceptual confidence data (Koriat 2012, Caziot & Mamassian 2021, Mamassian & de Gardelle 2022, Boundy-Singer et al. 2023).

Decision reliability models of confidence capture the observation that subjects typically report more confidence in correct than incorrect decisions. As a case in point, consider the behavior of a monkey performing an orientation discrimination task with direct binary confidence reports (**Figure 2*b***). The subject's confident and not confident perceptual choices are shown in green and red, respectively (**Figure 2*b***). Clearly, confident choices tend to be more accurate than not confident choices, resulting in a steeper relationship between stimulus orientation and perceptual choice (**Figure 2*b***, red versus green symbols). A decision reliability model describes these effects well (**Figure 2*b***, red versus green lines). These models also capture another prominent feature of choice confidence data. Across many repeated trials, the average level of reported confidence is often lawfully related to the consistency of the primary choice. This is evident in the data of a human subject who performed a stimulus categorization task with direct confidence reports (four levels). Both stimulus strength and stimulus reliability varied considerably across trials (Adler & Ma 2018). A single confidence-consistency relationship neatly summarizes the data across all conditions (**Figure 2*c***). Key to the decision reliability hypothesis is that confidence arises from an evaluation of the quality of the primary decision, not from a direct evaluation of the sensory input as such. Perceptual interpretations of the environment typically reflect an interaction between previous experience and current sensory input. Contextual manipulations that alter the distribution of experiences can therefore impact perceptual decisions. If perceptual confidence truly reflects an estimate of decision reliability, these manipulations ought to impact confidence reports in a similar fashion. This prediction is correct (Locke et al. 2020, Caziot & Mamassian 2021, Mihali et al. 2023), as can be seen in the data of a human subject who performed an orientation discrimination task with confidence comparison reports under two different stimulus distributions (**Figure 2*d***, left prior versus right prior).

Decision reliability models illuminate not only the computations that underlie decision confidence but also the factors that constrain its quality. Recall that the confidence computation in these models involves the uncertainty of the DV that informed the primary choice (**Figure 2*a***, middle). To accurately estimate the reliability of a choice, a subject thus needs to know this uncertainty. If there is uncertainty about this uncertainty (meta-uncertainty), the decision reliability estimate will be noisy. This idea is formalized in the confidence as a noisy decision reliability estimate (CASANDRE) model (Boundy-Singer et al. 2023). In the CASANDRE framework, meta-uncertainty is the sole factor that determines the quality of metacognition. The higher the level of meta-uncertainty, the weaker the association between confidence and decision reliability. This framework makes the unique prediction that meta-uncertainty will be higher in experiments that involve more levels of stimulus reliability. In other words, metacognitive abilities may depend on the specifics of the task, just like perceptual abilities do. Boundy-Singer et al. (2023) compared

results from six confidence studies and found that meta-uncertainty tends to grow with the number of reliability levels (**Figure 2*e***). Other decision reliability models have proposed alternative sources of confidence noise, such as instability of the confidence criteria across trials (Shekhar & Rahnev 2021). Some model variants additionally include a confidence boost component, which has the opposite effect of noise and captures information about decision reliability acquired after the decision has been committed to (Mamassian & de Gardelle 2022).

Meta-uncertainty provides a theoretically pure measure of metacognitive ability that expresses how well a decision-maker can discriminate reliable from unreliable choices, regardless of their level of perceptual sensitivity or response biases. Importantly, it is anchored in a process model of the decision that underlies behavioral confidence measurements. Psychologists have long sought to measure the quality of the sense of confidence (Nelson 1984). This is typically done by using metrics that are agnostic about the generative process that underlies confidence reports. For example, one popular statistic (meta-d′) seeks to measure how well confidence judgments distinguish correct from incorrect decisions using a sensitivity metric that resembles conventional d′ (Maniscalco & Lau 2012). Another recently introduced statistic (meta-*I*) is instead based on information theory (Dayan 2023). While these metrics are convenient, they not only reflect metacognitive ability but also depend on perceptual sensitivity and response biases (Guggenmos 2021, Xue et al. 2021, Vuorre & Metcalfe 2022, Arnold et al. 2023, Boundy-Singer et al. 2023, Dayan 2023, Rahnev 2025). We think that process models of confidence have matured enough for principled metrics of metacognitive ability to become the norm in the near future. Boundy-Singer et al. (2025) recently used such a model-based approach to compare the quality of confidence reports in humans and monkeys who performed the same perceptual confidence task under analogous incentives. The monkeys were well-trained specialists, while the humans were novices. Still, it is surprising that the monkeys initially outperformed the humans in their ability to judge the quality of perceptual orientation judgments (**Figure 2*f***, humans 1 versus monkeys Z and F). It took the humans two more 1,100-trial sessions to catch up with the monkeys (**Figure 2*f***, humans 3 versus monkeys Z and F). Note that these human data constitute a rare empirical indication of metacognitive learning; this topic seems ripe for experimental investigation but will likely require the use of principled metrics (Carpenter et al. 2019, Rouy et al. 2022).

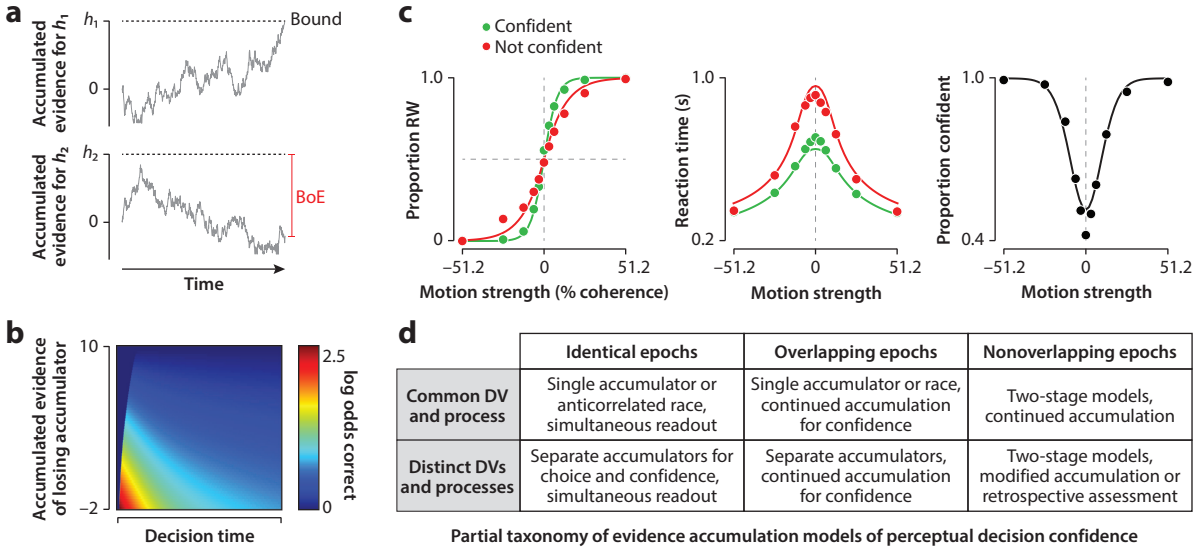
Sequential Sampling Models

While static SDT provides an elegant and tractable framework for building confidence models, the theory lacks an explicit representation of time. As a consequence, SDT-based models make no direct prediction about the relation between decision confidence and the time it takes to report a decision. Yet, in many perceptual tasks, a prominent relation is obvious: Confident decisions are reported faster. This suggests that decision confidence is deeply connected to decision time (Henmon 1911, Audley 1960, Vickers 1979). Dynamic models seek to explain this relationship, as well as the aforementioned confidence-choice relationships that can be captured by static models (**Figure 2*b–d***).

Most dynamic confidence models are derived from the broader framework of bounded SS, commonly referred to as evidence accumulation, and inclusive of drift-diffusion and race models (Ratcliff & Rouder 1998, Gold & Shadlen 2007). These models can be considered an extension of SDT in time, motivated by tasks in which evidence bearing on the decision is available in the form of a continuous stream or a discrete sequence. This temporal structure can be an explicit property of the external stimulus to be judged, as is the case for random dot motion and event-rate or counting tasks (Shadlen et al. 1996, Raposo et al. 2012, Brunton et al. 2013). These tasks simply cannot be solved based on a single click or video frame; integration across time is required. But even when

not explicit, temporal integration is almost always an implicit property of sensory processing. Of course, peripheral receptors and central neurons have their own intrinsic time constants, but sensory responses are also corrupted by momentary noise (Schiller et al. 1976, Shadlen & Newsome 1998, Faisal et al. 2008, Goris et al. 2024). It follows that even for static stimuli, integration of evidence for periods longer than the biophysical time constant may improve the signal-to-noise ratio of the DV (Bloch 1885, Watson 1979, Osborne et al. 2004, Goris et al. 2018, Langlois et al. 2025).

The classic normative rule for confidence judgments in a dynamic framework is attributed to Vickers, who observed that confidence should depend on the balance of evidence supporting each of the two options (Vickers et al. 1972, Vickers 1979). Intuitively, if the collected evidence favors a Category A decision, confidence in this decision should also take into account how well a Category B choice is supported by this evidence. The larger the difference in support for the two options, the more likely the choice is correct, and the higher the confidence in the decision should be. An elegant implementation of this notion is offered by race models (Vickers 1979, Kiani et al. 2014) (**Figure 3*a***). In these models, decision-making is portrayed as a race between two accumulators, each favoring one of the categorical options. The race ends when the first accumulator reaches a terminating bound, and confidence is inversely related to the amount



	Identical epochs	Overlapping epochs	Nonoverlapping epochs
Common DV and process	Single accumulator or anticorrelated race, simultaneous readout	Single accumulator or race, continued accumulation for confidence	Two-stage models, continued accumulation
Distinct DVs and processes	Separate accumulators for choice and confidence, simultaneous readout	Separate accumulators, continued accumulation for confidence	Two-stage models, modified accumulation or retrospective assessment

Figure 3

Dynamic models link decision accuracy, speed, and confidence under an evidence accumulation framework. (*a*) A leading model posits a race between two partially anticorrelated accumulators, each representing one categorical choice option. The winning accumulator determines the choice and RT, while confidence can be read out from the losing accumulator, which reflects the BoE at decision time. (*b*) The BoE rule can be formalized by calculating the log posterior odds of being correct as a function of the state of the losing race. When stimulus difficulty is unpredictable, this relationship is time dependent. (*c*) Monkeys were trained to report the direction of motion in a random dot display and to report their choice and confidence simultaneously in an RT paradigm using a target configuration similar to **Figure 1*a***. Panels show choice, RT, and confidence (proportion high bet) as functions of motion strength (% coherence), where positive values indicate rightward motion and negative leftward. Smooth curves are fits to the race model. (*d*) Partial taxonomy of dynamic models defined by whether choice and confidence make use of the same or distinct accumulation process and/or time epoch. From left to right starting in the top row, examples of each model class can be found in the following references: Kiani et al. 2014, Desender et al. 2021a, Pleskac & Busemeyer 2010, Balsdon et al. 2021, Maniscalco et al. 2021, Navajas et al. 2016. Abbreviations: BoE, balance of evidence; DV, decision variable; RT, reaction time; RW, rightward. Panel *b* adapted from Kiani et al. (2014); copyright 2014 Elsevier. Panel *c* adapted from Vivar-Lazo & Fetsch (2025) (CC BY-NC-ND 4.0).

of evidence favoring the losing accumulator. Although it is difficult to precisely define Bayes optimality in a dynamic model, the balance of evidence rule can be linked to Bayesian confidence by positing a mapping between the state of the losing race and the log posterior odds that the decision is correct (Kiani et al. 2014) (**Figure 3*b***). How such a mapping could be acquired and represented in the brain is an open question (Le Denmat et al. 2024). Regardless, the framework makes explicit predictions for the relationships between choice accuracy, reaction time, and decision confidence. In many tasks, these predictions have been found to capture behavioral data well (Kiani et al. 2014, van Den Berg et al. 2016, Vivar-Lazo & Fetsch 2025).

In a recent study, Vivar-Lazo & Fetsch (2025) trained macaques in a binary motion discrimination task with explicit confidence reports. The monkeys were tasked with judging whether the global stimulus was dominated by leftward or rightward motion and simultaneously reported their confidence in each decision (i.e., high versus low). Importantly, the subjects were free to report their decision (and confidence) when ready, resulting in fast and slow trials. The animals' choice confidence reports resemble those of the previous example (**Figure 2*b***)—easier task conditions were associated with better performance, as were high confidence reports (**Figure 3*c***, left and right). Additionally, the animals' average reaction time lawfully depended on task difficulty and was therefore inversely related to decision confidence (**Figure 3*c***, middle). These data are representative of observations made in human subjects (Kiani et al. 2014) and are well captured by a race model (**Figure 3*a*,*b***). Note that an inverse relationship between response time and confidence does not entail that confidence computations incorporate the passage of time per se. This was the case in the model and data of Kiani et al. (2014), but in general, decision speed may correlate with confidence simply because both are correlated with accuracy. Whether elapsed time plays a causal role likely depends on details of the task and was not directly tested in the monkey study of Vivar-Lazo & Fetsch (2025). This question may seem esoteric but it bears on deeper questions about how the brain performs inference under uncertainty (Hanks et al. 2011, Shadlen & Kiani 2013, Khalvati et al. 2021, Langlois et al. 2025), the intriguing possibility being that time itself is used by the brain as a proxy for evidence reliability (Shadlen & Kiani 2013). We might go a step further and say that time informs an estimate of decision reliability; after all, the distribution of the DV is affected not only by the strength of sensory evidence but by nonsensory factors (attention, biases, etc.) that limit the quality of the decision. An interesting hypothesis that follows is that imprecision in the estimate of accumulated evidence and/or elapsed time in a dynamic model may be the conceptual analog of the meta-uncertainty component in static models like CASANDRE.

Taking a step back, dynamic models allow us to ask more fundamental questions about the relative timing of decision and confidence computations (Baranski & Petrusic 1998, Xue et al. 2023). Is confidence determined only after choice commitment, or might it be available during decision formation? Under a race model, the balance of evidence could in principle be read out continuously and would furnish a provisional choice confidence report: If forced to choose right now, I would choose this option and my confidence would be this high. Under some alternative models, confidence can only be determined following an epoch of additional accumulation after choice commitment (Pleskac & Busemeyer 2010, Moran et al. 2015, Navajas et al. 2016). Indeed, there is compelling evidence for postdecisional processing in confidence judgments (Murphy et al. 2015, Desender et al. 2021b), but there is no reason to see this as mutually exclusive with a provisional confidence estimate emerging during the decision process (Gherman & Philiastides 2015, Balsdon et al. 2021, Vivar-Lazo & Fetsch 2025).

A partial taxonomy of dynamic models (**Figure 3*d***) can be defined by whether choice and confidence make use of the same or different information, both in terms of when the information is used and also what information is used. For example, choice and confidence could be computed in

parallel but with the latter governed by a distinct process that accumulates independent evidence for each choice rather than a comparison signal or relative evidence (Maniscalco et al. 2021). Alternatively, Balsdon & Philiastides (2024) propose an unbounded relative evidence accumulator for confidence and a separate motor accumulator dictating the choice, where the former controls the leakiness of the latter to adapt to dynamic changes in evidence strength or reliability. We return to this issue below in the discussion of human electroencephalography (EEG) studies and macaque neurophysiology. What we can say from behavioral work in humans (Kiani et al. 2014) and monkeys (Vivar-Lazo & Fetsch 2025) is that postdecision accumulation is not necessary for rational confidence-mediated behavior. Whether and when an online confidence estimate is accessible has further implications for sequential or hierarchical tasks (Sarafyazd & Jazayeri 2019, Zylberberg 2021), where a prediction of accuracy for a given decision or action informs the selection and execution of the next one. More broadly, models and experiments that jointly resolve the what and when of confidence formation are essential if we want to understand the neural operations and information flow underlying metacognition (Fleming 2024).

In summary, process models of visual confidence offer three key insights. First, the brain's confidence computation acts on the same sensory inputs that inform perceptual decisions. Second, the brain's confidence computation extracts an estimate of both stimulus strength and stimulus reliability to estimate something akin to decision reliability. And third, choice confidence computations unfold over time rather than being implemented as a snapshot process, and they appear to do so in parallel, facilitating a role for confidence in modulating current or subsequent decision processes. Armed with these insights, we now turn to the question of how neural circuits actually implement these computations.

NEURAL CORRELATES OF CONFIDENCE COMPUTATIONS IN NONHUMAN ANIMALS

Anatomically, the primate visual system is organized as a hierarchical network composed of parallel processing streams (Felleman & Van Essen 1991). Computationally, visual processing can therefore be understood as a series of transformations that extract behaviorally useful information from raw visual input. This perspective on neural computation readily explains how the same photoreceptor responses inform our ability not only to recognize objects (DiCarlo et al. 2012, Yamins et al. 2014) but also to perceive form (Van Essen & Gallant 1994), color (Gegenfurtner & Kiper 2003), motion (Simoncelli & Heeger 1998, Mineault et al. 2012), and depth (DeAngelis 2000). It even explains how these responses enable us to predict the future state of the visual environment (Hénaff et al. 2019, 2021). Might visual confidence also belong in this list? In other words, is it possible to specify a cascade of neural operations that converts low-level sensory responses into a perceptual decision confidence variable? If so, could we recognize signatures of these downstream operations in neural activity in the sensory cortex? This is the question that Boundy-Singer et al. (2025) sought to address for a fine orientation discrimination task with direct confidence reports (**Figure 4*a***).

In the primary visual cortex (V1), neurons are tuned for local image orientation (Hubel & Wiesel 1962), making this area well suited to inform perceptual orientation judgments. Does V1 activity also inform confidence in these decisions? To gain an intuition for downstream transformations that could accomplish this, consider a pair of hypothetical V1 neurons (**Figure 4*b***). One of these neurons prefers clockwise-oriented stimuli; the other prefers counterclockwise orientations. Their joint activity pattern therefore contains information about whether a stimulus is more or less likely to have a clockwise orientation (**Figure 4*c***, left). To convert this population response into a perceptual decision, a downstream decision-making circuit could in principle

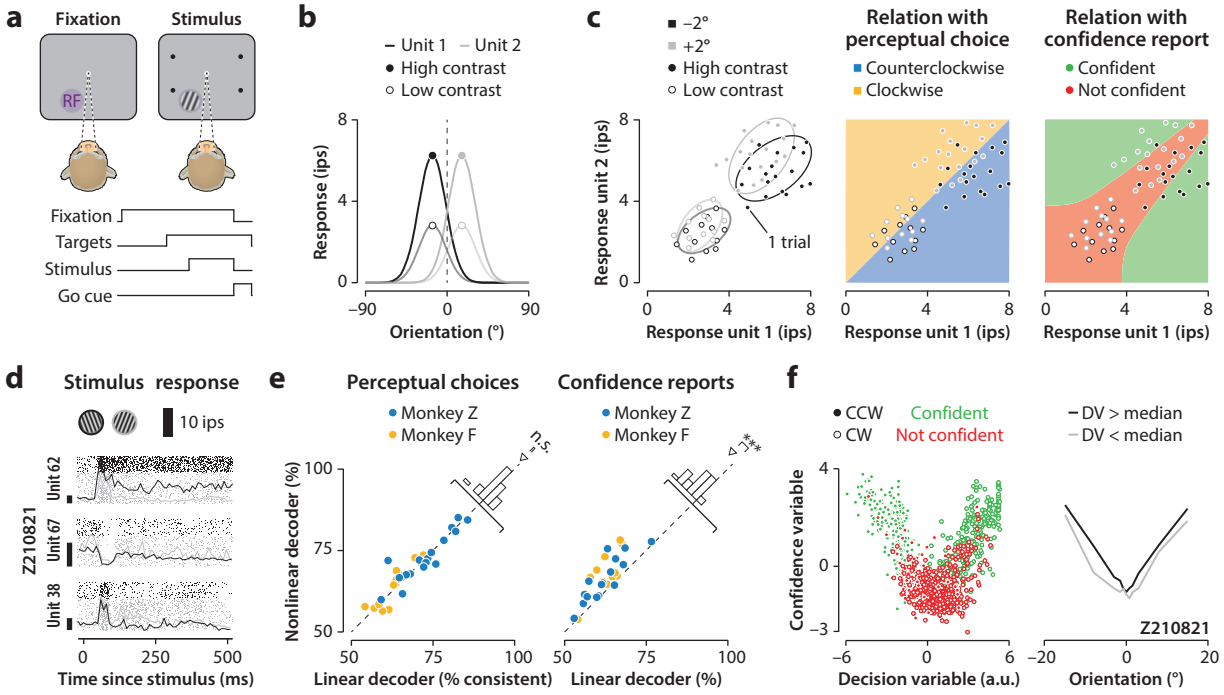


Figure 4

Neuronal operations that convert low-level sensory responses into a high-level confidence representation. (*a*) Sequence of events in the orientation discrimination task studied by Boundy-Singer et al. (2025). After the observer acquires fixation, four choice targets appear, followed by the stimulus. The stimulus is placed in the recorded neurons' visual receptive field (RF). The observer judges whether the stimulus is rotated CW or CCW relative to vertical. They jointly communicate this orientation judgment and their decision confidence with a saccade toward one of the four choice targets (as in **Figure 1*a***). (*b*) Orientation tuning functions for two model neurons (*black* versus *gray*) at high and low stimulus contrast (*open* versus *closed symbols*). (*c*, *left*) Joint responses of the pair of model neurons to repeated presentations of four stimuli that differ in orientation and contrast. (*Middle*) Illustration of a mapping rule that converts the pairwise activity into a perceptual decision. (*Right*) Illustration of a mapping rule that converts the same responses into a confidence report. (*d*) Spike rasters (*dots*) and peristimulus time histogram (*lines*) of three example units during presentation of a CW (*gray*) and CCW (*black*) stimulus. (*e*) Direct comparison of linear and nonlinear decoders. (*Left*) Comparison of the proportions of correctly predicted perceptual choices by linear (abscissa) and nonlinear (ordinate) choice decoders. (*Right*) Comparison of the proportions of correctly predicted confidence reports by linear (abscissa) and nonlinear (ordinate) confidence decoders. The asterisks indicate $P < 0.001$. (*f*) Relationship between the decoded confidence and decision variables. (*Left*) Each point represents a single trial in an example recording session. (*Right*) The mean confidence level plotted against stimulus orientation for trials with a DV value that is more (*black*) or less (*gray*) extreme than the stimulus-specific median. Abbreviations: a.u., arbitrary units; CCW, counterclockwise; CW, clockwise; DV, decision variable; ips, inches per second; n.s., not significant; RF, receptive field. Figure adapted from Boundy-Singer et al. (2025) (CC BY-NC-ND 4.0).

use a linear hyperplane to separate the response patterns associated with both stimulus categories (**Figure 4*c***, middle). The resulting decisions will not be flawless. There is overlap between both response distributions, making errors inevitable. The critical insight is that the structure of the population activity also contains information about the likelihood of such an error. First, responses that are close to the hyperplane are more likely to be caused by weak stimuli and thus more likely to yield flawed perceptual decisions (**Figure 4*c***, middle). Second, response patterns that reside near the bottom left corner of this space are more likely to be caused by less reliable low-contrast stimuli and thus more likely to yield perceptual errors (Mareschal & Shapley 2004, Hénaff et al. 2020, Boundy-Singer et al. 2024). It follows that a downstream confidence assignment circuit in

principle can use a nonlinear hyperplane to convert this sensory population activity into a confidence assessment.

To examine the relationship between sensory population activity and decision confidence, Boundy-Singer et al. (2025) recorded from diversely tuned V1 populations (**Figure 4*d***). They then trained linear and nonlinear decoders to predict the monkeys' perceptual choices and confidence reports from this activity. In both cases, the decoders performed above chance, demonstrating that the same sensory activity might inform perceptual and metacognitive judgments. However, there was an interesting distinction. The nonlinear decoders consistently outperformed their linear counterparts in predicting confidence reports but not perceptual decisions (**Figure 4*e***). This pattern is consistent with the hypothesis that the confidence assignment process involves a nonlinear transformation of sensory activity that is distinct from the linear transformation that gives rise to perceptual decisions. This does not mean that V1 contains a representation of confidence, nor that confidence can be trivially decoded from any site in the visual pathway, including the retina, if nonlinear operations are permitted (Pouget et al. 2016). Confidence is a latent cognitive variable used to guide behavior, not an abstract mathematical quantity. Boundy-Singer and colleagues' key advancement was linking V1 activity to a direct behavioral readout of confidence on a trial-by-trial basis (see also Geurts et al. 2022).

To further elucidate the transformations of V1 activity into a behavioral confidence report, Boundy-Singer et al. (2025) investigated the relationship between the latent variables used by the choice and confidence decoders. As expected from decision reliability models of confidence, the neurally decoded confidence variable and DV exhibited a U-shaped relationship (**Figure 4*f***, left). This means that trials that elicited a stronger DV value also tended to elicit a higher level of confidence. Crucially, this effect was also evident within fixed stimulus conditions (**Figure 4*f***, right, black versus gray line). Together, these results provide the first empirical evidence that for simple perceptual decisions, a set of specialized nonlinear transformations converts sensory population activity into a decision reliability estimate, which in turn informs confidence-mediated behavior. Stepping back, these findings highlight how sophisticated behavior can arise from a cascade of simple operations. This notion underlies much of the success of modern artificial intelligence (LeCun et al. 2015, Tuckute et al. 2024). We think it also applies to certain components of biological intelligence.

If perceptual decisions and decision confidence arise from distinct transformations, it is natural to ask whether both computations unfold simultaneously, as suggested by the behavioral studies described above. The capacity to compute multiple things at once is a key feature of parallel hierarchies and is evident across visual modalities: We simultaneously perceive form, motion, and depth. Is this also true for visual confidence? More specifically, does the brain's confidence computation unfold in parallel with the decision-making process itself? To address this question, Vivar-Lazo & Fetsch (2025) investigated neural activity in a decision-making circuit downstream of the visual cortex during a motion discrimination task with direct confidence reports. They focused on the lateral intraparietal area (LIP), a region involved in visuospatial attention and the planning of saccadic eye movements (Bisley & Goldberg 2010, Snyder et al. 2000), as well as other aspects of visual cognition, such as category judgments (Freedman & Assad 2006). LIP has been shown to represent an evolving DV that explains choice and response time (Roitman & Shadlen 2002, Kira et al. 2015, Steinemann et al. 2024) and predicts implicit confidence reports in an opt out task (Kiani & Shadlen 2009). The strategy is to exploit the spatial and oculomotor properties of LIP to expose the computations underlying the decision of where to move the eyes; thus, Vivar-Lazo & Fetsch (2025) used a task configuration very similar to Boundy-Singer et al. (2025) and recorded from LIP neurons whose response fields (RFs) overlapped with one of the four choice targets (**Figure 5*a***).

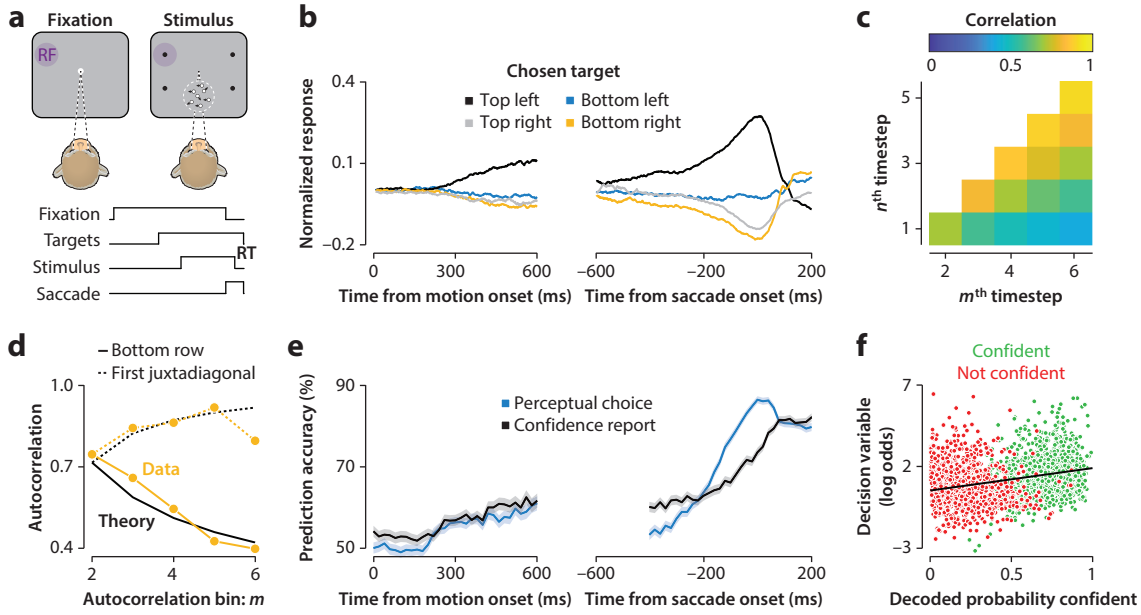


Figure 5

Neural dynamics in LIP support concurrent evolution of choice and confidence signals within a common population. (*a*) Sequence of events in the motion discrimination task studied by Vivar-Lazo & Fetsch (2025). After the observer acquires fixation, four choice targets appear, followed by the stimulus. One of the choice targets is placed in the recorded neurons' response field (RF). The observer judges whether dot motion is predominantly in the left or right direction. They jointly communicate this motion judgment and their decision confidence with a saccade toward one of the four choice targets (as in **Figure 1*a***). (*b*) Average firing rate (normalized and baseline subtracted) of neurons with a spatial RF overlapping the contralateral (*left*) high-confidence target, aligned to stimulus onset and saccade onset. Colored traces indicate the eventual choice of the monkey. Activity traces for high versus low choices separate at the same time as those for left versus right, consistent with parallel deliberation for choice and confidence. (*c*) Theoretical autocorrelation matrix of an ideal noisy accumulation process [Churchland et al. (2011) were the first to derive the expected autocorrelation matrix; their prediction applies to Vivar-Lazo & Fetsch's (2025) experiment]. Correlation relative to the initial time bin drops as a function of time (*bottom row*), whereas neighboring time bins show greater correlation over time as the accumulation continues (*first juxtadiagonal*). (*d*) Comparison of theoretical predictions with the corresponding time bins in the LIP spike count data. (*e*) Classification performance of logistic decoders trained to predict the perceptual choice (*blue*) and confidence report (*black*) as a function of time. Both decoders begin to ramp up simultaneously, although the choice decoder peaks slightly earlier (peri-saccade) than the confidence decoder (post-saccade). (*f*) Trial-by-trial decoding strength (model decision variable in units of log odds) for the choice decoder as a function of the strength of the confidence decoder prediction (probability of high confidence). Data are shown for only in-RF (contralateral) choices and color-coded by the monkey's behavioral confidence report. Abbreviations: LIP, lateral intraparietal area; RF, response field; RT, reaction time. Figure adapted from Vivar-Lazo & Fetsch (2025) (CC BY-NC-ND 4.0).

The use of a reaction time paradigm conferred several advantages. It allowed for trial-by-trial comparison of three key behavioral measures (choice, reaction time, and confidence; **Figure 3*c***) and served to isolate the temporal window over which neural activity could support the choice and confidence report. It also encouraged the monkeys to adopt a parallel strategy (i.e., to resolve the perceptual choice and confidence judgment concurrently), which was borne out by the behavioral results. What about the neurons? Many cells in LIP show ramping activity prior to a saccade made into their RF, with dynamics and statistical properties consistent with an evidence accumulation process (Churchland et al. 2011, Steinemann et al. 2024). Vivar-Lazo & Fetsch (2025) found that these accumulator-like neural signatures were present simultaneously for both the left versus right choice and the high versus low wager (**Figure 5*b–d***). Population decoding supported this conclusion (**Figure 5*e***). As with the V1 results from Boundy-Singer et al. (2025), there was

a trial-by-trial relationship between the strength of the decoded choice variable and the decoded degree of confidence, even within trials of fixed stimulus strength (**Figure 5*f***). Interestingly, this correlation held only for contralateral (in-RF) choices, which would correspond to the winning accumulator under the idea that LIP populations represent accumulated evidence favoring in-RF choices. This result [and a related finding from Zylberberg & Shadlen (2025)] appears to contradict the theoretical expectation that the losing accumulator contributes to confidence (Kiani et al. 2014), although caution is warranted: The correspondence between neural populations and modeled accumulators may not be as direct as is commonly assumed (Mante et al. 2013, Meister et al. 2013). Exactly how neural circuits map decision and confidence computations onto specific motor behaviors remains a topic of ongoing research.

Together, both recent studies in macaques that used direct confidence reports provide support for the hypothesis that visual confidence arises from a hierarchical transformation of sensory activity that unfolds in parallel with decision formation. Earlier findings obtained in different task paradigms offer indirect support for this hypothesis. Specifically, Fetsch et al. (2014) found that microstimulating sensory neurons in the middle temporal area during a motion discrimination opt out task simultaneously biased the monkeys' perceptual choices and opt out behavior as if the animals experienced a change in the sensory signal. These results suggest that the same sensory signals inform perceptual decisions and confidence in these decisions, as do similar results obtained with optogenetic inactivation by Fetsch et al. (2018). A different study employing the opt out paradigm found that pulvinar neurons represent decision confidence but not perceptual choice (Komura et al. 2013). Inactivating the pulvinar affected the confidence-mediated behavior but not perceptual sensitivity (Komura et al. 2013). These results suggest that in some tasks, different brain circuits implement the decision formation and confidence assignment computations, as is clearly possible (though not necessary) under the parallel hierarchical transformations hypothesis.

NEURAL CORRELATES OF CONFIDENCE COMPUTATIONS IN NONINVASIVE HUMAN EXPERIMENTS

Naturally, if we are interested in higher functions like metacognition, an essential animal model is the one reading this paragraph. The toolkit for exploring the neural basis of perceptual confidence in humans is limited by the coarse and indirect nature of blood oxygenation level–dependent (BOLD) and positron emission tomography (PET) imaging and by the scarcity of opportunistic intracranial recordings in human patients (but see below). Noninvasive electrophysiology, namely electro- and magnetoencephalography, complements the poor temporal resolution of neuroimaging, but until recently it was unclear how one could extract neural signatures of an evolving decision and confidence judgment using these methods. After all, EEG measures the aggregate electrical activity of large volumes of cortical tissue, filtered by the skull, whereas the underlying neural computations likely take place at a finer spatial scale and are transmitted through trains of action potentials. Nevertheless, aided by model-based approaches and novel signal processing techniques, recent noninvasive experiments (reviewed in O'Connell & Kelly 2021) have assembled a surprisingly rich account of perceptual decision processes and associated metacognitive evaluation in humans.

One approach is to identify an EEG correlate of evidence accumulation (O'Connell & Kelly 2021), then test whether and how the dynamics of that signal predict subjective confidence reports. This approach was taken by Gherman & Philiastides (2015), who decoded single-trial EEG signals during a perceptual categorization task with an opt out option. They found that the slope of the resulting accumulation-like ramping signal was greater for trials in which the participant was

given the chance to opt out of the decision but waived this option (high-confidence trials) than for trials in which the opt out choice was unavailable (a mixture of low- and high-confidence trials). The pattern was strikingly similar to findings in monkeys (Kiani & Shadlen 2009) and is consistent with the proposal that confidence is influenced by the dynamics of an evolving DV that also underlies the choice. A follow-up study by the same group used functional magnetic resonance imaging combined with EEG to localize a region of the ventromedial prefrontal cortex that reflects an early-arising confidence signal (Gherman & Philiastides 2018). The implication is that the early EEG correlate is not simply a reflection of post hoc sorting of trials but may actually be read out to establish a stimulus-independent representation of provisional confidence prior to decision termination. This conclusion was recently reinforced by Dou et al. (2024), who found that EEG decision signals predict confidence independently of accuracy, reaction time, and evidence strength.

Other investigators used a similar approach but with a focus on postdecision signals. Murphy et al. (2015) found that the EEG-derived DV continues to evolve after the initial choice commitment and predicts the timing of self-reported errors. The idea of continued accumulation after initial bound crossing (van Den Berg et al. 2016) raises the question of how two distinct termination rules can be applied to the same DV. An alternative is to posit separate accumulators for choice and confidence (Balsdon et al. 2021, Maniscalco et al. 2021). As alluded to above, this architecture has the intriguing property of allowing online confidence computations to regulate the decision process itself (Balsdon et al. 2020, Balsdon & Philiastides 2024). Anatomically, the recent study by Balsdon & Philiastides (2024) localizes the choice (motor) accumulator to the contralateral motor cortex, whereas the confidence (primary) accumulator is associated with central parietal positivity, a prominent signal that is the focus of most EEG studies of evidence accumulation. Inspired by these findings, future work in invasive preparations (human and animal) could evaluate the dual-accumulator hypothesis at the circuit and population levels, for instance, by testing for DV representations simultaneously evolving in orthogonal subspaces (e.g., Charlton & Goris 2024) and by comparing single-trial dynamics across multiple brain areas (Bondy et al. 2024, Khilkevich et al. 2024).

CONFIDENCE AND PERFORMANCE MONITORING

At the beginning of this review, we laid out examples of why perceptual confidence is critical for adaptive behavior in the moment (e.g., cross the street or not?). There is also a large literature on how the quality of decisions can be monitored (Ullsperger et al. 2014, Fu et al. 2023) to decide whether to maintain or adjust the current strategy or policy to adapt to changing environments (for a detailed review, see Egner 2023). Like perceptual confidence, performance monitoring involves tracking internally generated representations—task sets, goals, motor plans, efference copies—that are private to the individual brain, so it, too, certainly qualifies as metacognitive. Apart from both involving metacognitive monitoring of decisions, confidence judgments and performance monitoring appear to activate overlapping brain regions in humans, particularly the medial frontal cortex (MFC) (Shenhav et al. 2013, Ullsperger et al. 2014, Morales et al. 2018, Fu et al. 2023), raising the question: What is the relationship between the two?

A key function of performance monitoring is detecting errors or identifying failures to achieve internal goals. Errors lead to measurable behavioral adaptations, such as post-error slowing of reaction times on subsequent trials (Rabbitt 1966, Laming 1968), and evoke a robust signal in noninvasive EEG recordings, known as error-related negativity (Falkenstein et al. 1991, Gehring et al. 1993). The amplitude of error-related negativity correlates with neuronal firing in the MFC in both macaques (Sajad et al. 2019) and humans (Fu et al. 2019). While this error detection

function at face value might suggest conceptual overlap with confidence, there are fundamental differences. In error monitoring, the notion of correctness is derived from endogenous task set or goal representations, which are encoded in the prefrontal cortex and provide strong input to the MFC to compute error signals (Miller et al. 2002; Mian et al. 2014; Helfrich & Knight 2016; Sajad et al. 2019, 2022; Fu et al. 2023). This means that the ground truth against which performance is judged is internal, whereas perceptual confidence ought to ultimately relate to an external world state, notwithstanding that it is estimated by internal monitoring of the reliability of a decision process. However, confidence reports can also be elicited in scenarios where the ground truth is ambiguous or only partially accessible to decision-making circuits. For instance, people can rate their confidence in episodic memory retrieval (Rutishauser et al. 2015), value-based choices (De Martino et al. 2013), and perceptual tasks involving fully ambiguous stimuli, as we have described above. An intriguing question for future research is whether confidence in the absence of an external ground truth is computed similarly to the detection (or prediction) of errors relative to an internal goal or standard.

In the cognitive control literature, action selection refers to a process in which goal-directed and goal-irrelevant actions compete for behavioral output, analogous to the deliberation process during perceptual decision-making. Conflict monitoring theory (Botvinick et al. 2001, Shenhav et al. 2013) proposes that conflict between response options in action selection signals the need for enhanced cognitive control. This conflict, like errors, is explicitly monitored and serves as an internal feedback signal for cognitive control. Consistent with this theory, human and macaque single-unit recordings have found such conflict signals during (Sheth et al. 2012) and after (Sajad et al. 2019, 2022; Fu et al. 2022; Corrigan et al. 2024) action selection in the MFC. Such a monitoring process, which generates the error and conflict signals, could then inform domain-general confidence judgments and simultaneously inform adjustments of behavioral strategy. Supporting this, humans can rate the confidence of error occurrence in the action just performed, and an error positivity EEG signal (Van Veen & Carter 2002) is correlated with such confidence reports (Nieuwenhuis et al. 2001, Boldt & Yeung 2015).

Interestingly, a classical behavioral study provides evidence that performance monitoring and explicit report are at least partially dissociable (Logan & Crump 2010). Skilled typists performed a typing task with visual feedback and were later asked to decide whether errors had occurred or not. As they made errors, some errors were displayed as is, some were covertly corrected, and some fictitious errors on the screen were covertly inserted by the experimenter. Genuine errors made by the typist led to post-error slowing regardless of what was on the screen, yet the typist reported only errors that they visually perceived (both real and inserted). This suggests that while performance monitoring provides error signals for explicit report, it is outweighed by the detection of an error based on visual feedback. Perhaps if the visual feedback was degraded, the conscious report would rely more on error detection, analogous to how signals are weighted by reliability in sensory cue combination (Fetsch et al. 2013; Ernst & Banks 2002), but this speculation remains untested.

A key advantage of human experiments is the ability to perform multiple tasks with minimal training (Fu et al. 2022). Recording single neurons in humans in epilepsy monitoring settings provides a powerful way to test the generalizability of hypotheses generated by neuroimaging studies and nonhuman primate work across several tasks and thus complement such work (Fu et al. 2023). This platform creates new opportunities to evaluate the domain generality of confidence computations by eliciting judgments from a range of inputs, including both perceptual signals and performance monitoring. This approach not only complements the detailed mechanistic studies in macaques, which have traditionally focused on a narrower set of cognitive domains, but it has also already yielded fresh insights into the general principles of performance monitoring. The

representational geometry of MFC neurons enables error monitoring to generalize across different cognitive tasks by factorizing into a task dimension and error dimension (Fu et al. 2022). It remains to be tested whether a similar geometry may also support domain-general confidence based on performance monitoring and perception.

CONCLUDING REMARKS AND OPEN QUESTIONS

Seeking to understand how the brain evaluates the quality of its own perceptual judgments and goal-directed action plans is one of systems neuroscience's more ambitious undertakings. By combining research traditions from psychology and neuroscience in creative ways, considerable progress has been made in this endeavor (see the section titled Summary Points). Modern computational models of perceptual decision confidence articulate explicit hypotheses about the mental processes at work and provide quantitative predictions for confidence in binary perceptual decisions. These models capture many intriguing aspects of confidence reports, including the relationships between decision confidence and choice consistency and between decision confidence and reaction time. In doing so, these models can provide insight into the relative timing of decision formation and confidence assignment and illuminate the factors that constrain the quality of the sense of confidence. Importantly, key model components have been either directly or indirectly recognized in the neural activity of human and nonhuman subjects generating confidence-mediated behavior. In our view, we have now reached the point where the basic principles of the causal chain of neural events that transform photons into visual metacognition are understood for some simple perceptual tasks. Despite this progress, it is obvious that our current knowledge is still lacking in depth and breadth. The study of the neurobiological basis of visual confidence has so far been limited to a small number of perceptual tasks and brain areas. Much remains to be learned (see the section titled Future Issues). Future work should aim to identify how the neural mechanisms that compute perceptual decision confidence operate across different sensory modalities, confidence-mediated behaviors, and task contexts. The behavioral paradigms, computational models, and physiological discoveries highlighted in this review offer a promising starting point for this endeavor.

SUMMARY POINTS

1. Building on a rich history, the modern study of visual confidence primarily focuses on confidence in perceptual judgments of ambiguous stimuli. Confidence reports can be direct or indirect and incentivized or unincentivized.
2. Confidence experiments are conducted in humans and nonhuman animals such as monkeys and rats. Once fully trained, monkeys' behavioral responses can closely resemble human direct confidence reports.
3. Process models of decision confidence specify a causal chain of events that yield quantitative predictions for the relation between experimentally controlled variables and observed confidence reports. The literature is populated by a large variety of such models.
4. Static models that equate the sense of confidence to subjective decision reliability explain the relationship between choice consistency and decision confidence. These models also provide principled metrics for the quality of the sense of confidence (i.e., metacognitive ability).

5. Dynamic models that equate the sense of confidence to the balance of evidence across competing evidence accumulators explain the relationship between the time it takes to report a choice and decision confidence. These models also provide insight into the relative timing of decision and confidence computations.
6. A recent study showed that monkeys' perceptual orientation judgments and their confidence in these decisions can both be predicted from primary visual cortex population activity. The relationship between sensory responses and decision confidence differs from the one between sensory responses and decision content.
7. A recent study that employed a reaction time paradigm with simultaneous decision and confidence reports showed that neural signatures of decision formation and confidence assignment unfold simultaneously in the lateral intraparietal area of macaques.
8. The capacity to evaluate the quality of one's own perceptual decisions may be connected to the capacity to monitor internal action selection conflicts, an important aspect of cognitive control. This monitoring process is reflected in neural activity in the medial frontal cortex of humans and macaques.

FUTURE ISSUES

1. Can the strengths of current static and dynamic models be unified in a single process model that illuminates both the relative timing and quality of confidence computations?
2. How do the neural mechanisms underlying perceptual decision confidence operate across different sensory modalities, confidence-mediated behaviors, and task contexts?
3. In some tasks, a parallel confidence computation appears to guide sensory evidence sampling and integration. What is the neurophysiological substrate of this dynamic interplay between decision formation and confidence assignment?
4. Researchers have proposed several coding schemes by which populations of sensory neurons represent sensory uncertainty. Is there a mechanistic link between these probabilistic representations of sensory information and the subjective feeling of decision confidence?
5. How do neural representations of decision confidence guide changes in learning and subsequent decisions in hierarchical tasks and strategic planning?
6. What is the circuit-level relation between metacognitive monitoring of perceptual decision quality and action selection conflicts?

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported by the US National Science Foundation (CAREER award 2146369 to R.L.T.G.), the US National Institutes of Health (grants EY032999 to R.L.T.G. and RF1NS132910 to C.R.F.), the Whitehall Foundation (C.R.F.), the France-Merrick Foundation

(C.R.F.), and a NARSAD Young Investigator Award (Z.F.). The authors wish to thank Zoe Boundy-Singer and Corey Ziemba for their feedback on an earlier draft of the article and Miguel Vivar-Lazo for discussions.

LITERATURE CITED

Adams WJ, Graf EW, Ernst MO. 2004. Experience can change the 'light-from-above' prior. *Nat. Neurosci.* 7(10):1057–58

Adler WT, Ma WJ. 2018. Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLOS Comput. Biol.* 14(11):e1006572

Arnold DH, Johnston A, Adie J, Yarrow K. 2023. On why we lack confidence in some signal-detection-based analyses of confidence. *Conscious. Cogn.* 113:103532

Audley R. 1960. A stochastic model for individual choice behavior. *Psychol. Rev.* 67(1):1–15

Bahrami B, Olsen K, Latham PE, Roepstorff A, Rees G, Frith CD. 2010. Optimally interacting minds. *Science* 329(5995):1081–85

Balsdon T, Mamassian P, Wyart V. 2021. Separable neural signatures of confidence during perceptual decisions. *eLife* 10:e68491

Balsdon T, Philiastides MG. 2024. Confidence control for efficient behaviour in dynamic environments. *Nat. Commun.* 15:9089

Balsdon T, Wyart V, Mamassian P. 2020. Confidence controls perceptual evidence accumulation. *Nat. Commun.* 11(1):1753

Baranski JV, Petrusic WM. 1998. Probing the locus of confidence judgments: experiments on the time to determine confidence. *J. Exp. Psychol. Hum. Percept. Perform.* 24(3):929–45

Barthelmé S, Mamassian P. 2009. Evaluation of objective uncertainty in the visual system. *PLOS Comput. Biol.* 5(9):e1000504

Bertana A, Chetverikov A, van Bergen RS, Ling S, Jehee JFM. 2021. Dual strategies in human confidence judgments. *J. Vis.* 21(5):21

Bisley JW, Goldberg ME. 2010. Attention, intention, and priority in the parietal lobe. *Annu. Rev. Neurosci.* 33:1–21

Bloch AM. 1885. Experiences sur la vision. *C. R. Soc. Biol.* 37(28):493–95

Boldt A, Yeung N. 2015. Shared neural markers of decision confidence and error detection. *J. Neurosci.* 35(8):3478–84

Bondy AG, Charlton JA, Luo TZ, Kopec CD, Stagnaro WM, et al. 2024. Coordinated cross-brain activity during accumulation of sensory evidence and decision commitment. Preprint, bioRxiv. **https://www.biorxiv.org/content/10.1101/2024.08.21.609044v3.abstract**

Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD. 2001. Conflict monitoring and cognitive control. *Psychol. Rev.* 108(3):624–52

Boundy-Singer ZM, Ziemba CM, Goris RLT. 2023. Confidence reflects a noisy decision reliability estimate. *Nat. Hum. Behav.* 7(1):142–54

Boundy-Singer ZM, Ziemba CM, Goris RLT. 2025. Sensory population activity reveals downstream confidence computations in the primate visual system. *PNAS* 122(26):e2426441122

Boundy-Singer ZM, Ziemba CM, Hénaff OJ, Goris RLT. 2024. How does V1 population activity inform perceptual certainty? *J. Vis.* 24(6):12

Brunton BW, Botvinick MM, Brody CD. 2013. Rats and humans can optimally accumulate evidence for decision-making. *Science* 340(6128):95–98

Carpenter J, Sherman MT, Kievit RA, Seth AK, Lau H, Fleming SM. 2019. Domain-general enhancements of metacognitive ability through adaptive training. *J. Exp. Psychol. Gen.* 148(1):51–64

Caziot B, Mamassian P. 2021. Perceptual confidence judgments reflect self-consistency. *J. Vis.* 21(12):8

Charlton JA, Goris RLT. 2024. Abstract deliberation by visuomotor neurons in prefrontal cortex. *Nat. Neurosci.* 27:1167–75

Charlton JA, Młynarski WF, Bai YH, Hermundstad AM, Goris RLT. 2023. Environmental dynamics shape perceptual decision bias. *PLOS Comput. Biol.* 19:e1011104

Churchland AK, Kiani R, Chaudhuri R, Wang X-J, Pouget A, Shadlen MN. 2011. Variance as a signature of neural computations during decision making. *Neuron* 69:818–31

Corrigan BW, Errington SP, Sajad A, Schall JD. 2024. Conflict neurons in cingulate cortex of macaques. Preprint, bioRxiv. **https://www.biorxiv.org/content/10.1101/2024.10.03.616355v1.abstract**

Croner LJ, Purpura K, Kaplan E. 1993. Response variability in retinal ganglion cells of primates. *PNAS* 90(17):8128–30

Dayan P. 2023. Metacognitive information theory. *Open Mind Discov. Cogn. Sci.* 7:392–411

de Gardelle V, Mamassian P. 2014. Does confidence use a common currency across two visual tasks? *Psychol. Sci.* 25(6):1286–88

De Martino B, Fleming SM, Garrett N, Dolan RJ. 2013. Confidence in value-based choice. *Nat. Neurosci.* 16(1):105–10

DeAngelis GC. 2000. Seeing in three dimensions: the neurophysiology of stereopsis. *Trends Cogn. Sci.* 4(3):80–90

Denison RN, Adler WT, Carrasco M, Ma WJ. 2018. Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *PNAS* 115(43):11090–95

Desender K, Donner TH, Verguts T. 2021a. Dynamic expressions of confidence within an evidence accumulation framework. *Cognition* 207:104522

Desender K, Ridderinkhof KR, Murphy PR. 2021b. Understanding neural signals of post-decisional performance monitoring: an integrative review. *eLife* 10:e67556

DiCarlo JJ, Zoccolan D, Rust NC. 2012. How does the brain solve visual object recognition? *Neuron* 73(3):415–34

Dou W, Martinez Arango LJ, Castaneda OG, Arellano L, Mcintyre E, et al. 2024. Neural signatures of evidence accumulation encode subjective perceptual confidence independent of performance. *Psychol. Sci.* 35(7):760–79

Egner T. 2023. Principles of cognitive control over task focus and task switching. *Nat. Rev. Psychol.* 2(11):702–14

Ernst MO, Banks MS. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415(6870):429–33

Faisal AA, Selen LPJ, Wolpert DM. 2008. Noise in the nervous system. *Nat. Rev. Neurosci.* 9(4):292–303

Falkenstein M, Hohnsbein J, Hoormann J, Blanke L. 1991. Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalogr. Clin. Neurophysiol.* 78(6):447–55

Felleman DJ, Van Essen DC. 1991. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1(1):1–47

Festinger L. 1943. Studies in decision: I. Decision-time, relative frequency of judgment and subjective confidence as related to physical stimulus difference. *J. Exp. Psychol.* 32(4):291–306

Fetsch CR, DeAngelis GC, Angelaki DE. 2013. Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. *Nat. Rev. Neurosci.* 14(6):429–42

Fetsch CR, Kiani R, Newsome WT, Shadlen MN. 2014. Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron* 83(4):797–804

Fetsch CR, Odean NN, Jeurissen D, El-Shamayleh Y, Horwitz GD, Shadlen MN. 2018. Focal optogenetic suppression in macaque area MT biases direction discrimination and decision confidence, but only transiently. *eLife* 7:e36523

Fleming SM. 2024. Metacognition and confidence: a review and synthesis. *Annu. Rev. Psychol.* 75:241–68

Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G. 2010. Relating introspective accuracy to individual differences in brain structure. *Science* 329(5998):1541–43

Foote AL, Crystal JD. 2007. Metacognition in the rat. *Curr. Biol.* 17(6):551–55

Freedman DJ, Assad JA. 2006. Experience-dependent representation of visual categories in parietal cortex. *Nature* 443(7107):85–88

Fu Z, Beam D, Chung JM, Reed CM, Mamelak AN, et al. 2022. The geometry of domain-general performance monitoring in the human medial frontal cortex. *Science* 376(6593):eabm9922

Fu Z, Sajad A, Errington SP, Schall JD, Rutishauser U. 2023. Neurophysiological mechanisms of error monitoring in human and non-human primates. *Nat. Rev. Neurosci.* 24(3):153–72

Fu Z, Wu D-AJ, Ross I, Chung JM, Mamelak AN, et al. 2019. Single-neuron correlates of error monitoring and post-error adjustments in human medial frontal cortex. *Neuron* 101(1):165–77.e5
Gegenfurtner KR, Kiper DC. 2003. Color vision. *Annu. Rev. Neurosci.* 26:181–206
Gehring WJ, Goss B, Coles MGH, Meyer DE, Donchin E. 1993. A neural system for error detection and compensation. *Psychol. Sci.* 4(6):385–90
Geurts LS, Cooke JRH, van Bergen RS, Jehee JFM. 2022. Subjective confidence reflects representation of Bayesian probability in cortex. *Nat. Hum. Behav.* 6(2):294–305
Gherman S, Philiastides MG. 2015. Neural representations of confidence emerge from the process of decision formation during perceptual choices. *NeuroImage* 106:134–43
Gherman S, Philiastides MG. 2018. Human VMPFC encodes early signatures of confidence in perceptual decisions. *eLife* 7:e38293
Glaze CM, Kable JW, Gold JI. 2015. Normative evidence accumulation in unpredictable environments. *eLife* 4:e08825
Gold JI, Shadlen MN. 2007. The neural basis of decision making. *Annu. Rev. Neurosci.* 30:535–74
Goris RLT, Coen-Cagli R, Miller KD, Priebe NJ, Lengyel M. 2024. Response sub-additivity and variability quenching in visual cortex. *Nat. Rev. Neurosci.* 25(4):237–52
Goris RLT, Ziemba CM, Movshon JA, Simoncelli EP. 2018. Slow gain fluctuations limit benefits of temporal integration in visual cortex. *J. Vis.* 18(8):8
Green DM, Swets JA. 1966. *Signal Detection Theory and Psychophysics*, Vol. 1. Wiley
Guggenmos M. 2021. Measuring metacognitive performance: type 1 performance dependence and test-retest reliability. *Neurosci. Conscious.* 2021(1):niab040
Hahn M, Wei XX. 2024. A unifying theory explains seemingly contradictory biases in perceptual estimation. *Nat. Neurosci.* 27(4):793–804
Hampton RR. 2001. Rhesus monkeys know when they remember. *PNAS* 98(9):5359–62
Hanks TD, Mazurek ME, Kiani R, Hopp E, Shadlen MN. 2011. Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *J. Neurosci.* 31(17):6339–52
Helfrich RF, Knight RT. 2016. Oscillatory dynamics of prefrontal cognitive control. *Trends Cogn. Sci.* 20(12):916–30
Hénaff OJ, Bai Y, Charlton J, Nauhaus I, Simoncelli EP, Goris RLT. 2021. Primary visual cortex straightens natural video trajectories. *Nat. Commun.* 12(1):5982
Hénaff OJ, Boundy-Singer ZM, Meding K, Ziemba CM, Goris RLT. 2020. Representation of visual uncertainty through neural gain variability. *Nat. Commun.* 11(1):2513
Hénaff OJ, Goris RLT, Simoncelli EP. 2019. Perceptual straightening of natural videos. *Nat. Neurosci.* 22(6):984–91
Henmon VAC. 1911. The relation of the time of a judgment to its accuracy. *Psychol. Rev.* 18(3):186–201
Hubel DH, Wiesel TN. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160(1):106–54
Johnson DM. 1939. *Confidence and speed in the two-category judgment*. MS Thesis, Columbia University
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(1594):1322–37
Kepecs A, Uchida N, Zariwala HA, Mainen ZF. 2008. Neural correlates, computation and behavioural impact of decision confidence. *Nature* 455(7210):227–31
Khalvati K, Kiani R, Rao RPN. 2021. Bayesian inference with incomplete knowledge explains perceptual confidence and its deviations from accuracy. *Nat. Commun.* 12(1):5704
Khilkevich A, Lohse M, Low R, Orsolic I, Bozic E, et al. 2024. Brain-wide dynamics linking sensation to action during decision-making. *Nature* 634:890–900
Kiani R, Corthell L, Shadlen MN. 2014. Choice certainty is informed by both evidence and decision time. *Neuron* 84(6):1329–42
Kiani R, Shadlen MN. 2009. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* 324(5928):759–64
Kilpatrick ZP, Holmes WR, Eissa TL, Josić K. 2019. Optimal models of decision-making in dynamic environments. *Curr. Opin. Neurobiol.* 58:54–60

Kira S, Yang T, Shadlen MN. 2015. A neural implementation of Wald's sequential probability ratio test. *Neuron* 85(4):861–73
Knill DC, Pouget A. 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27(12):712–19
Knill DC, Richards W, eds. 1996. *Perception as Bayesian Inference*. Cambridge University Press
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(6):749–55
Koriat A. 2012. The self-consistency model of subjective confidence. *Psychol. Rev.* 119(1):80–113
Laming DRJ. 1968. *Information Theory of Choice-Reaction Times*. Academic Press
Langlois TA, Charlton JA, Goris RLT. 2025. Bayesian inference by visuomotor neurons in the prefrontal cortex. *PNAS* 122(13):e2420815122
Latimer KW, Freedman DJ. 2023. Low-dimensional encoding of decisions in parietal cortex reflects long-term training history. *Nat. Commun.* 14(1):1010
Le Denmat P, Verguts T, Desender K. 2024. A low-dimensional approximation of optimal confidence. *PLOS Comput. Biol.* 20(7):e1012273
LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521(7553):436–44
Li H-H, Ma WJ. 2020. Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. *Nat. Commun.* 11(1):2004
Locke SM, Gaffin-Cahn E, Hosseinizaveh N, Mamassian P, Landy MS. 2020. Priors and payoffs in confidence judgments. *Atten. Percept. Psychophys.* 82(6):3158–75
Logan GD, Crump MJC. 2010. Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. *Science* 330(6004):683–86
Mamassian P. 2016. Visual confidence. *Annu. Rev. Vis. Sci.* 2:459–81
Mamassian P, de Gardelle V. 2022. Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychol. Rev.* 129(5):976–98
Maniscalco B, Lau H. 2012. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn.* 21(1):422–30
Maniscalco B, Odegaard B, Grimaldi P, Cho SH, Basso MA, et al. 2021. Tuned inhibition in perceptual decision-making circuits can explain seemingly suboptimal confidence behavior. *PLOS Comput. Biol.* 17(3):e1008779
Mante V, Sussillo D, Shenoy KV, Newsome WT. 2013. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* 503(7474):78–84
Mareschal I, Shapley RM. 2004. Effects of contrast and size on orientation discrimination. *Vis. Res.* 44(1):57–67
Meister MLR, Hennig JA, Huk AC. 2013. Signal multiplexing and single-neuron computations in lateral intraparietal area during decision-making. *J. Neurosci.* 33(6):2254–67
Meyniel F, Sigman M, Mainen ZF. 2015. Confidence as Bayesian probability: from neural origins to behavior. *Neuron* 88(1):78–92
Mian MK, Sheth SA, Patel SR, Spiliopoulos K, Eskandar EN, Williams ZM. 2014. Encoding of rules by neurons in the human dorsolateral prefrontal cortex. *Cereb. Cortex* 24(3):807–16
Middlebrooks PG, Sommer MA. 2012. Neuronal correlates of metacognition in primate frontal cortex. *Neuron* 75(3):517–30
Mihali A, Broeker M, Ragalmuto FDM, Horga G. 2023. Introspective inference counteracts perceptual distortion. *Nat. Commun.* 14:7826
Miller EK, Freedman DJ, Wallis JD. 2002. The prefrontal cortex: categories, concepts and cognition. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 357(1424):1123–36
Mineault PJ, Khawaja FA, Butts DA, Pack CC. 2012. Hierarchical processing of complex motion along the primate dorsal visual pathway. *PNAS* 109(16):E972–80
Młynarski WF, Hermundstad AM. 2018. Adaptive coding for dynamic sensory inference. *eLife* 7:e32055
Morales J, Lau H, Fleming SM. 2018. Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *J. Neurosci.* 38(14):3535–46
Moran R, Teodorescu AR, Usher M. 2015. Post choice information integration as a causal determinant of confidence: novel data and a computational account. *Cogn. Psychol.* 78:99–147

Murphy PR, Robertson IH, Harty S, O'Connell RG. 2015. Neural evidence accumulation persists after choice to inform metacognitive judgments. *eLife* 4(6593):e11946

Navajas J, Bahrami B, Latham PE. 2016. Post-decisional accounts of biases in confidence. *Curr. Opin. Behav. Sci.* 11:55–60

Navajas J, Hindocha C, Foda H, Keramati M, Latham PE, Bahrami B. 2017. The idiosyncratic nature of confidence. *Nat. Hum. Behav.* 1(11):810–18

Nelson TO. 1984. A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychol. Bull.* 95(1):109–33

Nieuwenhuis S, Ridderinkhof KR, Blom J, Band GPH, Kok A. 2001. Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology* 38(5):752–60

Norton EH, Acerbi L, Ma WJ, Landy MS. 2019. Human online adaptation to changes in prior probability. *PLOS Comput. Biol.* 15:e1006681

O'Connell RG, Kelly SP. 2021. Neurophysiology of human perceptual decision-making. *Annu. Rev. Neurosci.* 44:495–516

Osborne LC, Bialek W, Lisberger SG. 2004. Time course of information about motion direction in visual area MT of macaque monkeys. *J. Neurosci.* 24(13):3210–22

Peirce CS, Jastrow J. 1884. On small differences in sensation. *Mem. Natl. Acad. Sci.* 3:73–83

Peixoto D, Verhein JR, Kiani R, Kao JC, Nuyujukian P, et al. 2021. Decoding and perturbing decision states in real time. *Nature* 591(7851):604–9

Persaud N, McLeod P, Cowey A. 2007. Post-decision wagering objectively measures awareness. *Nat. Neurosci.* 10(2):257–61

Peters MAK, Thesen T, Ko YD, Maniscalco B, Carlson C, et al. 2017. Perceptual confidence neglects decision-incongruent evidence in the brain. *Nat. Hum. Behav.* 1(7):0139

Pleskac TJ, Busemeyer JR. 2010. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychol. Rev.* 117(3):864–901

Pouget A, Drugowitsch J, Kepecs A. 2016. Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* 19(3):366–74

Purcell BA, Kiani R. 2016. Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *PNAS* 113(31):E4531–40

Rabbitt PM. 1966. Errors and error correction in choice-response tasks. *J. Exp. Psychol.* 71(2):264–72

Rahnev D. 2025. A comprehensive assessment of current methods for measuring metacognition. *Nat. Commun.* 16(1):701

Raposo D, Sheppard JP, Schrater PR, Churchland AK. 2012. Multisensory decision-making in rats and humans. *J. Neurosci.* 32(11):3726–35

Ratcliff R, Rouder JN. 1998. Modeling response times for two-choice decisions. *Psychol. Sci.* 9(5):347–56

Rausch M, Zehetleitner M, Steinhauser M, Maier ME. 2020. Cognitive modeling reveals distinct electrophysiological markers of decision confidence and error monitoring. *NeuroImage* 218:116963

Roitman JD, Shadlen MN. 2002. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci.* 22(21):9475–89

Rouy M, de Gardelle V, Reyes G, Sackur J, Vergnaud J, et al. 2022. Metacognitive improvement: disentangling adaptive training from experimental confounds. *J. Exp. Psychol. Gen.* 151(9):1939–2222

Rutishauser U, Ye S, Koroma M, Tudusciuc O, Ross IB, et al. 2015. Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nat. Neurosci.* 18(7):1041–50

Sajad A, Errington SP, Schall JD. 2022. Functional architecture of executive control and associated event-related potentials in macaques. *Nat. Commun.* 13(1):6270

Sajad A, Godlove DC, Schall JD. 2019. Cortical microcircuitry of performance monitoring. *Nat. Neurosci.* 22(2):265–74

Sanders JI, Hangya B, Kepecs A. 2016. Signatures of a statistical computation in the human sense of confidence. *Neuron* 90(3):499–506

Sarafyazd M, Jazayeri M. 2019. Hierarchical reasoning by neural circuits in the frontal cortex. *Science* 364(6441):eaav8911

Schiller PH, Finlay BL, Volman SF. 1976. Short-term response variability of monkey striate neurons. *Brain Res*. 105(2):347–49
Shadlen MN, Britten KH, Newsome WT, Movshon JA. 1996. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J. Neurosci.* 76(4):1486–10
Shadlen MN, Kiani R. 2013. Decision making as a window on cognition. *Neuron* 80(3):791–806
Shadlen MN, Newsome WT. 1998. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J. Neurosci.* 18(10):3870–96
Shekhar M, Rahnev D. 2021. The nature of metacognitive inefficiency in perceptual decision making. *Psychol. Rev.* 128(1):45–70
Shenhav A, Botvinick MM, Cohen JD. 2013. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron* 79(2):217–40
Sheth SA, Mian MK, Patel SR, Asaad WF, Williams ZM, et al. 2012. Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation. *Nature* 488(7410):218–21
Simoncelli EP, Heeger DJ. 1998. A model of neuronal responses in visual area MT. *Vis. Res.* 38(5):743–61
Smith JD. 2009. The study of animal metacognition. *Trends Cogn. Sci.* 13(9):389–96
Smith JD, Shields WE, Schull J, Washburn DA. 1997. The uncertain response in humans and animals. *Cognition* 62(1):75–97
Snyder LH, Batista AP, Andersen RA. 2000. Intention-related activity in the posterior parietal cortex: a review. *Vis. Res.* 40(10–12):1433–41
Steinemann N, Stine GM, Trautmann E, Zylberberg A, Wolpert DM, Shadlen MN. 2024. Direct observation of the neural computations underlying a single decision. *eLife* 12:RP90859
Stocker AA, Simoncelli EP. 2006. Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* 9(4):578–85
Stone M. 1960. Models for choice-reaction time. *Psychometrika* 25(3):251–60
Strasburger H, Rentschler I, Jüttner M. 2011. Peripheral vision and pattern recognition: a review. *J. Vis.* 11(5):13
Tanner WP Jr., Birdsall TG, Clarke F. 1960. *The concept of the ideal observer in psychophysics*. Tech. Rep. 98, University of Michigan. **https://deepblue.lib.umich.edu/bitstream/handle/2027.42/7891/bad2543.0001.001.pdf?sequence=5**
Tanner WP Jr., Swets JA. 1954. A decision-making theory of visual detection. *Psychol. Rev.* 61(6):401–9
Treisman M, Faulkner A. 1984. The setting and maintenance of criteria representing levels of confidence. *J. Exp. Psychol. Hum. Percept. Perform.* 10(1):119–39
Tuckute G, Kanwisher N, Fedorenko E. 2024. Language in brains, minds, and machines. *Annu. Rev. Neurosci.* 47:277–301
Ullsperger M, Danielmeier C, Jocham G. 2014. Neurophysiology of performance monitoring and adaptive behavior. *Physiol. Rev.* 94(1):35–79
van Den Berg R, Anandalingam K, Zylberberg A, Kiani R, Shadlen MN, Wolpert DM. 2016. A common mechanism underlies changes of mind about decisions and confidence. *eLife* 5:e12192
Van Essen DC, Gallant JL. 1994. Neural mechanisms of form and motion processing in the primate visual system. *Neuron* 13(1):1–10
Van Veen V, Carter CS. 2002. The timing of action-monitoring processes in the anterior cingulate cortex. *J. Cogn. Neurosci.* 14(4):593–602
Vickers D. 1979. *Decision Processes in Visual Perception*. Academic Press
Vickers D, Nettelbeck T, Willson RJ. 1972. Perceptual indices of performance: the measurement of 'inspection time' and 'noise' in the visual system. *Perception* 1(3):263–95
Vivar-Lazo M, Fetsch CR. 2025. Neural basis of concurrent deliberation toward a choice and confidence judgment. *Nat. Neurosci*. In press; Vivar-Lazo M, Fetsch CR. 2025. Neural basis of concurrent deliberation toward a choice and confidence judgment. Preprint, bioRxiv. **https://www.biorxiv.org/content/10.1101/2024.08.06.606833v3**
von Helmholtz H. 1948. Concerning the perceptions in general, 1867. In *Readings in the History of Psychology*, ed. W Dennis. Appleton-Century-Crofts
Vuorre M, Metcalfe J. 2022. Measures of relative metacognitive accuracy are confounded with task performance in tasks that permit guessing. *Metacogn. Learn.* 17(2):269–91

Watson AB. 1979. Probability summation over time. *Vis. Res.* 19(5):515–22
Webb TW, Miyoshi K, So TY, Rajananda S, Lau H. 2023. Natural statistics support a rational account of confidence biases. *Nat. Commun.* 14(1):3992
Weiss Y, Simoncelli EP, Adelson EH. 2002. Motion illusions as optimal percepts. *Nat. Neurosci.* 5(6):598–604
Xue K, Shekhar M, Rahnev D. 2021. Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. *Conscious. Cogn.* 95:103196
Xue K, Shekhar M, Rahnev D. 2024. Challenging the Bayesian confidence hypothesis in perceptual decision-making. *PNAS* 121(48):e2410487121
Xue K, Zheng Y, Rafiei F, Rahnev D. 2023. The timing of confidence computations in human prefrontal cortex. *Cortex* 168:167–75
Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS* 111(23):8619–24
Zylberberg A. 2021. Decision prioritization and causal reasoning in decision hierarchies. *PLOS Comput. Biol.* 17(12):e1009688
Zylberberg A, Shadlen MN. 2025. A population representation of the confidence in a decision in the parietal cortex. *Cell Rep.* 44(4):115526