Confidence as a noisy decision reliability estimate

Zoe M. Boundy-Singer[1,*], **Corey M. Ziemba**[1,*], **Robbe L. T. Goris**[1]

[1] Center for Perceptual Systems, The University of Texas at Austin, Austin, TX USA. * These authors contributed equally

Decisions vary in difficulty. Humans know this and typically report more confidence in easy than in difficult decisions. However, confidence reports do not perfectly track decision accuracy, but also reflect response biases and difficulty misjudgments. To isolate the quality of confidence reports, we developed a model of the decision-making process underlying choice-confidence data. In this model, confidence reflects a subject's estimate of the reliability of their decision. The quality of this estimate is limited by the subject's uncertainty about the uncertainty of the variable that informs their decision ("meta-uncertainty"). This model provides an accurate account of choice-confidence data across a broad range of perceptual and cognitive tasks, revealing that meta-uncertainty varies across subjects, is stable over time, generalizes across some domains, and can be manipulated experimentally. The model offers a parsimonious explanation for the computational processes that underlie and constrain the sense of confidence.

Humans are aware of the fallibility of perception and cognition. When we experience a high degree of confidence in a perceptual or cognitive decision, that decision is more likely to be correct than when we feel less confident[1]. This "metacognitive" ability helps us to learn from mistakes[2], to plan future actions[3], and to optimize group decision-making[4]. There is a long-standing interest in the mental operations underlying our sense of confidence[5–7], and the rapidly expanding field of metacognition seeks to understand how metacognitive ability varies across domains[8], individuals[9], clinical states[10], and development[11].

Quantifying a subject's ability to introspect about the correctness of a decision is a challenging problem[12–14]. There exists no generally agreed-upon method[15]. Even in the simplest decision-making tasks, several distinct factors influence a subject's confidence reports. Consider a subject jointly reporting a binary decision about a sensory stimulus (belongs to "Category A" vs "Category B") and their confidence in this decision. Confidence reports will reflect the subject's ability to discriminate between both stimulus categories – the higher this ability, the higher the reported confidence[16]. They will also reflect the subject's response bias (e.g., a large willingness to declare "high confidence" in their decisions)[17–19]. Yet, neither of these factors characterizes the subject's metacognitive ability[13].

Here, we introduce a new method to quantify metacognitive ability on the basis of choice-confidence data. We propose that confidence reflects a subject's estimate of the reliability of their decision[20], expressed in units of signal-to-noise ratio. This estimate results from a computation involving the uncertainty of the decision variable that informed the subject's choice[21]. It follows that metacognitive ability is determined by the subject's knowledge about this uncertainty, or lack thereof (i.e., uncertainty about uncertainty, hereafter termed "meta-uncertainty"). The more certain a subject is about the uncertainty of the decision variable, the lower their meta-uncertainty, and the better they are able to assess the reliability of a decision. We leverage modern computational techniques to formalize this hypothesis in a two-stage process model that is rooted in traditional signal detection theory[22] and that can be fit to choice-confidence data (the "CASANDRE" or "Confidence AS A Noisy Decision Reliability Estimate" model). The model predicts a systematic dependency of confidence on choice consistency[20] and naturally separates metacognitive ability from discrimination ability and response bias.

We found that this model provides an excellent account of choice-confidence data reported in a large set of previously published studies[23–28]. Our analysis suggests that meta-uncertainty provides a better metric for metacognitive ability than the non-process-model based alternatives that currently prevail in the literature[13,15]. Specifically, meta-uncertainty has higher test-retest reliability, is less affected by discrimination ability and response bias, and has higher cross-domain generalizability. We show that our method differentiates experimental manipulations that affect meta-uncertainty from those that do not. Together, these results illuminate the mental operations that give rise to our sense of confidence, and they suggest that metacognitive ability is fundamentally limited by subjects' uncertainty about the reliability of their decisions.

Results

In simple decision-making tasks, human confidence reports lawfully reflect choice consistency[20]. Consider two example subjects who performed a two-alternative forced choice (2-AFC) categorization task in which they judged on every trial whether a visual stimulus belonged to category A or B, and additionally reported their confidence in this decision using a four-point rating scale. Categories were characterized by distributions of stimulus orientation that were predominantly smaller (A) or larger (B) than zero degrees. Stimuli varied in orientation and contrast (Fig. 1a). Because the category distributions overlap, errors are

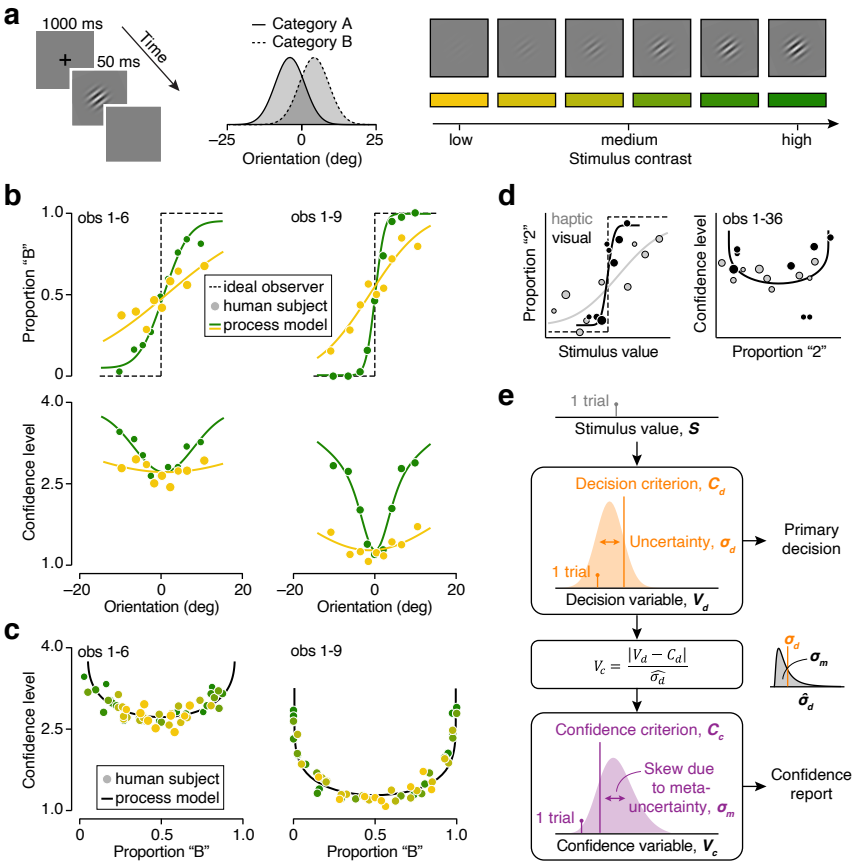


Figure 1 A two-stage process model of decision confidence accounts for the relation between confidence reports and choice consistency. (**a**) Experimental design employed by ref.[26]. (**b**) Top: Proportion of "Category B" choices is plotted against stimulus orientation, split out by stimulus contrast (green vs yellow), for two example subjects (left vs right). Bottom: Same for mean confidence level. Symbols summarize observed choice behavior, the dotted line illustrates the theoretical optimum, and the full lines show the fit of the CASANDRE model. Symbol size is proportional to the number of trials. All conditions are shown in Supplementary Fig. 1. (**c**) Observed and predicted confidence-consistency relationship for two example subjects. (**d**) Observed and predicted choice-confidence data for an example subject performing a visuo-haptic two-interval forced choice (2-IFC) categorization task (see Arbuzova and Filevich in the Confidence Database[23]). (**e**) Schematic of the hierarchical decision-making process underlying choice-confidence data in the CASANDRE model.

inevitable. The most accurate strategy is to choose category A for all stimuli whose orientation is smaller than zero degrees, and category B for all stimuli whose orientation exceeds zero degrees (Fig. 1b, top, dotted line). As can be seen from the aggregated choice behavior, the more the stimulus orientation deviates from zero, the more closely human subjects approximate this ideal (Fig. 1b, top, symbols). As can also be seen, this relationship is modulated by stimulus contrast – the lower the stimulus contrast, the weaker the association between orientation and choice (Fig. 1b, top, green vs yellow symbols). The distinct effects of orientation and contrast on choice consistency are evident in the subjects' confidence reports. Confidence is minimal for conditions associated with a choice proportion near 0.5 (i.e., the most difficult conditions), and monotonically increases as choice proportions deviate more from 0.5 (Fig. 1b, bottom). We found that the association between choice consistency and confidence is so strong, that plotting average confidence level against the aggregated choice behavior reveals a single relationship across all stimulus conditions (Fig. 1c). This is true of both example subjects, although their confidence-consistency relationships differ in shape, offset, and range. A lawful confidence-consistency relationship is not a coincidental feature of this experiment, but a widespread phenomenon in confidence studies (Fig. 1d).

A monotonically increasing relation between confidence reports and choice consistency implies that subjects can assess the reliability of their decisions. However, whether their ability to do so is excellent or poor cannot be deduced from empirical measurements alone. One possibility is that subjects accurately assess decision reliability on every single trial. This would indicate excellent metacognitive ability. Alternatively, the confidence-consistency relationship may only become evident across

many repeated measurements due to a high degree of variability in confidence reports. This would be indicative of limited metacognitive ability. Of course, given the variability of the primary choice behavior, some variability in confidence reports is expected, even for flawless introspection. How much exactly? And what might be the origin of excess variance? Answering these questions requires a quantitative model that provides an analogy for the mental operations that underlie a subject's primary decisions and confidence reports. In the following section, we develop such a process model.

A two-stage process model of decision-making

Assume that a subject solves a binary decision-making task by comparing a noisy, one-dimensional decision variable, V_d, to a fixed criterion, C_d (Fig. 1e, top). For some tasks, it is convenient to think of this decision variable as representing a direct estimate of a stimulus feature (e.g., orientation for the task shown in Fig. 1a). For other tasks, it is more appropriate to think of it as representing the accumulated evidence that favors one response alternative over the other (e.g., *"Have I heard this song before?"*). Regardless of the interpretation of the decision variable, the process model specified by these assumptions has proven very useful in the study of perception and cognition. It readily explains why repeated presentations of the same stimulus often elicit variable choices. In doing so, it clarifies how choices reflect a subject's underlying ability to solve the task as well as their response bias[22].

We expand this framework with an analogous second processing stage that informs the subject's confidence report. Assume that the subject is presented with a set of stimuli that elicit the same level of cross-trial variability in the decision variable. The smaller the overlap of the stimulus-specific decision variable distribution with the decision criterion, the "stronger" the associated stimulus is, and the more consistent choices will be. On any given trial, the distance between the decision variable and the decision criterion provides an instantaneously available proxy for stimulus strength, and hence for choice reliability[29–33]. However, in many tasks, the decision variable's dispersion, σ_d, will vary across conditions, resulting in different amounts of stimulus "uncertainty" (the larger σ_d, the greater this uncertainty). To be a useful proxy for choice reliability, the stimulus strength estimate must therefore be normalized by this factor[20]. This operation yields a unitless, positive-valued variable, V_c, which represents the subject's confidence in the decision:

$$V_c = \frac{|V_d - C_d|}{\hat{\sigma_d}} \tag{1}$$

where V_d is the decision variable, C_d the decision criterion, and $\hat{\sigma_d}$ the subject's estimate of σ_d. We assume that the subject is unsure about the exact level of stimulus uncertainty. Repeated trials will thus not only elicit different values of the decision variable, but will also elicit different estimates of stimulus uncertainty. Specifically, we assume that $\hat{\sigma_d}$ is on average correct (i.e., its mean value equals σ_d), but varies from trial-to-trial with standard deviation σ_m, resulting in "meta-uncertainty" (the larger σ_m, the greater this meta-uncertainty). As we shall see, variability in the decision variable is the critical model component that limits stimulus discriminability, while variability in the uncertainty estimate similarly limits metacognitive ability. Finally, comparing the confidence variable with a fixed criterion, C_c, yields a confidence report (Fig. 1e, bottom).

To fit this model to data, the form of the noise distributions must be specified. A common choice for the first stage noise is the normal distribution. This choice is principled, as the normal distribution is the maximum entropy distribution for real-valued signals with a specified mean and variance[34]. It is also convenient, as it results in fairly simple data-analysis recipes[22]. The second stage noise describes variability of a positive-valued signal (σ_d cannot be smaller than zero by definition). A suitable maximum entropy distribution for such a variable is the log-normal distribution[34,35]. Under these assumptions, the confidence variable is a probability distribution constructed as the distribution of the ratio of a normally and log-normally distributed variable. There exists no closed form description of this ratio distribution, ruling out simple data-analysis recipes. However, we can leverage modern computational tools to quickly compute the confidence variable's probability density function by describing it as a mixture of Gaussian distributions (see Methods). This mathematical street-fighting maneuver enables us to fit this two-stage process model to choice data (Fig. 1b-d, full lines). Before doing so, we first derive a set of basic model predictions.

Deriving model predictions

To gain a deeper understanding of the impact of the different model components on confidence reports, we investigated the model's behavior in a continuous 2-AFC discrimination task with binary confidence report options ("confident" or "not confident"). We assumed the decision variable's mean value to be stimulus-dependent (in this simulation, it is identical to the true stimulus value). All other model components were varied independently of the stimulus (see Methods). Altering the first-stage decision criterion (Fig. 2a, top left, orange vs grey line) affects the confidence variable distribution by shifting its mode and, in the presence of meta-uncertainty, its spread and skew (Fig. 2a, bottom left, purple vs grey distribution). At the level of ob-

servables, this manipulation results in a horizontal shift of the "psychometric function" that characterizes how choices depend on stimulus value (Fig. 2a, top right). This shift is accompanied by an identical shift of the "confidence function" that characterizes how confidence reports depend on stimulus value (Fig. 2a, bottom right). Effects of this kind have been documented for human[20,36,37] and animal[38,39] subjects. Altering the level of first stage noise (Fig. 2b, top left, orange vs grey distribution) affects the confidence variable distribution by changing its mode and, in the presence of meta-uncertainty, its spread and skew (Fig. 2b, bottom left, purple vs grey distribution). At the level of choice behavior, this manipulation changes the slope of the psychometric function (Fig. 2b, top right) as well as the overall fraction of "confident" reports (Fig. 2b, bottom right). In contrast, the parameters that control the model's second stage operations do not affect the primary decision behavior but only the confidence reports. Specifically, changing the confidence criterion (Fig. 2c, bottom left, purple vs grey lines) mainly impacts the confidence function by shifting it vertically (Fig. 2c, bottom right). Changing the level of meta-uncertainty alters the confidence variable distribution's mode, variance, and skew (Fig. 2d, bottom left, purple vs grey distribution), resulting in a complex pattern of changes in the confidence function (Fig. 2d, bottom right).

What does it mean to say that someone has good or bad self-knowledge? The CASANDRE model provides a principled answer that is at once intuitive and revealing. Everything held equal, increasing meta-uncertainty makes the confidence variable distribution more heavy-tailed (Fig. 2d, bottom left). This in turn leads to an increase in the fraction of "confident" reports for weak stimuli, but has the opposite effect for strong stimuli (Fig. 2d, bottom right). As a consequence, the dynamic range of the confidence-consistency relation decreases (Fig. 2e). However, these effects are not balanced. In particular, when meta-uncertainty is high, there is a dramatic increase in "confident" reports for the most difficult conditions (Fig. 2e, full black line). This increase does not reflect an actual change in task performance (Fig. 2d, top right). Rather, the association between confidence and choice consistency has weakened. This can be appreciated by inspecting the psychometric function split out by confidence report. When meta-uncertainty is low, "confident" decisions tend to be much more reliable than "not confident" decisions (Fig. 2f, left, green vs red). As meta-uncertainty increases, this distinction weakens and eventually disappears (Fig. 2f, middle-right). In sum, under the CASANDRE model, a lack of self-knowledge means having a limited capacity to distinguish reliable from unreliable decisions (note that this is not the same as distinguishing correct from incorrect decisions)[20]. This is a practical and useful insight. However, the magnitude of the effects shown in Fig. 2e,f depends on the other model components as well (e.g., Fig. 2g). These components will rarely be constant across tasks, individuals, or the life-span. Determining the level of meta-uncertainty therefore requires directly fitting the model to data. Finally, note that whenever we report values for σ_m, we use the coefficient of variation (σ_m/σ_d), as this ratio is identifiable under the model (the absolute level of meta-uncertainty is not).

Connecting the model to data

We seek to quantify a subject's ability to introspect about the reliability of a decision in simple decision-making tasks. Our method consists of interpreting human choice-confidence data through the lens of a principled two-stage process model of decision-making. This undertaking may fail to yield meaningful results for two reasons. First, it is not a given that conventional experiments provide the appropriate set of measurements to identify the parameters that control the model's behavior. This is a statistical estimation issue. Second, it remains to be seen whether the CASANDRE model provides a useful analog for the mental operations that underlie human choice-confidence data. After all, *all* models are wrong, and only *some* are useful[40]. This is a fundamental issue. To address these concerns, we investigated the performance of our method in synthetic best-case scenarios and in real bench-marking experiments.

What kind of measurements are required to obtain robust and reliable estimates of the model's parameters? We used the model to generate synthetic data sets for five model subjects performing a continuous 2-AFC discrimination task with binary confidence report options (see Methods). The model subjects only differed in their level of meta-uncertainty, which ranged from negligible to considerable (Fig. 3a, colored lines). Empirical studies of confidence employ a variety of experimental designs. Two key aspects that differ widely across studies are the number of trials per subject and the number of unique conditions. To investigate the consequences of these design choices, we simulated data for each model subject using experimental designs that varied in the number of trials (100 vs 1,000) and in the number of conditions (2 vs 20; Fig. 3a, top). Figure 3b summarizes an example synthetic experiment. The model parameters (σ_d, C_d, σ_m, C_c) specify the relation between stimulus value and the probability of each response option ("Category A – confident", "Category A – not confident", "Category B – not confident", and "Category B – confident"; Fig. 3b, left). We used these probabilities to simulate a synthetic data set of 1,000 trials distributed across 20 conditions (Fig. 3b, middle). We then identified the set of parameter values that best describes these data using a maximum likelihood estimation method (Fig. 3b, right). We repeated this procedure 100 times for each simulated experiment.

Our method yields robust estimates of meta-uncertainty: for all model subjects and all experimental designs, the median estimate closely approximates the ground truth value (Fig. 3c, symbols). The reliability of these estimates is much higher when they are based on 1,000 instead of 100 trials (Fig. 3c, error bars in right vs left column). The reliability also tends to be higher

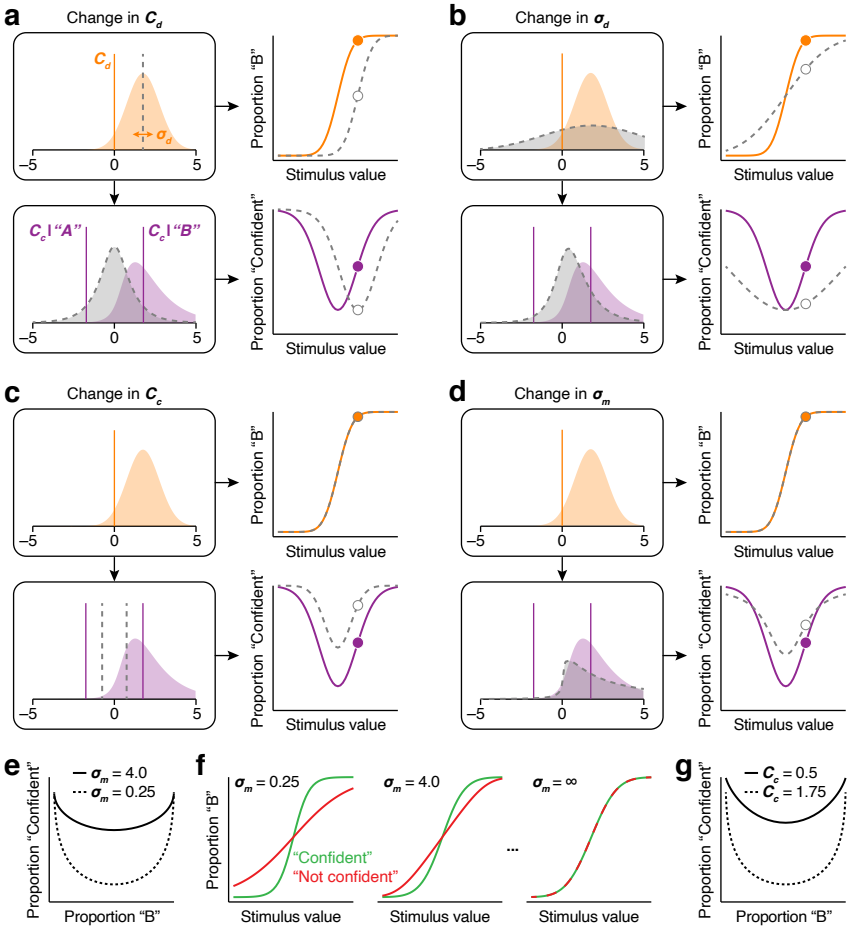


Figure 2 Impact of the different model components on primary choice behavior and confidence reports. (**a**) Top left: illustration of the decision criterion (orange line) and the decision variable distribution elicited by repeated presentations of the same stimulus (orange distribution). Bottom left: the associated confidence variable distribution (purple distribution). V_c is a positive-valued variable. As plotting convention, we reserve negative values for "Category A" choices, and positive values for "Category B" choices. The confidence criterion (purple line) therefore shows up twice in this graph. Top right: the resulting psychometric function over a range of stimulus values (orange line). The filled symbol corresponds to the condition depicted on the left hand side. Bottom right: same for the resulting confidence function. All panels: the grey dotted line illustrates how the model predictions change when the decision criterion is altered. The open symbol corresponds to the condition depicted on the left hand side. (**b**) Increasing the level of stimulus uncertainty affects both primary decisions and confidence reports. (**c**) Lowering the confidence criterion yields more "confident" reports at all stimulus values. (**d**) Increasing meta-uncertainty increases the fraction of "confident" reports for weak stimuli, but has the opposite effect for strong stimuli. (**e**) The confidence-consistency relation for two levels of meta-uncertainty. (**f**) The psychometric function, split out by confidence report ("confident" in green vs "not confident" in red), for three levels of meta-uncertainty. (**g**) The confidence-consistency relation under a liberal vs a conservative confidence criterion.

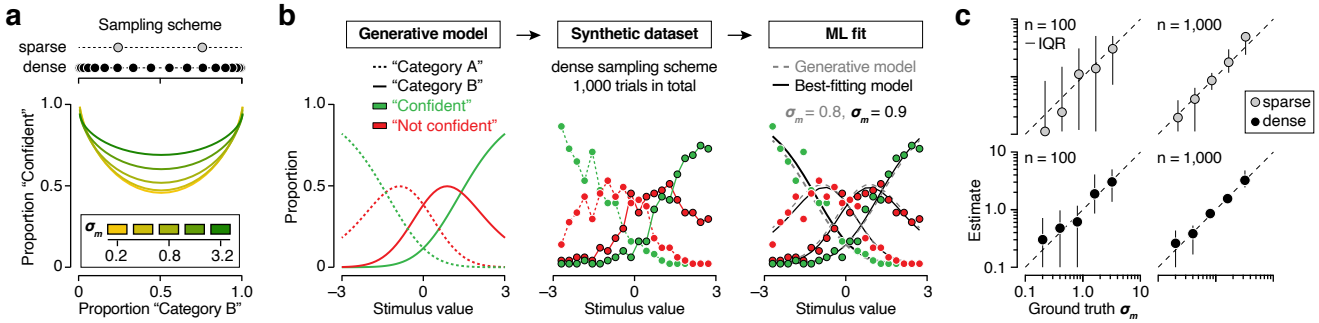


Figure 3 Model recovery analysis. (**a**) We simulated choice-confidence data for five model subjects who differed in their level of meta-uncertainty (colored lines) using experimental designs that varied in the number of trials (100 vs 1,000) and in the number of conditions (2 vs 20, grey and black symbols). (**b**) An example synthetic experiment and model-based analysis. (**c**) The median estimate of meta-uncertainty is plotted against the ground truth value for four experimental designs. Error bars illustrate the interquartile range (IQR) computed from 100 simulated data sets.

for experimental designs that involve 20 rather than 2 conditions, but this effect is more subtle (Fig. 3c, error bars in bottom vs top row). We conclude that the CASANDRE model typically can be identified in conventional experimental designs. In addition, we wish to underscore that the above procedure exploits the benefits conferred by a generative model. These benefits are not limited to studying model recovery. Our procedure can be leveraged to pursue other important goals as well (e.g., it can be used to optimize experimental design in light of a hypothesized effect size, or to conduct an absolute goodness-of-fit test to decide on the right level of model complexity, as parametric bootstrap method to derive confidence intervals on a model-based estimate, or as a tool to investigate optimal policies in incentivized confidence reporting tasks).

Meta-uncertainty: construct reliability and validity

What is a useful model? In our view, useful models afford insight into mental operations and their neural substrate beyond what can be learned from direct observations. Here, this means that meta-uncertainty ought to capture an aspect of confidence reporting behavior that is a stable property of the subject and that limits their metacognitive ability. To test these general notions, we analyzed data from several previously published studies suitably designed to illuminate the properties of meta-uncertainty as a psychological construct.

The most straightforward form of stability is repeatability. If we were to measure a subject's meta-uncertainty on two different occasions using the same experimental paradigm, we should obtain similar estimates. Navajas et al. (2017) conducted a perceptual confidence experiment in which 14 subjects performed the same task twice with approximately one month in between test sessions[25]. We used the CASANDRE model to analyze their data (see Methods and Supplementary Fig. 2). Measured and predicted choice-confidence data were strongly correlated, indicating that the model describes the data well (condition-specific proportion correct choices: Spearman's rank correlation coefficient $r = 0.96$, $P < 0.001$; condition-specific mean confidence level: $r = 0.99$, $P < 0.001$). Critically, we found meta-uncertainty estimates to be strongly correlated across both sessions as well ($r = 0.78$, $P = 0.002$; Fig. 4a). This suggests that meta-uncertainty measures a stable characteristic of human confidence reporting behavior.

Under the CASANDRE model, meta-uncertainty provides a measure of metacognitive ability, not of response bias. To investigate whether this idealized pattern holds true in human choice-confidence data, we analyzed data from 43 sessions where subjects either performed a perceptual or a cognitive confidence task. They reported their confidence in a binary decision using a six-point rating scale[25]. We artificially biased these confidence reports by mapping them onto a liberal and a conservative 4-point rating scale (see Methods)[41]. This manipulation resulted in a mean confidence level of 2.89 and 2.43 – a substantial difference in light of the standard deviation (the effect size, expressed as Cohen's d, is 3.16). We then used the model to analyze both perturbed versions of the data (see Methods). Meta-uncertainty estimates were strongly correlated ($r = 0.84$, $P < 0.001$; Fig. 4b), though note that they were on average higher for the conservatively biased version of the data (mean increase: 47%, median increase: 0%, $P = 0.002$, Wilcoxon signed rank test). This suggests that meta-uncertainty is largely, but not fully, independent of subjects' response bias.

We wondered whether meta-uncertainty depends on the absolute level of stimulus uncertainty[42]. We analyzed data from 43 sessions where subjects either performed a perceptual or cognitive confidence task. In both tasks, stimulus uncertainty was manipulated by varying the variance of the category distributions over four levels[25]. We used the CASANDRE model to

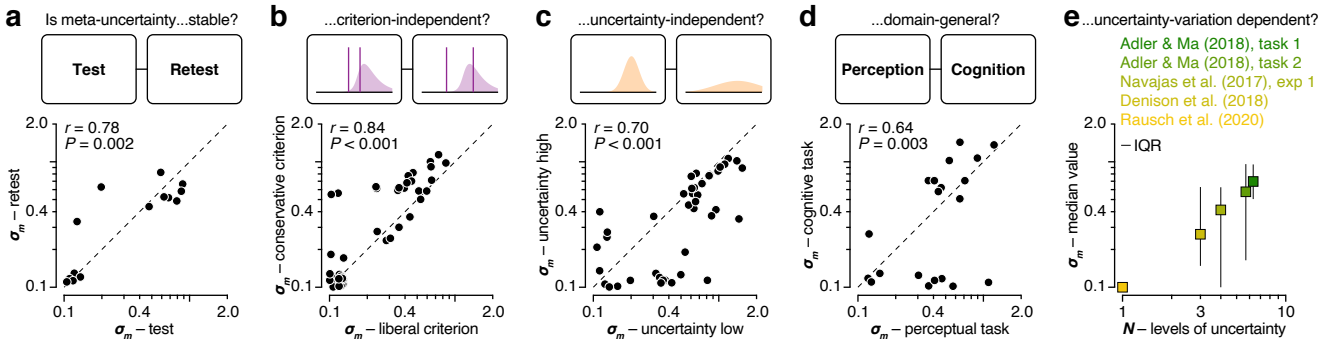


Figure 4 Evaluating meta-uncertainty as a psychological construct. (**a**) Comparison of meta-uncertainty estimates for 14 subjects who performed the same perceptual confidence task on two different occasions, separated by 1 month. We added a small amount of jitter to get a better view of overlapping data points in the lower left region of the plot. (**b**) Comparison of meta-uncertainty estimates for 43 sessions whose 6-point confidence ratings were mapped onto a liberal and conservative 4-point rating scale. (**c**) Comparison of meta-uncertainty estimates for 43 sessions where subjects performed a confidence task involving low and high levels of stimulus uncertainty. (**d**) Comparison of meta-uncertainty estimates for 20 subjects who performed a perceptual and cognitive confidence task. (**e**) Median level of meta-uncertainty plotted against number of uncertainty levels for 5 confidence experiments. Error bars illustrate the interquartile range (IQR) across subjects.

analyze these data and estimated meta-uncertainty separately for the two lowest and the two highest levels of stimulus variance (see Methods). The former conditions resulted in a much higher task performance than the latter (average proportion correct decisions: 87% vs 70%). According to the model, the corresponding underlying levels of stimulus uncertainty, σ_d, averaged 2.61 and 8.71. While increasing stimulus variance tripled stimulus uncertainty, meta-uncertainty estimates did not change much (median change: –14.76%, $P = 0.004$, Wilcoxon signed rank test). Moreover, meta-uncertainty estimates were strongly correlated across both sets of conditions ($r = 0.70$, $P < 0.001$; Fig. 4c). This suggests that meta-uncertainty is largely, but not fully, independent of the absolute level of stimulus uncertainty.

Whether metacognitive ability is domain-specific or domain-general is a debated question[8,43–45]. We analyzed data from 20 subjects who performed a perceptual and cognitive confidence task. Both tasks had the same experimental design. Stimulus categories were either defined by the average orientation of a series of rapidly presented gratings, or by the average value of a series of rapidly presented numbers[25]. Subjects' performance level was correlated across both tasks (condition-specific proportion correct choices: $r = 0.69$, $P < 0.001$), and so were their reported confidence levels, albeit to a lesser degree ($r = 0.53$, $P < 0.001$). We used the CASANDRE model to analyze both data-sets (see Methods). Meta-uncertainty estimates were strongly correlated ($r = 0.64$, $P = 0.003$; Fig. 4d). Thus, meta-uncertainty appears to capture an aspect of confidence-reporting behavior that generalizes across at least some domains.

Finally, we asked which experimental manipulations might affect meta-uncertainty. We reasoned that meta-uncertainty will be minimal in tasks that involve a single level of stimulus uncertainty. To appreciate our logic, consider the factors that make these tasks difficult. If there were no stimulus variation, the primary decision would be trivially easy. The subject could even learn to solve the task without taking the stimulus into consideration (the correct response is always the same). The primary decision only becomes difficult if the task involves multiple confusable stimulus conditions. Likewise, in the absence of uncertainty-variation, the distance between the decision variable and the decision criterion provides a direct proxy for choice consistency. There is no need to take uncertainty into account (the correct normalizing factor is always the same). Confidence judgments only become difficult if the task involves multiple confusable uncertainty levels (example shown in Fig. 1a, right). In such tasks, meaningful confidence judgments require the incorporation of trial-specific uncertainty estimates (see equation 1). We expect these estimates to become more variable as the number of uncertainty levels grows. This, in turn, will manifest as an increase in meta-uncertainty. To test this hypothesis, we used the CASANDRE model to analyze five confidence experiments that varied in the number of uncertainty levels (see Methods). These experiments utilized different stimuli and employed different experimental designs[25–28]. Yet, as expected, meta-uncertainty appears to grow lawfully with the number of uncertainty levels (Fig. 4e).

Comparison with other metrics for metacognitive ability

Our method to analyze choice-confidence data is built on the hypothesis that metacognitive ability is solely determined by meta-uncertainty. It is natural to ask how this metric of metacognitive ability relates to currently employed alternatives. We

approach this question in two ways. First, by investigating this relationship in silico whilst using the CASANDRE model as generative model of choice-confidence reports. And second, by comparing performance of these different candidate-metrics on a set of real bench-marking experiments (the tests shown in Fig. 4a-d).

One historically popular approach to quantify metacognitive ability consists of measuring the trial-by-trial correlation between choice accuracy and the confidence report (this metric is sometimes termed "phi")[12]. From the vantage point of the CASANDRE model, this approach is sensible given that meta-uncertainty shapes the association between confidence and consistency (Fig. 2e). Its simplicity is appealing. However, the other model components also affect the confidence-consistency relationship (Fig. 2g), thereby limiting phi's usefulness as metric of meta-uncertainty[12,13]. Consider an analysis of the choice-confidence reports of five model subjects who differed in their level of meta-uncertainty. We additionally varied the other model components in a step-wise fashion and computed phi for each simulated experiment. This analysis revealed a complex interdependence of the effects of the different model components on phi (Fig. 5a, top).

An alternative method to quantify metacognitive ability that has gained popularity in recent years seeks to estimate how well confidence judgements distinguish correct from incorrect decisions[13,46]. This estimate is expressed in signal-to-noise units and often termed "meta-d'". The ratio of meta-d' and stimulus discrimination ability (d') theoretically provides a measure of metacognitive efficiency and is often considered the quantity of interest[46]. Under the CASANDRE model, meta-uncertainty governs the capacity to distinguish decisions that are more likely to be correct from decisions that are less likely to be correct (Fig. 2f). However, the other model components will also affect the exact level of discrimination between correct and incorrect choices. As a result, the meta-d'/d' metric does not provide a direct measurement of meta-uncertainty, but instead reflects a complex mixture of model components (Fig. 5a, bottom).

Now consider the relationship between these metrics and meta-uncertainty for the three experiments performed by Navajas et al. (2017). Meta-uncertainty and phi are well correlated ($r = -0.60$, $P < 0.001$, Spearman correlation, Fig. 5b, top). But two other model components are also correlated with phi: stimulus uncertainty ($r = -0.59$, $P < 0.001$) and the confidence criterion ($r = -0.24$, $P = 0.028$). Likewise, meta-uncertainty and meta-d'/d' are well correlated ($r = -0.52$, $P < 0.001$). But the confidence criterion is also correlated with meta-d'/d' ($r = -0.29$, $P = 0.008$). To identify the relative importance of the different model components, we decomposed the variance of both metrics using the averaging-over-orderings technique (see Methods)[47,48]. For phi, the dominant model component is stimulus uncertainty (explained variance: 27%), followed by meta-uncertainty (22%). For meta-d'/d', the dominant model component is meta-uncertainty (24%). In other words, for both metrics, about three quarters of the variance arises from factors other than meta-uncertainty (Fig. 5c).

Our analysis suggest that phi and meta-d'/d' do not isolate the factors that limit metacognitive ability but instead measure a complex mixture of factors underlying choice-confidence data. We wondered how the performance of these mixtures in benchmarking experiments compares to that of meta-uncertainty. We computed phi and meta-d'/d' for the data sets shown in Fig. 4a-d. For each test, we first asked whether the manipulation induced a systematic change in the range of the different metrics. This was generally not the case. Change, expressed in units of standard deviation, tended to be small for all three metrics (Fig. 5d, top). The most extreme shift was observed for phi in the uncertainty-dependence test (median: –0.69, $P < 0.001$, Wilcoxon signed rank test), followed by meta-d'/d' in the test-retest experiment (median: –0.30, $P = 0.808$). We then asked for each test whether the different metrics were correlated across both test conditions. Correlations ranged from moderate to high levels, with only one test failing to reach statistical significance (domain generality of meta-d'/d': $r = 0.44$, $P = 0.056$). Overall, meta-uncertainty compared favorably to both alternative metrics. The mean correlation value across the four tests was 0.74 for meta-uncertainty, 0.67 for phi, and 0.52 for meta-d'/d' (all correlations are shown in Fig. 5d, bottom).

Intriguing stimulus, task, and state manipulations

We have argued that confidence reflects a subject's estimate of stimulus strength, normalized by their estimate of stimulus uncertainty. This notion is challenged by experimental manipulations that change confidence reporting behavior without directly altering the stimulus' signal-to-noise ratio[24,26,27,49–52]. Such effects may expose fundamental shortcomings of our hypothesis. Alternatively, these effects may result from underlying changes in the decision-making process commensurate with our proposal. The CASANDRE model provides a useful vehicle to investigate this latter option. In this final section, we analyze data from three previously published studies which kept the stimulus' signal-to-noise ratio constant, but manipulated a crucial aspect of the stimulus, the task, or the subject.

Zylberberg et al. (2016) conducted a perceptual confidence study in which three human subjects judged the direction of motion of dynamic random dot kinematograms – a stochastic motion stimulus[24]. Stimuli varied in signal-to-noise ratio ("motion strength") and in the volatility of this statistic (low vs high). This latter manipulation introduces a second layer of stochasticity. When volatility was high, subjects performed slightly worse, yet they were more confident in their decisions (Fig. 6a). Why is this so? In a sense, volatility crystallizes the concept of meta-uncertainty in stimulus-space. It does not alter the average

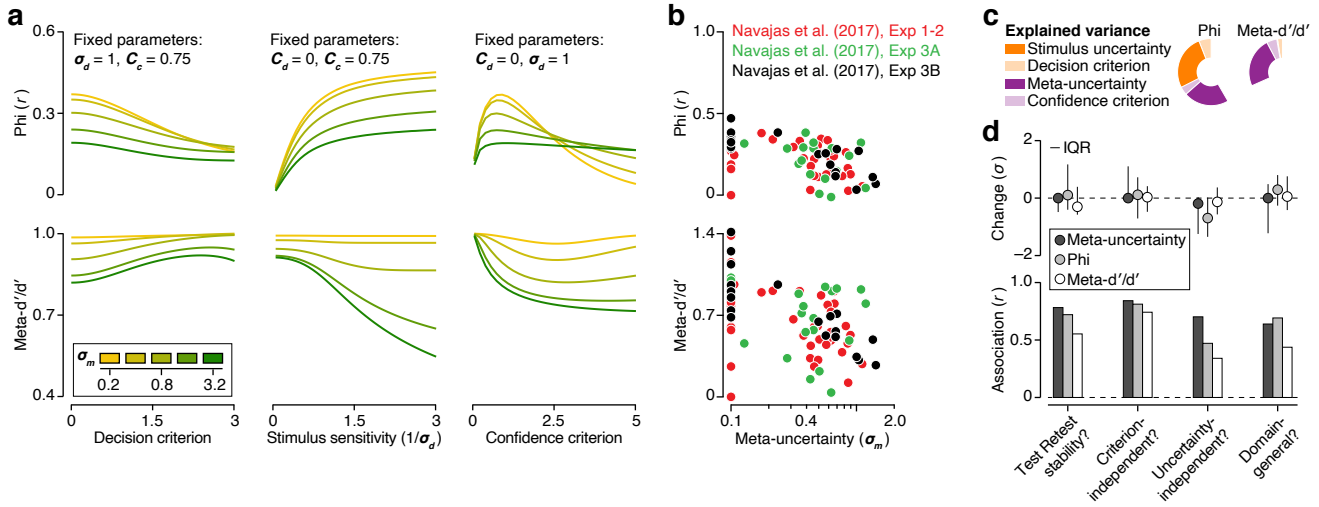


Figure 5 Comparing meta-uncertainty with two existing metrics for metacognitive ability. (**a**) We simulated choice-confidence data for a set of model observers who differed in their level of meta-uncertainty (colored lines) and additionally varied the decision criterion (left), the level of stimulus uncertainty (middle), and the confidence criterion (right). We computed phi (top) and meta-d'/d' (bottom) for each simulated experiment. (**b**) Phi (top) and meta-d'/d' (bottom) plotted against meta-uncertainty estimates for three confidence experiments. Each symbol summarizes data from a single session. (**c**) Wedges indicate the proportion of variance in phi (left) and meta-d'/d' (right) explained by each model component. (**d**) Comparison of the performance of three metrics of metacognitive ability in four bench-marking tests. Error bars illustrate the interquartile range (IQR) across subjects.

signal-to-noise ratio of the stimulus, but solely controls its variability within a trial. We therefore expect that higher levels of volatility will yield more variable estimates of stimulus uncertainty, and thus will manifest as an increase in meta-uncertainty. To test this hypothesis, we used the CASANDRE model to analyze these data and estimated parameters separately for the low and high volatility conditions (see Methods and and Supplementary Fig. 3). When stimulus volatility was high, the dynamic range of every subject's confidence-consistency relation was reduced (Fig. 6b, red vs green). This shift was well captured by the model and at least partly reflects an increase in meta-uncertainty (subject 1: +196%, $P < 0.001$; subject 2: +67.1%, $P = 0.038$; subject 3: +8.9%, $P = 0.374$; non-parametric bootstrap based test). Our analysis also revealed that volatility decreased stimulus sensitivity for two out of three subjects (subject 1: –16%, $P = 0.111$; subject 2: +1.78%, $P = 0.871$; subject 3: –25.1%, $P = 0.001$), and that it lowered confidence criteria (subject 1: –22.3%, $P = 0.049$; subject 2: –5.8%, $P = 0.68$; subject 3: –34.1%, $P < 0.001$). This final effect may be real, but it could also reflect an imperfect differentiation of the effects of meta-uncertainty and confidence criterion due to finite data (subjects completed about 2,000 trials in total). In summary, we suggest that stimulus volatility primarily changes confidence reporting behavior because it increases meta-uncertainty.

Adler and Ma (2018) conducted a perceptual confidence study in which subjects classified oriented visual stimuli in two different 2-AFC categorization tasks (task 1 and 2)[26]. In task 1, categories were characterized by two distributions of stimulus orientation that had a different mean but identical spread. In task 2, both categories had the same mean but differed in their spread. Stimuli varied in orientation and contrast. Subjects performed worse in task 2 (average proportion correct: 66.4% in task 1 vs 62.6% in task 2), yet claimed to be slightly more confident in their decisions (median change: +2.61% in task 2, $P = 0.005$, Wilcoxson signed rank test; Fig. 6c). Why might this be so? The experimental manipulation does not alter the stimulus or the subject's attentional state in a specific manner. A change in stimulus sensitivity therefore seems improbable. By the same logic, we see no reason to expect a change in meta-uncertainty. However, task 2 is designed to invite the subject to adjust their primary decision criteria in a contrast-specific manner[53]. Moreover, there was no incentive for the subjects to adopt the same set of confidence criteria in both tasks. We therefore hypothesized that changes in decision and confidence criteria accounted for the change in confidence reports. We used the CASANDRE model to analyze these data and estimated parameters separately for both tasks. Stimulus sensitivity did not differ much across both tasks (median change: +2.92% in task 2, $P = 0.011$, Wilcoxson signed rank test; $r = 0.96$; Fig. 6d, left). Meta-uncertainty was estimated to be somewhat higher in task 2 than in task 1 (median: +26.3%, $P = 0.033$; $r = 0.34$; Fig. 6d, middle). Consistent with previous analyses of these data[26], we also found that the decision criteria changed with stimulus contrast in task 2, but not in task 1 (not shown). However, the most notable effect was that subjects adopted much more liberal confidence criteria in task 2 than in task 1 (median: -48.7%, $P < 0.001$; $r = 0.80$; Fig. 6d, right). In conclusion, we suggest that this specific task structure manipulation changes confidence reporting behavior because it invites different response biases.

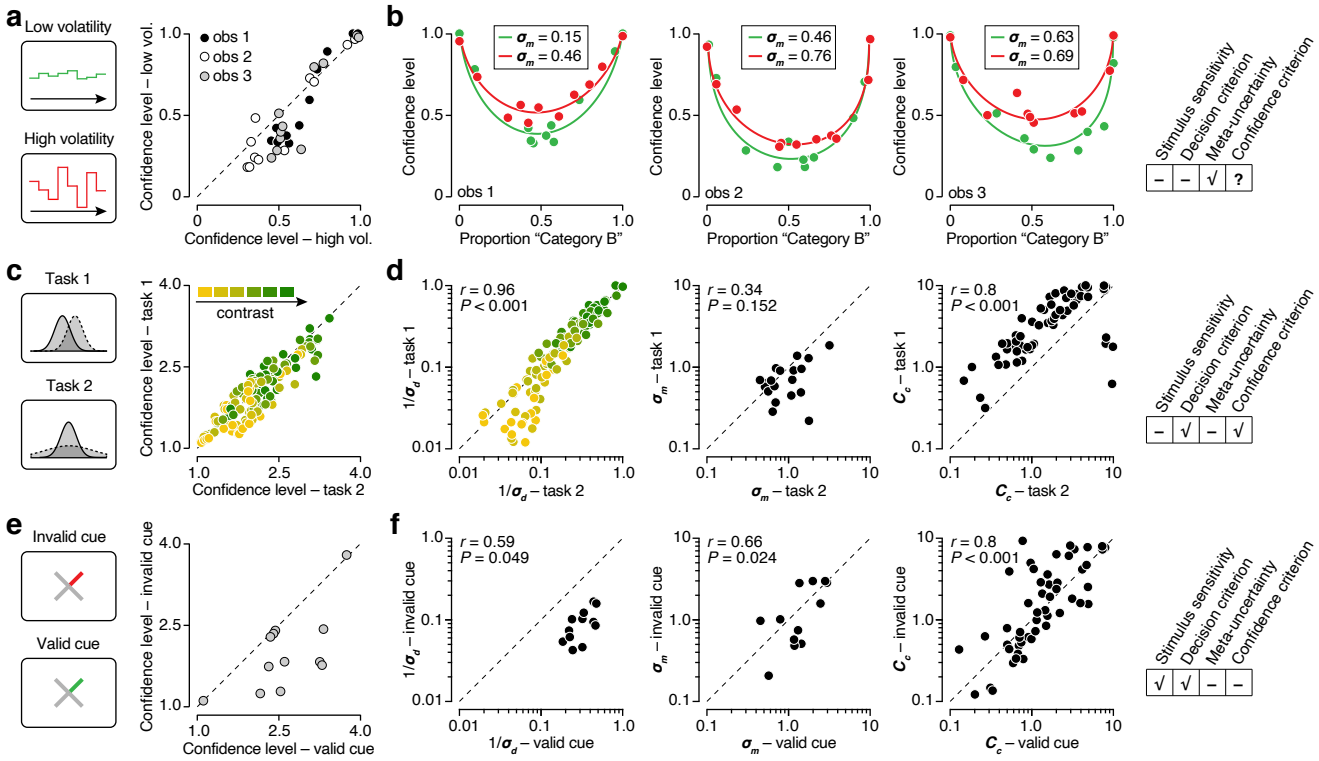


Figure 6 Model-based analysis of three confidence studies. (**a**) Zylberberg et al. (2016) found that stimulus volatility increases confidence in a perceptual decision[24]. (**b**) The confidence-consistency relationship for three subjects, measured using low vs high stimulus volatility. (**c**) Adler and Ma (2018) showed that task structure can change confidence in a perceptual decision[26]. (**d**) Cross-task comparison of stimulus sensitivity, meta-uncertainty, and confidence criteria. (**e**) Denison et al. (2018) showed that cue validity affects confidence in a perceptual decision[27]. (**f**) Comparison of stimulus sensitivity, meta-uncertainty, and confidence criteria for the valid and invalid cue condition.

Denison et al. (2018) conducted a perceptual confidence study in which subjects classified oriented visual stimuli while their spatial attention was either directed towards or away from the stimulus location ("valid cue" vs "invalid cue")[27]. Subjects performed better when their attention was directed towards the stimulus location, and their confidence reports reflected this (Fig. 6e). As such, this finding is consistent with our confidence-computation hypothesis, as attention is known to reduce perceptual uncertainty[54–56]. However, we wondered whether attention might also affect meta-uncertainty. To address this question, we used the CASANDRE model to analyze these data and estimated parameters separately for the valid and invalid cue conditions (see Methods). When the cue was valid, stimulus sensitivity was much higher (median change: +227%, $P < 0.001$; $r = 0.59$; Fig. 6f, left). However, cue validity did not alter meta-uncertainty (median change: +29.6%, $P = 0.339$; $r = 0.66$; Fig. 6f, middle), nor did it change confidence criteria (median change: +0.9%, $P = 0.629$; $r = 0.80$; Fig. 6f, right). Consistent with previous analyses of these data[27], we also found that the decision criteria changed with cue validity (not shown). Together, these results suggest that endogenous spatial attention manipulations primarily change confidence reporting behavior because they impact perceptual uncertainty[57].

Discussion

It has long been known that humans and other animals can meaningfully introspect about the quality of their decisions and actions[5–7,31,58]. Quantifying this ability has remained a significant challenge, even for simple binary decision-making tasks[12,13,15,35,41,42]. The core problem is that observable choice-confidence data reflect metacognitive ability as well as task difficulty and response bias. To overcome this problem, we introduced a metric that is anchored in an explicit hypothesis about the decision-making process that underlies behavioral reports. Our method is based on likening choice-confidence data to the outcome of an abstract mathematical process in which confidence reflects a subject's noisy estimate of their choice reliability, expressed in signal-to-noise units. This framework allowed us to specify the effects of factors that limit metacognitive ability and to summarize this loss in a single, interpretable parameter: meta-uncertainty. We showed that this process model (which we term the CASANDRE model) can explain the effects of stimulus strength and stimulus reliability on confidence reports and that meta-uncertainty can be estimated from conventional experimental designs. We found that a subject's level of meta-uncertainty is stable over time and across at least some domains. Meta-uncertainty can be manipulated experimentally: it is higher in tasks that involve more levels of stimulus reliability across trials and more volatile stimuli within trials. Meta-uncertainty appears to be mostly independent of task difficulty, task structure, response bias, and attentional state. Currently employed metrics for metacognitive ability are poor proxies for meta-uncertainty. Thus the CASANDRE model represents a notable advance toward realizing crucial medium and long-term goals in the field of metacognition[59].

The mental operations underlying confidence in a decision have long intrigued psychologists. Some authors have proposed that decision outcome and confidence both arise from a single stimulus strength estimation process[31,60–62]. Such models can explain the effects of stimulus strength, but not of stimulus reliability. Others have argued in favor of a dual process in which decision outcome and confidence are based on different stimulus strength estimates[51,63,64,64,65]. This may be the appropriate framework for cases in which subjects acquire additional task-relevant information after reporting their choice[65–67] (also see ref.[68] for a reevaluation of some evidence for dual processes). For all other cases, it appears overly complex. Instead, we have modeled confidence judgements as arising from a hierarchical process[69]. The first stage determines the choice, the second stage determines confidence. Key to our proposal is that assessing the reliability of a decision requires the use of additional information (stimulus uncertainty)[21] that in most tasks has no relevance for the choice as such. This is a plausible suggestion, as subjects can incorporate a stimulus uncertainty estimate into their decision-making when required to do so[26,53,70,71]. Additionally, there is considerable evidence that neural activity in sensory areas of the brain conveys information about stimulus features as well as the uncertainty of those features[72–78]. An important future direction will be to investigate how the precision of neural uncertainty representation informs confidence and meta-uncertainty.

The CASANDRE model provides a static description of the outcome of a hierarchical decision-making process. However, making a decision requires time. The more difficult the decision, the more time it requires[79,80]. For this reason, some authors have suggested that decision time directly informs confidence[50,66]. This proposal enjoys strong empirical support[24,38,50,81]. It is related to our proposed confidence computation, provided that the decision variable results from a mechanism that resembles bounded evidence accumulation[7]. For these mechanisms, time to reach the bound reflects the drift rate of a drift diffusion process. Drift rate is governed by stimulus strength, normalized by stimulus uncertainty and thus matches our confidence computation. Moreover, just like our second stage involves an additional factor to reflect on the quality of the decision (the uncertainty estimate), time measurements are not inherent to bounded accumulation. Finally, note that neural and behavioral time measurements are noisy[82–84]. This noisiness provides the natural dynamic analogue for meta-uncertainty in our static model. How it affects confidence reporting behavior in the diffusion-to-bound framework has not yet been studied. It remains to be seen whether choice outcome, reaction time, and metacognitive ability can all be modeled simultaneously.

Our work offers a new perspective on the mechanism underlying decision confidence. In doing so, it raises several new questions. For example, one may question the generality of the observation that stimulus strength and stimulus reliability solely impact confidence reports through their effect on choice consistency (Fig. 2b-d). This phenomenon provides the direct motivation for our proposed confidence computation, yet only a few existing studies provide a strong test for its existence. Likewise, one would be justified in questioning our knowledge of the statistical nature of meta-uncertainty. Recall that we model confidence as a noisy estimate of stimulus strength, normalized by a noisy estimate of stimulus uncertainty. It follows that the confidence variable is a ratio distribution. This class of distributions is notoriously difficult to work with. Ratio distributions rarely have closed form descriptions and are often "pathological" in the sense that their mean, variance, and other statistical moments are undefined. To circumvent this problem, we exploited the property that the family of normal distributions is closed under linear transformations. This makes it possible to describe the confidence variable distribution as a mixture of Gaussian distributions (see Methods). We used an inverted log-normal as mixing variable. However, our approach is not limited to the log-normal. Natural alternatives are the Gamma and Weibull distribution. We have not explored whether the form of the second-stage noise distribution meaningfully impacts the model behavior. This is certainly possible. If so, it would be important to know which candidate best describes behavioral confidence data.

The usefulness of process models far exceeds our current application. When coupled to an explicit goal such as maximizing choice accuracy, process models can be leveraged to derive the optimal task-strategy. The resulting predictions offer a critical point of reference for human behavior[85]. This approach has revealed that humans improve the quality of uncertain decisions by accumulating evidence over time[79], by combining information acquired through different sensory modalities[70], and by exploiting knowledge of statistical regularities in the environment[86]. Might the same hold true for uncertain confidence judgments? Stated more generally: Does our brain attempt to maximize the precision of our sense of confidence? This is a fundamental question that is ripe to be addressed. Doing so will require experiments that manipulate meta-uncertainty and incentivize the confidence reporting strategy (e.g.,[31,51,58,62,81,87–89]). The process model we have developed provides a vehicle to derive the reward-maximizing strategy and to evaluate whether human meta-uncertainty changes as expected for theoretically ideal introspection.

Methods

Modeling: Hierarchical decision-making process

We model choice-confidence data in binary decision-making tasks as arising from a hierarchical process. The first stage follows conventional signal detection theory applications[22] and describes the primary decision as resulting from the comparison of a one-dimensional decision variable, V_d, with a fixed criterion, C_d. The decision variable is subject to zero-mean Gaussian noise and hence follows a normal distribution with mean μ_d and standard deviation σ_d. The decision variable is converted into a signed confidence variable, V_c', by taking the difference of V_d and C_d, and dividing this difference by $\hat{\sigma_d}$, the subject's estimate of σ_d. The family of normal distributions is closed under linear transformations. This means that, if $\hat{\sigma_d}$ were a constant, V_c' would also follow a normal distribution with mean $\mu_c' = (\mu_d - C_d)/\hat{\sigma_d}$ and standard deviation $\sigma_c' = \sigma_d/\hat{\sigma_d}$. The confidence report results from the comparison of the confidence variable with a single fixed criterion, C_c (or with a set of criteria if the confidence scale has more than two levels). It follows that the probability of a "confident" judgement given a "Category A" decision is given by $P(C = 1|D = 0) = \Phi(-C_c)$, where $\Phi(.)$ is the cumulative normal distribution with mean μ_c' and standard deviation σ_c'. By the same logic, $P(C = 0|D = 0) = \Phi(0) - \Phi(-C_c)$, $P(C = 0|D = 1) = \Phi(C_c) - \Phi(0)$, and $P(C = 1|D = 1) = 1 - \Phi(C_c)$. Key to the CASANDRE model is that $\hat{\sigma_d}$ is not a constant, but a random variable that follows a log-normal distribution with mean σ_d and standard deviation σ_m. Consequently, the signed confidence variable is a mixture of normal distributions, with mixing weights determined by σ_m. To obtain the probability of each response option under this mixture, we sample $\hat{\sigma_d}$ in steps of constant cumulative density (using the Matlab function 'logninv'), compute the probability of each response option under each sample's resulting normal distribution (using the Matlab function 'normcdf'), and average these probabilities across all samples. We found that this procedure yields stable probability estimates once the number of samples exceeds 25 (i.e., sampling the log-normal distribution in steps no greater than 4%). For all applications in this paper, we used 100 samples, thus sampling $\hat{\sigma_d}$ at a cumulative density of 0.5%, 1.5%, 2.5%,..., and 99.5%.

Modeling: Parameterization, simulations, and fitting

We analyzed data from a large set of previously published studies that employed different task designs. The simplest designs involve the combination of a 2-AFC categorization decision and a binary confidence report (i.e. the model simulations shown in Fig. 2 and 3). Under the CASANDRE model, the predicted probability of each response option is fully specified by five parameters: the mean of the decision variable (μ_d), the standard deviation of the decision variable (σ_d), the decision criterion (C_d), the level of meta-uncertainty (σ_m), and the confidence criterion (C_c). It is not possible to estimate each of these parameters

for every unique experimental condition. To make the model identifiable, we generally assume that μ_d is identical to the true stimulus value, that σ_d is constant for a given level of stimulus reliability, and that C_d, σ_m and C_c are constant across multiple conditions. Figure 2 shows how each of the parameters affects the model's behavior. Finally, when fitting data, we use one additional parameter, λ, to account for stimulus-independent lapses[90], which we assume to be uniformly distributed across all response options. We fit the model on a subject-by-subject basis. For each subject, we compute the log-likelihood of a given set of model parameters across all choice-confidence reports and use an iterative procedure to identify the most likely set of parameter values (specifically, the interior point algorithm used by the Matlab function 'fmincon'). Figure 3b shows an example model fit to a synthetic data set whereby we used 5 free parameters (λ, σ_d, C_d, σ_m, and C_c) to capture data across 20 experimental conditions.

Some studies used a task design that combined a 2-AFC categorization decision with a multi-level confidence rating scale (i.e., ref.[25,26,28]). To model these data, we used the same approach as described above but we used multiple confidence criteria (one less than the number of confidence levels). We modeled the data from ref.[28] using seven free parameters: λ, σ_d, C_d, σ_m, and C_c (4-point confidence rating scale, thus three in total) (see Fig. 4e). We modeled some data from ref.[26] (task 1) using seventeen free parameters: λ, σ_d (one per contrast level, six in total), C_d (one per contrast level, six in total), σ_m, and C_c (4-point confidence rating scale, thus three in total). Example fits are shown in Fig. 1b,c and in Supplementary Fig. 1 (also see Fig. 4e, task 1). We modeled the data from ref.[25] using twelve free parameters: λ, σ_d (one per stimulus variance level, four in total), C_d, σ_m, and C_c (6-point confidence rating scale, thus five in total). Example fits are shown in Supplementary Fig. 2 (also see Fig. 4 and Fig. 5b-d).

Some studies used a task design in which the 2-AFC categorization decision pertained to two category distributions with the same mean but different spread (i.e., ref.[26,27]). To model these data, we assumed that the primary decision results from a comparison of the decision variable with two decision criteria, and that the confidence estimate is based on the distance between the decision variable and the nearest decision criterion. We modeled some data from ref.[26] (task 2) using twenty-three free parameters: λ, σ_d (one per contrast level, six in total), C_d (two per contrast level, twelve in total), σ_m, and C_c (4-point confidence rating scale, thus three in total). See Fig. 4e, task 2. Example fits are shown in Supplementary Fig. 4. We modeled data from ref.[27] using sixteen free parameters: λ, σ_d (one per attention level, three in total), C_d (two per attention level, six in total), σ_m (one per attention level, three in total), and C_c (4-point confidence rating scale, thus three in total). See Fig. 4e.

For some data sets, it was evident that subjects' confidence criteria depended on the primary decision. Such asymmetry can for example arise when the motor response used to communicate the confidence report differs depending on the primary decision (e.g., ref.[24,26,27]). To model such data, we used two sets of confidence criteria: one for each decision-outcome. For the sake of consistency, we used this approach for all analyses documented in Fig. 6. Visually, asymmetric confidence criteria manifest as an asymmetric confidence-consistency relation (e.g., Fig. 6a, obs 3; also see example fits shown in Supplementary Fig. 3 and 4).

Some studies used a task design that combined a 2-IFC categorization decision with a confidence report (i.e., Arbuzova and Filevich, unpublished but available in the Confidence Database[23]). In these tasks, a subject is shown two stimulus intervals and judges which interval contained the "signal" stimulus. To model such data, we assume that the decision is based on a comparison of the evidence provided by each stimulus interval. The one-dimensional decision variable, V_d, reflects the outcome of this comparison, which we model as a difference operation[22]. The difference of two Gaussian distributions is itself a Gaussian with mean equal to the difference of the means and standard deviation equal to the square root of the sum of the variances. Everything else is the same as for the 2-AFC task. When different from zero, C_d now reflects an interval bias (e.g., a preference for "interval 1" choices). See example fit in Fig. 1d.

Datasets

We obtained most experimental data from the Confidence Database[23] (available at: `https://osf.io/s46pr/`). Additional experimental data from Navajas et al. (2017) were provided by an author[25]. Given that the CASANDRE model yields more reliable parameter estimates for longer experiments with more stimulus conditions (Fig. 3), we restricted our analyses to experiments which involved several hundred trials per subject and multiple levels of stimulus strength and/or stimulus reliability. The majority of subjects we analyzed came from two studies[25,26], but we applied the CASANDRE model to data from four additional studies with fewer subjects to demonstrate the breadth of possible applications[23,24,27,28]. All detailed experimental designs and procedures are available in the original publications or in abbreviated form in the Confidence Database. We briefly describe each data set below.

We analyzed data from all 3 experiments in ref.[26]. All subjects in experiments 1 and 2 performed both task 1 (discriminating categories of orientation distributions with different means but the same standard deviation; their "Task A") and task 2 (discriminating categories of orientation distributions with the same mean but different standard deviations; their "Task B"). Since

stimulus orientations were drawn from a continuous distribution, to plot the data we grouped nearby orientations into 9 bins with similar numbers of trials. Data and model fits from two example subjects performing task 1 in experiment 1 are shown in Fig. 1b-c and Supplementary Fig. 1. Fitted parameters from all 19 subjects who performed experiments 1 and 2 are included in Fig. 4e (task 1) and Fig. 6d. Subjects in experiment 3 performed only task 2. Data and model fits from an example subject performing task 2 in experiment 3 are shown in Supplementary Fig. 4. Fitted parameters from all 34 subjects who performed task 2 in experiments 1, 2, and 3 are included in Fig. 4e (task 2).

We analyzed data from all 3 experiments in ref.[25]. 30 subjects performed experiment 1. 14 of those 30 subjects returned about a month after their first session to perform the same task again as experiment 2. Finally, 20 subjects performed experiment 3, participating in a perceptual (experiment 3A) and cognitive (experiment 3B) task in two different sessions. We analyzed each of these 84 different experimental sessions independently. Data and model fits from an example subject are shown in Supplementary Fig. 2. Fitted parameters and alternative metacognitive metrics from all 14 subjects who performed both experiments 1 and 2 are included in Fig. 4a and Fig. 5d (Test-retest stability). Fitted parameters and alternative metacognitive metrics from all 20 subjects who performed experiment 3 are included in Fig. 4d and Fig. 5d (Domain generality). Fitted parameters from 30 subjects performing experiment 1 are included in Fig. 4e, and fitted parameters from all 84 sessions are included in Fig. 5b. Further analyses using these data to test the independence of meta-uncertainty from confidence bias and uncertainty are explained in the next section.

We analyzed unpublished data from Arbuzova and Filevich (available in the Confidence Database under the name Arbuzova_unpub_1)[23]. This experiment demonstrates the generalization of the CASANDRE model to a visuomotor estimation task as well as 2-IFC experimental designs. Data and model fits from a representative subject are shown in Fig. 1d.

Fitted parameters from all 25 subjects from ref.[28] are included in Fig. 4e. We analyzed data from 12 subjects performing a version of task 2 in ref.[26] with an added attention manipulation from ref.[27]. To get the single estimate of meta-uncertainty included in Fig. 4e for each subject, we averaged the values estimated from all three attention conditions, as these were not significantly different (Fig. 6f). To obtain the parameter estimates shown in Fig. 6f, the model was fit independently to the valid and invalid cue conditions, with only the lapse rate shared across conditions.

We analyzed data from 3 subjects from ref.[24]. Here, subjects reported their confidence on a continuous scale through the end point of a saccadic eye movement directed toward an elongated target (encoded as a value between 0 and 1). We transformed these data to a binary confidence report by categorizing all saccade end points above the mid point of the target (0.5) as high confidence, and all saccade end points below the mid point as low confidence. Confidence data and model fits from all 3 subjects are shown in Fig. 6b, and data and model fits for an example subject are shown in Supplementary Fig. 3.

Construct validity analyses

To test the independence between confidence bias and measures of metacognitive ability, we manipulated the confidence reporting behavior of subjects across all sessions from ref.[25] (following an analysis developed by ref.[41]). In these experiments, confidence reports were measured using a six-point rating scale. We remapped responses into a four point rating scale using two different grouping rules (one conservative, one liberal). The conservative mapping is [1|2 3 4|5|6], the liberal mapping is [1|2|3 4 5|6]. To limit the model comparison to the second stage of the decision making process, the lapse rate, stimulus sensitivity, and perceptual criterion were shared across both model variants. Only the meta-uncertainty and confidence criteria differed across both model variants. To obtain adequately constrained and stable model fits to these manipulated data, we only included a session in the analysis if at least 10 responses were recorded at the highest level of the confidence scale. This reduced a total of 84 sessions to 43, shown in Fig. 4b.

To test the independence between stimulus uncertainty and measures of metacognitive ability, we split experimental data from each session in half[25]. We estimated meta-uncertainty independently for the two easiest and the two hardest stimulus conditions. To limit the model comparison to the question of whether meta-uncertainty is independent of stimulus reliability, all other model parameters were fixed across conditions. For consistency with the criterion analysis, we applied the same inclusion criteria, yielding data from 43 sessions included in Fig. 4c.

Calculating alternative metacognitive metrics

We probed the relation between meta-uncertainty and two alternative metrics of meta-cognitive ability under the CASANDRE model. We used two distinct procedures for this. First, to obtain estimates of "meta-d'", we used the CASANDRE model to specify the probability of each response option in a 2-AFC discrimination task with binary confidence report options for an experiment that included two stimulus conditions. We calculated meta-d' following[46]. Briefly, we searched for the level of sensory noise and the confidence criterion that best explained the distribution of confidence reports conditioned on the primary

choice, assuming a normally distributed confidence variable. The ratio of the ground truth sensory noise level and this estimate is plotted in the bottom panels of Fig. 5a. Second, to obtain the expected value of phi, we simulated 200,000 trials in an experiment that included 20 levels of stimulus strength. We then calculated the Pearson correlation between the resulting choice accuracy and confidence vectors (Fig 5a, top panels).

We also computed these alternative metrics for each session from ref.[25] (see Fig. 5b-d). As is conventional, we estimated d' for each stimulus condition from the observed hit and false alarm rates[22]. To obtain estimates of "meta-d'", we searched for the decision criterion, the set of confidence criteria, and the level of sensory noise that best explained the choice-conditioned data, assuming a normally distributed confidence variable. To obtain a single meta-d'/d' estimate per session, we computed the arithmetic mean across the four stimulus conditions. We computed phi for each session by calculating the Pearson correlation between choice accuracy and raw confidence report (see Fig. 5b-d)

To compute the proportion of variance in each alternative metric across 84 sessions[25] explained by different components of the CASANDRE model, we used the averaging-over-orderings technique[47,48]. We used multiple linear regression to obtain the variance in a metric explained by the CASANDRE model. Then, for each model parameter we compute the difference in explained variance when the parameter is included and when it is not. The resulting estimates of explained variance for each parameter are plotted in Fig. 5c.

References

1. Florent Meyniel, Mariano Sigman, and Zachary F. Mainen. Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron*, 88(1):78–92, October 2015.
2. Jan Drugowitsch, André G. Mendonça, Zachary F. Mainen, and Alexandre Pouget. Learning optimal decisions with confidence. *Proceedings of the National Academy of Sciences*, 116(49):24872–24880, December 2019.
3. Braden A. Purcell and Roozbeh Kiani. Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proceedings of the National Academy of Sciences*, 113(31):E4531–E4540, August 2016.
4. Bahador Bahrami, Karsten Olsen, Peter E. Latham, Andreas Roepstorff, Geraint Rees, and Chris D. Frith. Optimally Interacting Minds. *Science*, 329(5995):1081–1085, August 2010.
5. Charles Sanders Peirce and Joseph Jastrow. On small differences in sensation. *Memoirs of the National Academy of Sciences*, 3, 1884.
6. Roger Ratcliff. A theory of memory retrieval. *Psychological Review*, 85(2):59–108, 1978.
7. Douglas Vickers. *Decision processes in visual perception*. Academic Press, New York, 1979.
8. Vincent de Gardelle, François Le Corre, and Pascal Mamassian. Confidence as a Common Currency between Vision and Audition. *PLOS ONE*, 11(1):e0147901, January 2016.
9. Stephen M. Fleming, Rimona S. Weil, Zoltan Nagy, Raymond J. Dolan, and Geraint Rees. Relating Introspective Accuracy to Individual Differences in Brain Structure. *Science*, 329(5998):1541–1543, September 2010.
10. Marion Rouault, Tricia Seow, Claire M. Gillan, and Stephen M. Fleming. Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry*, 84(6):443–451, September 2018.
11. Deanna Kuhn. Theory of mind, metacognition, and reasoning: A life-span perspective. In *Children's reasoning and the mind*, pages 301–326. Psychology Press, 2000.
12. Thomas O. Nelson. A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 95(1):109–133, 1984.
13. Stephen M. Fleming and Hakwan C. Lau. How to measure metacognition. *Frontiers in Human Neuroscience*, 8:443, 2014.
14. Pascal Mamassian. Visual Confidence. *Annual Review of Vision Science*, 2(1):459–481, 2016.
15. Matthias Guggenmos. Validity and reliability of metacognitive performance measures. Technical report, PsyArXiv, March 2021. type: article.
16. L. Festinger. Studies in decision: I. Decision-time, relative frequency of judgment and subjective confidence as related to physical stimulus difference. *Journal of Experimental Psychology*, 32(4):291–306, 1943.
17. Jinoos Hosseini and William R. Ferrell. Detectability of correctness: A measure of knowing that one knows. *Instructional Science*, 11(2):113–127, August 1982.
18. Thomas S. Critchfield. Signal-Detection Properties of Verbal Self-Reports. *Journal of the Experimental Analysis of Behavior*, 60(3):495–514, 1993.
19. Susan J. Galvin, John V. Podd, Vit Drga, and John Whitmore. Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, 10(4):843–876, December 2003.
20. Baptiste Caziot and Pascal Mamassian. Perceptual confidence judgments reflect self-consistency. *Journal of Vision*, 21(12):8, November 2021.
21. A. Pouget, J. Drugowitsch, and A. Kepecs. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience*, 19(3):366–374, 2016.
22. David Marvin Green and John A. Swets. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966.
23. Dobromir Rahnev, Kobe Desender, Alan L. F. Lee, William T. Adler, David Aguilar-Lleyda, Başak Akdoğan, Polina Arbuzova,

Lauren Y. Atlas, Fuat Balcı, Ji Won Bang, Indrit Bègue, Damian P. Birney, Timothy F. Brady, Joshua Calder-Travis, Andrey Chetverikov, Torin K. Clark, Karen Davranche, Rachel N. Denison, Troy C. Dildine, Kit S. Double, Yalçın A. Duyan, Nathan Faivre, Kaitlyn Fallow, Elisa Filevich, Thibault Gajdos, Regan M. Gallagher, Vincent de Gardelle, Sabina Gherman, Nadia Haddara, Marine Hainguerlot, Tzu-Yu Hsu, Xiao Hu, Iñaki Iturrate, Matt Jaquiery, Justin Kantner, Marcin Koculak, Mahiko Konishi, Christina Koß, Peter D. Kvam, Sze Chai Kwok, Maël Lebreton, Karolina M. Lempert, Chien Ming Lo, Liang Luo, Brian Maniscalco, Antonio Martin, Sébastien Massoni, Julian Matthews, Audrey Mazancieux, Daniel M. Merfeld, Denis O'Hora, Eleanor R. Palser, Borysław Paulewicz, Michael Pereira, Caroline Peters, Marios G. Philiastides, Gerit Pfuhl, Fernanda Prieto, Manuel Rausch, Samuel Recht, Gabriel Reyes, Marion Rouault, Jérôme Sackur, Saeedeh Sadeghi, Jason Samaha, Tricia X. F. Seow, Medha Shekhar, Maxine T. Sherman, Marta Siedlecka, Zuzanna Skóra, Chen Song, David Soto, Sai Sun, Jeroen J. A. van Boxtel, Shuo Wang, Christoph T. Weidemann, Gabriel Weindel, Michał Wierzchoń, Xinming Xu, Qun Ye, Jiwon Yeon, Futing Zou, and Ariel Zylberberg. The Confidence Database. *Nature Human Behaviour*, 4(3):317–325, March 2020.

24. Ariel Zylberberg, Christopher R Fetsch, and Michael N Shadlen. The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *eLife*, 5:e17688, October 2016.
25. Joaquin Navajas, Chandni Hindocha, Hebah Foda, Mehdi Keramati, Peter E. Latham, and Bahador Bahrami. The idiosyncratic nature of confidence. *Nature Human Behaviour*, 1(11):810–818, November 2017.
26. William T. Adler and Wei Ji Ma. Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLOS Computational Biology*, 14(11):e1006572, November 2018.
27. Rachel N. Denison, William T. Adler, Marisa Carrasco, and Wei Ji Ma. Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences*, 115(43):11090–11095, October 2018.
28. Manuel Rausch, Michael Zehetleitner, Marco Steinhauser, and Martin E. Maier. Cognitive modelling reveals distinct electrophysiological markers of decision confidence and error monitoring. *NeuroImage*, 218:116963, September 2020.
29. J. D. Balakrishnan and Roger Ratcliff. Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology: Human Perception and Performance*, 22(3):615–633, 1996.
30. William R. Ferrell. A model for realism of confidence judgments: Implications for underconfidence in sensory discrimination. *Perception & Psychophysics*, 57(2):246–254, January 1995.
31. Adam Kepecs, Naoshige Uchida, Hatim A. Zariwala, and Zachary F. Mainen. Neural correlates, computation and behavioural impact of decision confidence. *Nature*, 455(7210):227–231, September 2008.
32. Michel Treisman and Andrew Faulkner. The setting and maintenance of criteria representing levels of confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 10(1):119–139, 1984.
33. Thomas S. Wallsten and Claudia González-Vallejo. Statement verification: A stochastic model of judgment and response. *Psychological Review*, 101(3):490–504, 1994.
34. E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review*, 106(4):620–630, May 1957.
35. Medha Shekhar and Dobromir Rahnev. The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review*, 128(1):45–70, 2021.
36. Shannon M. Locke, Elon Gaffin-Cahn, Nadia Hosseinizaveh, Pascal Mamassian, and Michael S. Landy. Priors and payoffs in confidence judgments. *Attention, Perception, & Psychophysics*, 82(6):3158–3175, August 2020.
37. Andra Mihali, Marianne Broeker, and Guillermo Horga. Insightful inference compensates for distorted perception. *bioRxiv*, page 2021.11.13.468497, November 2021. Type: article.
38. Christopher R. Fetsch, Roozbeh Kiani, William T. Newsome, and Michael N. Shadlen. Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron*, 83(4):797–804, 2014.
39. Christopher R Fetsch, Naomi N Odean, Danique Jeurissen, Yasmine El-Shamayleh, Gregory D Horwitz, and Michael N Shadlen. Focal optogenetic suppression in macaque area MT biases direction discrimination and decision confidence, but only transiently. *eLife*, 7:e36523, July 2018.
40. George EP Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.
41. Kai Xue, Medha Shekhar, and Dobromir Rahnev. Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. *Consciousness and Cognition*, 95:103196, October 2021.
42. Ji Won Bang, Medha Shekhar, and Dobromir Rahnev. Sensory noise increases metacognitive efficiency. *Journal of Experimental Psychology. General*, 148(3):437–452, March 2019.
43. Li Yan McCurdy, Brian Maniscalco, Janet Metcalfe, Ka Yuet Liu, Floris P. de Lange, and Hakwan Lau. Anatomical Coupling between Distinct Metacognitive Systems for Memory and Visual Perception. *Journal of Neuroscience*, 33(5):1897–1906, January 2013.
44. Benjamin Baird, Matthew Cieslak, Jonathan Smallwood, Scott T. Grafton, and Jonathan W. Schooler. Regional White Matter Variation Associated with Domain-specific Metacognitive Accuracy. *Journal of Cognitive Neuroscience*, 27(3):440–452, March 2015.
45. Alan L. F. Lee, Eugene Ruby, Nathan Giles, and Hakwan Lau. Cross-Domain Association in Metacognitive Efficiency Depends on First-Order Task Types. *Frontiers in Psychology*, 9, 2018.
46. Brian Maniscalco and Hakwan Lau. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1):422–430, March 2012.

47. William Kruskal. Relative Importance by Averaging over Orderings. *The American Statistician*, 41(1):6–10, February 1987.
48. Ulrike Grömping. Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. *The American Statistician*, 61(2):139–147, May 2007.
49. Dobromir Rahnev, Brian Maniscalco, B. Luber, Hakwan Lau, and Sarah H. Lisanby. Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *Journal of neurophysiology*, 107(6):1556–1563, 2012.
50. Roozbeh Kiani, Leah Corthell, and Michael N. Shadlen. Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron*, 84(6):1329–1342, December 2014.
51. Brian Maniscalco, Megan A. K. Peters, and Hakwan Lau. Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception, & Psychophysics*, 78(3):923–937, April 2016.
52. Jason Samaha and Rachel Denison. The positive evidence bias in perceptual confidence is not post-decisional. *bioRxiv*, page 2020.03.15.991513, March 2020. Type: article.
53. Ahmad T. Qamar, R. James Cotton, Ryan G. George, Jeffrey M. Beck, Eugenia Prezhdo, Allison Laudano, Andreas S. Tolias, and Wei Ji Ma. Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proceedings of the National Academy of Sciences*, 110(50):20332–20337, December 2013.
54. Zhong-Lin Lu and Barbara Anne Dosher. External noise distinguishes attention mechanisms. *Vision Research*, 38(9):1183–1198, May 1998.
55. Marisa Carrasco, Cigdem Penpeci-Talgar, and Miguel Eckstein. Spatial covert attention increases contrast sensitivity across the CSF: support for signal enhancement. *Vision Research*, 40(10):1203–1215, June 2000.
56. Marisa Carrasco. Visual attention: The past 25 years. *Vision Research*, 51(13):1484–1525, July 2011.
57. Dobromir Rahnev, Brian Maniscalco, Tashina Graves, Elliott Huang, Floris P. de Lange, and Hakwan Lau. Attention induces conservative subjective biases in visual perception. *Nature Neuroscience*, 14(12):1513–1515, December 2011.
58. Wendy E. Shields, J. David Smith, Katarina Guttmannova, and David A. Washburn. Confidence Judgments by Humans and Rhesus Monkeys. *The Journal of general psychology*, 132(2):165–186, April 2005.
59. Dobromir Rahnev, Tarryn Balsdon, Lucie Charles, Vincent de Gardelle, Rachel N. Denison, Kobe Desender, Nathan Faivre, Elisa Filevich, Stephen Fleming, Janneke Jehee, Hakwan Lau, Alan L. F. Lee, Shannon M. Locke, Pascal Mamassian, Brian Odegaard, Megan A. K. Peters, Gabriel Reyes, Marion Rouault, Jérôme Sackur, Jason Samaha, Claire Sergent, Maxine Sherman, Marta Siedlecka, David Soto, Alexandra Vlassova, and Ariel Zylberberg. Consensus goals for the field of visual metacognition. Technical report, PsyArXiv, April 2021. type: article.
60. Yoshiaki Ko and Hakwan Lau. A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1401–1411, May 2012.
61. Yutaka Komura, Akihiko Nikkuni, Noriko Hirashima, Teppei Uetake, and Aki Miyamoto. Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nature Neuroscience*, 16(6):749–755, June 2013.
62. Sébastien Massoni, Thibault Gajdos, and Jean-Christophe Vergnaud. Confidence measurement in the light of signal detection theory. *Frontiers in Psychology*, 5, 2014.
63. Ariel Zylberberg, Pablo Barttfeld, and Mariano Sigman. The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, 6, 2012.
64. Megan A. K. Peters, Thomas Thesen, Yoshiaki D. Ko, Brian Maniscalco, Chad Carlson, Matt Davidson, Werner Doyle, Ruben Kuzniecky, Orrin Devinsky, Eric Halgren, and Hakwan Lau. Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, 1(7):1–8, July 2017.
65. Stephen M. Fleming and Nathaniel D. Daw. Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, 124(1):91–114, 2017.
66. Christopher R. Fetsch, Roozbeh Kiani, and Michael N. Shadlen. Predicting the accuracy of a decision: a neural mechanism of confidence. In *Cold Spring Harbor symposia on quantitative biology*, volume 79, pages 185–197. Cold Spring Harbor Laboratory Press, 2014.
67. Peter R Murphy, Ian H Robertson, Siobhán Harty, and Redmond G O'Connell. Neural evidence accumulation persists after choice to inform metacognitive judgments. *eLife*, 4:e11946, December 2015.
68. Koosha Khalvati, Roozbeh Kiani, and Rajesh P. N. Rao. Bayesian inference with incomplete knowledge explains perceptual confidence and its deviations from accuracy. *Nature Communications*, 12(1):5704, September 2021.
69. Brian Maniscalco and Hakwan Lau. The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness*, 2016(1), January 2016.
70. Marc O. Ernst and Martin S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429, 2002.
71. Christopher R. Fetsch, Alexandre Pouget, Gregory C. DeAngelis, and Dora E. Angelaki. Neural correlates of reliability-based cue weighting during multisensory integration. *Nature neuroscience*, 15(1):146, 2012.
72. Wei Ji Ma, Jeffrey M. Beck, Peter E. Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432, 2006.
73. Gergő Orbán, Pietro Berkes, József Fiser, and Máté Lengyel. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92(2):530–543, 2016.
74. Ruben S. van Bergen, Wei Ji Ma, Michael S. Pratte, and Janneke F. M. Jehee. Sensory uncertainty decoded from visual

cortex predicts behavior. *Nature Neuroscience*, 18(12):1728–1730, December 2015.
75. Ruben S. van Bergen and Janneke F. M. Jehee. Probabilistic Representation in Human Visual Cortex Reflects Uncertainty in Serial Decisions. *Journal of Neuroscience*, 39(41):8164–8176, October 2019.
76. Olivier J. Hénaff, Zoe M. Boundy-Singer, Kristof Meding, Corey M. Ziemba, and Robbe L. T. Goris. Representation of visual uncertainty through neural gain variability. *Nature Communications*, 11(1):2513, May 2020.
77. Edgar Y. Walker, R. James Cotton, Wei Ji Ma, and Andreas S. Tolias. A neural basis of probabilistic computation in visual cortex. *Nature Neuroscience*, 23(1):122–129, January 2020.
78. Dylan Festa, Amir Aschner, Aida Davila, Adam Kohn, and Ruben Coen-Cagli. Neuronal variability reflects probabilistic inference tuned to natural image statistics. *Nature Communications*, 12(1):3635, June 2021.
79. John Palmer, Alexander C. Huk, and Michael N. Shadlen. The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, 5(5):1, May 2005.
80. Timothy D. Hanks, Mark E. Mazurek, Roozbeh Kiani, Elisabeth Hopp, and Michael N. Shadlen. Elapsed Decision Time Affects the Weighting of Prior Probability in a Perceptual Decision Task. *Journal of Neuroscience*, 31(17):6339–6352, April 2011.
81. Roozbeh Kiani and Michael N. Shadlen. Representation of confidence associated with a decision by neurons in the parietal cortex. *science*, 324(5928):759–764, 2009.
82. John Gibbon. Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, 84(3):279–325, 1977.
83. Mehrdad Jazayeri and Michael N. Shadlen. Temporal context calibrates interval timing. *Nature Neuroscience*, 13(8):1020–1026, August 2010.
84. Mehrdad Jazayeri and Michael N. Shadlen. A Neural Mechanism for Sensing and Reproducing a Time Interval. *Current Biology*, 25(20):2599–2609, October 2015.
85. Wilson S. Geisler. Ideal observer analysis. In *The Visual Neurosciences*, volume 10, pages 825–837. MIT Press, Boston, 2003.
86. Yair Weiss, Eero P. Simoncelli, and Edward H. Adelson. Motion illusions as optimal percepts. *Nature neuroscience*, 5(6):598, 2002.
87. Navindra Persaud, Peter McLeod, and Alan Cowey. Post-decision wagering objectively measures awareness. *Nature Neuroscience*, 10(2):257–261, February 2007.
88. Zoltán Dienes and Anil Seth. Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition*, 19(2):674–681, June 2010.
89. Zahra Murad, Martin Sefton, and Chris Starmer. How do risk attitudes affect measured confidence? *Journal of Risk and Uncertainty*, 52(1):21–46, February 2016.
90. F. A. Wichmann and N. J. Hill. The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, 63(8):1293–1313, November 2001.

Acknowledgements

We thank the creators and contributors to the Confidence Database and Joaquin Navajas for making their data available. This work was supported by a Whitehall Foundation grant (to RLTG), U.S. National Institutes of Health grants T32 EY021462 (supporting CMZ) and K99 EY032102 (to CMZ), and a U.S. National Science Foundation Graduate Research Fellowship (to ZMB-S).

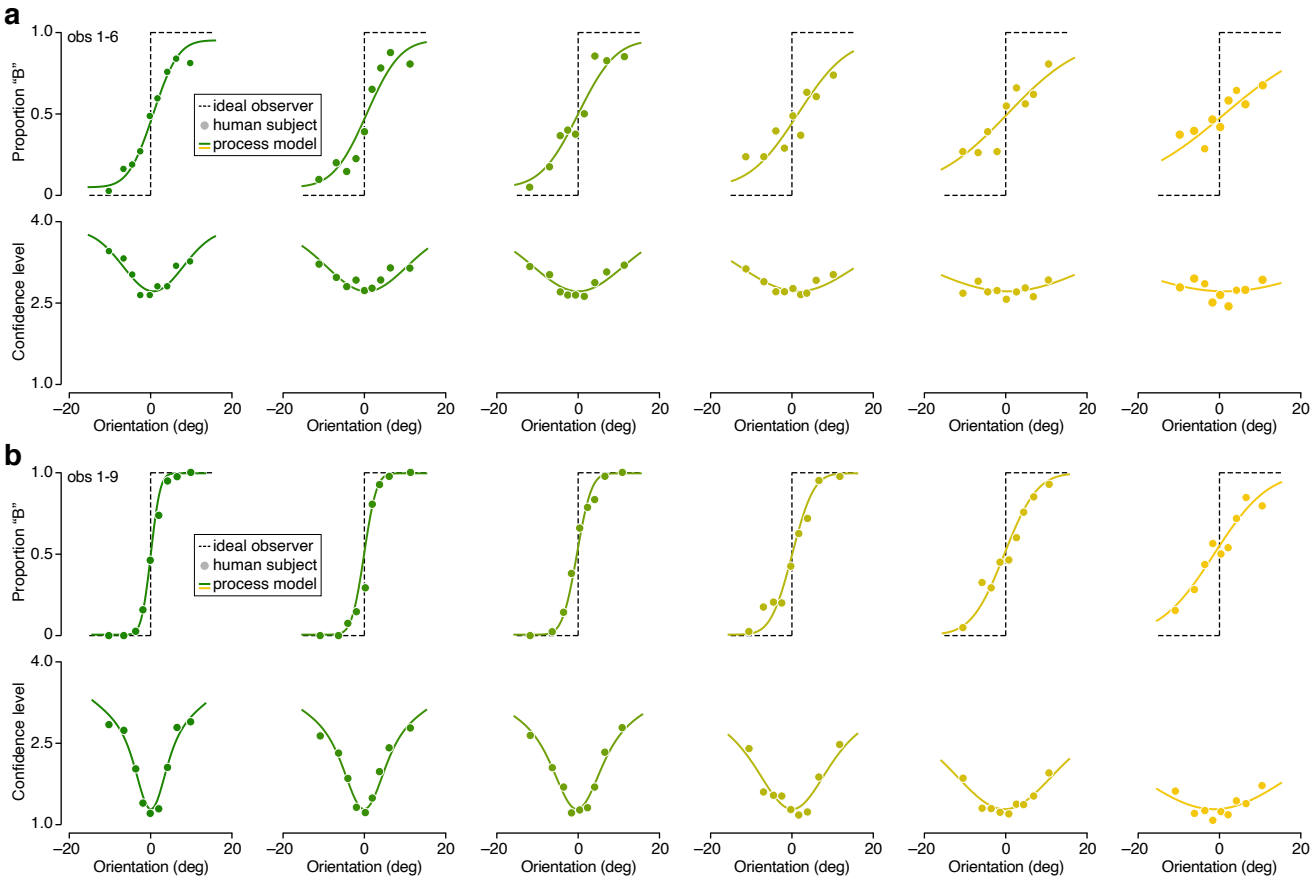
Data and code availability

Most data used in this study are available from the Confidence Database (available at: https://osf.io/s46pr/). The code supporting the findings of this study is available from the corresponding author upon reasonable request, and will be available in a public repository shortly.

Supplementary information

Adler and Ma (2018), task 1

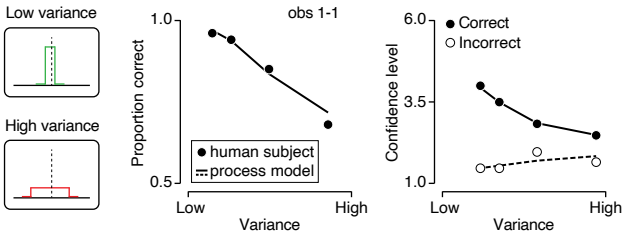
Figure 1b,c shows data from two subjects who performed a perceptual 2-AFC categorization task and additionally reported their confidence using a four-point rating scale. Data were collected by Adler and Ma (2018). Supplementary Figure 1a,b illustrates model fits by plotting the psychometric function (top row) and accompanying confidence function (bottom row) for each stimulus contrast (columns). Subjects completed 2,160 trials each. To model these data, we used one lapse rate parameter (obs 1-6: 5%; obs 1-9: 0.5%), one contrast-specific sensitivity parameter (obs 1-6: 0.21, 0.15, 0.14, 0.11, 0.07, and 0.06; obs 1-9: 0.55, 0.42, 0.36, 0.24, 0.16, and 0.10), one contrast-specific decision criterion parameter (obs 1-6: 0.62, 0.55, 0.00, 1.49, 0.45, and 1.28 degrees; obs 1-9: 0.03, -0.08, -0.32, 0.05, -0.35, and -1.33 degrees), one meta-uncertainty parameter (obs 1-6: 0.21; obs 1-9: 0.51), and three confidence criterion parameters (obs 1-6: 0.02, 0.40, and 1.92; obs 1-9: 1.42, 3.30, and 10.99). The log-probability of the data under the model was –3,462.1 for obs 1-6, and –2,654.0 for obs 1-9.



Supplementary Figure 1 Model fits for two example subjects from Adler and Ma (2018). Both subjects judged whether a stimulus belonged to category A or B. Category A stimuli typically had an orientation smaller than zero, while category B stimuli typically had an orientation larger than zero. Stimuli varied in orientation and contrast. (**a**) Top: Proportion of "Category B" choices is plotted against stimulus orientation, split out by stimulus contrast (columns, contrast decreases from left to right), for one example subject. Bottom: Same for mean confidence level. Symbols summarize observed choice behavior, the dotted line illustrates the theoretical optimum, and the full lines show the fit of a two-stage process model of decision-making. Symbol size is proportional to the number of trials. (**b**) Same for a different example subject.

Navajas et al. (2017), experiment 1

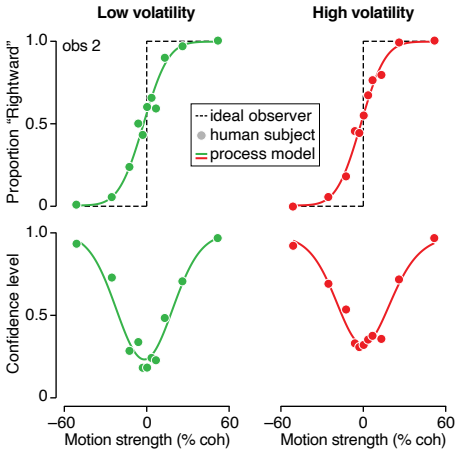
Figure 4a-d shows an analysis of data from subjects who performed either a perceptual or cognitive 2-AFC categorization task and additionally reported their confidence using a six-point rating scale. Data were collected by Navajas et al. (2017). Supplementary Figure 2 illustrates the model fit for an example subject by plotting the data in the format used in the original publication[25]. Proportion correct and confidence level are plotted against stimulus variance. The confidence reports are split out by decision accuracy. The experiment consisted of 400 trials. To model these data, we used one lapse rate parameter (0%), one stimulus variance-specific sensitivity parameter (0.67, 0.53, 0.33, and 0.19), one decision criterion parameter (–0.59 degrees), one meta-uncertainty parameter (0.10), and five confidence criterion parameters (0.35, 1.19, 1.64, 2.29, and 3.36). The log-probability of the data under the model was –733.50.



Supplementary Figure 2 Model fit for an example subject from Navajas et al. (2017). The subject judged whether the mean orientation of a sequence of 30 rapidly presented Gabor stimuli was tilted right or left. Left: Stimulus sequences were sampled from distributions with different orientation variance. Middle: Proportion correct choices is plotted against stimulus variance for an example subject. Right: Mean confidence level is plotted against stimulus variance, split out by decision accuracy. Symbols summarize observed choice behavior, the full line shows the fit of a two-stage process model of decision-making.

Zylberberg et al. (2016), confidence experiment

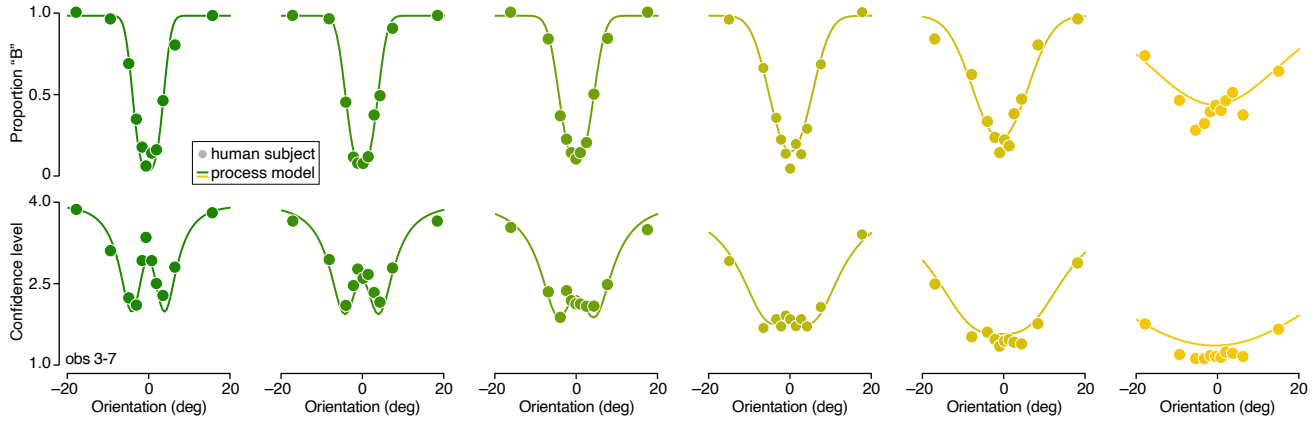
Figure 6a shows an analysis of data from subjects who performed a perceptual 2-AFC categorization task and additionally reported their confidence using a continuous rating scale. Data were collected by Zylberberg et al. (2016). Supplementary Figure 3 illustrates the model fit for an example subject. Note that we binarized the continuous confidence ratings. Proportion "rightward" choices and confidence level are plotted against motion coherence, split out by stimulus volatility. The subject completed 2,104 trials. To model these data, we used one lapse rate parameter (0.64%), one stimulus volatility-specific sensitivity parameter (7.09 and 7.22), one stimulus volatility-specific decision criterion parameter (-0.03 and -0.02 % coherence), one stimulus volatility-specific meta-uncertainty parameter (0.46 and 0.76), one stimulus volatility-specific confidence criterion parameter for leftward choices (1.35 and 1.17), and one stimulus volatility-specific confidence criterion parameter for rightward choices (1.44 and 1.46). The log-probability of the data under the model was –2,055.9.



Supplementary Figure 3 Model fit for an example subject from Zylberberg et al. (2016). The subject judged whether a random dot motion stimulus moved leftward or rightward. Stimuli differed in their volatility. Top: Proportion "rightward" choices is plotted against motion coherence, split out by stimulus volatility (left vs right column), for an example subject. Bottom: Same for mean confidence level. Symbols summarize observed choice behavior, the dotted line illustrates the theoretical optimum, and the full line shows the fit of a two-stage process model of decision-making.

Adler and Ma (2018), task 2

Figure 6b shows an analysis of data from subjects who performed two perceptual 2-AFC categorization tasks and additionally reported their confidence using a four-point rating scale. Data were collected by Adler and Ma (2018). Supplementary Figure 4 illustrates model fits for task 2 by plotting the psychometric function (top row) and accompanying confidence function (bottom row) for each stimulus contrast (columns). The subject completed 3,240 trials. To model these data, we used one lapse rate parameter (2.17%), one contrast-specific sensitivity parameter (0.68, 0.55, 0.46, 0.28, 0.20, and 0.07), one contrast-specific low decision criterion parameter (-4.05, -4.27, -4.35, -5.21, -7.30, and -11.34 degrees), one contrast-specific high decision criterion parameter (3.62, 3.76, 4.26, 5.36, 5.38, and 9.66 degrees), one meta-uncertainty parameter (0.69), three confidence criterion parameters for "Category A" choices (0.69, 1.15, and 1.89), and three confidence criterion parameters for "Category B" choices (0.65, 1.78, and 4.53). The log-probability of the data under the model was –4,842.5.



Supplementary Figure 4 Model fits for an example subject from Adler and Ma (2018). The subject judged whether a stimulus belonged to category A or B. Category A stimuli were drawn from a distribution with small orientation spread, category B stimuli were drawn from a distribution with large orientation spread. Stimuli varied in orientation and contrast. Top: Proportion of "Category B" choices is plotted against stimulus orientation, split out by stimulus contrast (columns, contrast decreases from left to right), for one example subject. Bottom: Same for mean confidence level. Symbols summarize observed choice behavior, the full lines show the fit of a two-stage process model of decision-making.