nature human behaviour

Article https://doi.org/10.1038/s41562-022-01464-x

Confidence reflects a noisy decision reliability estimate

Received: 7 January 2022

Accepted: 21 September 2022

Published online: 07 November 2022

Check for updates

Zoe M. Boundy-Singer[1,2], Corey M. Ziemba[1,2] & Robbe L. T. Goris[1]

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 misjudgements. 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, investigated in six previous studies. We find 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 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' versus '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 (for example, a large willingness to declare 'high confidence' or 'category A')[17–19]. Yet, neither of these factors characterizes the subject's metacognitive ability[13].

Here, we introduce a method to quantify metacognitive ability on the basis of choice-confidence data, building on an extensive body of previous work[13,20–27]. We propose that confidence reflects a subject's estimate of the reliability of their decision, expressed in units of signal-to-noise ratio[27]. This estimate results from a computation involving the uncertainty of the decision variable that informed the subject's choice[28]. It follows that metacognitive ability is determined by the subject's knowledge about this uncertainty, or lack thereof (that is, 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[29] 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[27,30] 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

[1]Center for Perceptual Systems, The University of Texas at Austin, Austin, TX, USA. [2]These authors contributed equally: Zoe M. Boundy-Singer, Corey M. Ziemba. e-mail: robbe.goris@utexas.edu

studies[22,25,31–34]. 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 comparable cross-domain generalizability. Meta-uncertainty is higher in tasks that involve more levels of stimulus uncertainty, implying that it can be manipulated experimentally. Together, these results illuminate the mental operations that give rise to our sense of confidence, and they provide evidence 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[27]. 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 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). The more the stimulus orientation deviates from zero, the more closely human subjects' aggregated choice behaviour approximates this ideal (Fig. 1b, top, symbols). This relationship is also modulated by stimulus contrast—the lower the stimulus contrast, the weaker the association between orientation and choice (Fig. 1b, top, green versus 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 (that is, the most difficult conditions), and monotonically increases as choice proportions deviate more from 0.5 (Fig. 1b, bottom). The association between choice consistency and confidence is so strong that plotting average confidence level against the aggregated choice behaviour 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. We speculate that a lawful confidence–consistency relationship is not a coincidental feature of this experiment, but a widespread phenomenon in confidence studies (Fig. 1d).

A single, increasing relation between confidence reports and choice consistency across many levels of uncertainty implies that subjects can assess the reliability of their decisions. However, whether this ability is excellent or poor cannot be deduced from empirical measurements alone. One possibility is that subjects accurately assess decision reliability on every single trial, indicating excellent metacognitive ability. Alternatively, there might be a high degree of cross-trial variability in confidence reports, implying less accurate decision reliability assessment and thus limited metacognitive ability. Of course, given the variability of primary choice behaviour, 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 (for example, 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 favours one response alternative over the other (for example, 'Have I heard this song before?'). 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 primary response bias[29].

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[14,35–39]. 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, and thus produce a single confidence–consistency relation, the stimulus strength estimate must therefore be normalized by this factor[27]. 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 (that is, 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[40]. It is also convenient, as it results in fairly simple data-analysis recipes[29]. 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[25,40]. 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 (Methods). This mathematical street-fighting manoeuvre[41] 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

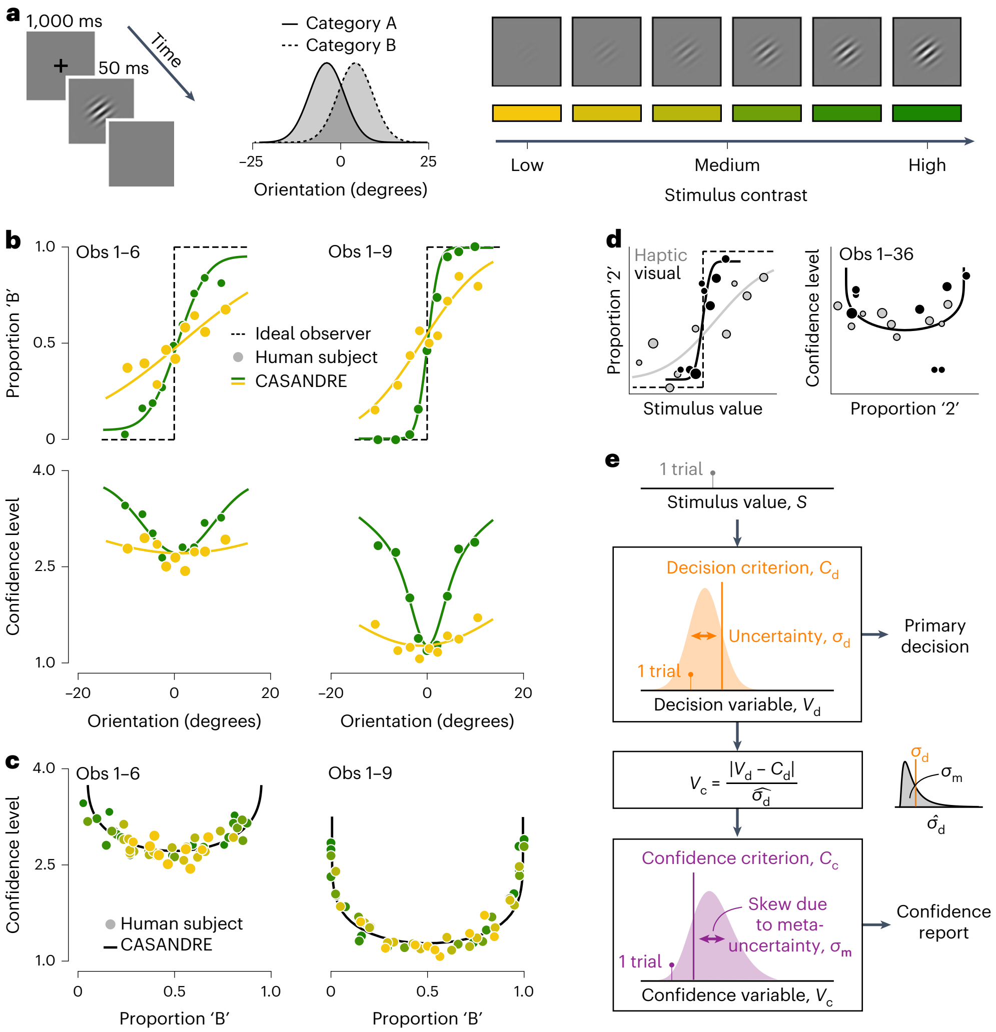


Fig. 1 | CASANDRE, a two-stage process model of decision confidence, accounts for the relation between confidence reports and choice consistency. a, Experimental design employed by ref. [22]. **b**, Top, proportion of 'category B' choices is plotted against stimulus orientation, split by stimulus contrast (green versus yellow), for two example subjects (left, obs 1–6: observer 6 in experiment 1 from ref. [22]; right, obs 1–9: observer 9 in experiment 1 from ref. [22]). Bottom, same for mean confidence level. Symbols summarize observed choice behaviour, 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. The model was fit to all data simultaneously using a maximum likelihood estimation method. Only two out of six contrasts are shown here. Fits to 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 (observer 36 in experiment 1 from Arbuzova and Filevich in the Confidence Database[31]). **e**, Schematic of the hierarchical decision-making process underlying choice-confidence data in the CASANDRE model.

behaviour 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 (Methods). Altering the first-stage decision criterion (Fig. 2a, top left, orange versus 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 versus grey distribution). At the level of observables, 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[27,42,43] and animal[44,45] subjects. Altering the level of first-stage noise (Fig. 2b, top left, orange versus 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 versus grey distribution). At the level of choice behaviour, 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). These first-stage parameters do not affect the shape of the confidence–consistency relationship, only where a particular stimulus value falls on this curve. In contrast, the parameters that control the model's second-stage operations do not affect primary choice behaviour but only confidence reports and thus the shape of the confidence–consistency relationship. Specifically, changing the confidence criterion (Fig. 2c, bottom left, purple versus 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 versus grey distribution), resulting in a complex pattern of changes in the confidence function (Fig. 2d, bottom right).

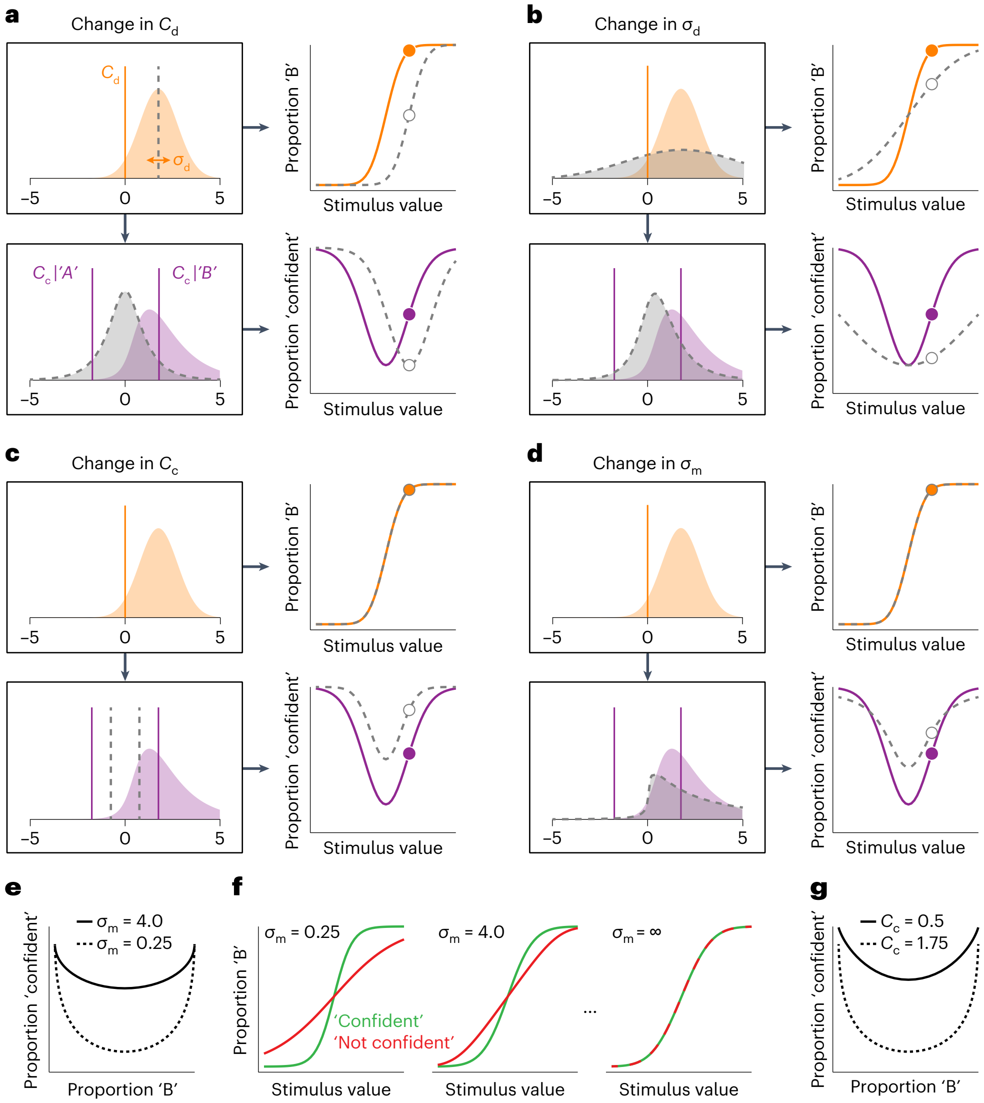


Fig. 2 | Impact of the different model components on primary choice behaviour and confidence reports. a, Top left, illustration of the decision criterion (orange line) and the decision variable distribution elicited by repeated presentations of the same stimulus (orange distribution). Bottom left, the 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. For all panels, the grey dotted line illustrates how the model predictions change when a specific model component (here, 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. All other model parameters held equal. **f**, The psychometric function, split by confidence report ('confident' in green versus 'not confident' in red), for three levels of meta-uncertainty. **g**, The confidence–consistency relation under a liberal versus a conservative confidence criterion. All other model parameters held equal, $\sigma_m = 0.25$.

What does it mean to say that someone has good or bad self-knowledge? The CASANDRE model provides a principled answer. 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 by confidence report. When meta-uncertainty is low, 'confident' decisions tend to be much more reliable than 'not confident' decisions (Fig. 2f, left, green versus 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)[27]. However, the magnitude of the effects shown in Fig. 2e,f depends on the other model components as well (for example, Fig. 2g). Determining the level of meta-uncertainty therefore requires directly fitting the model to data.

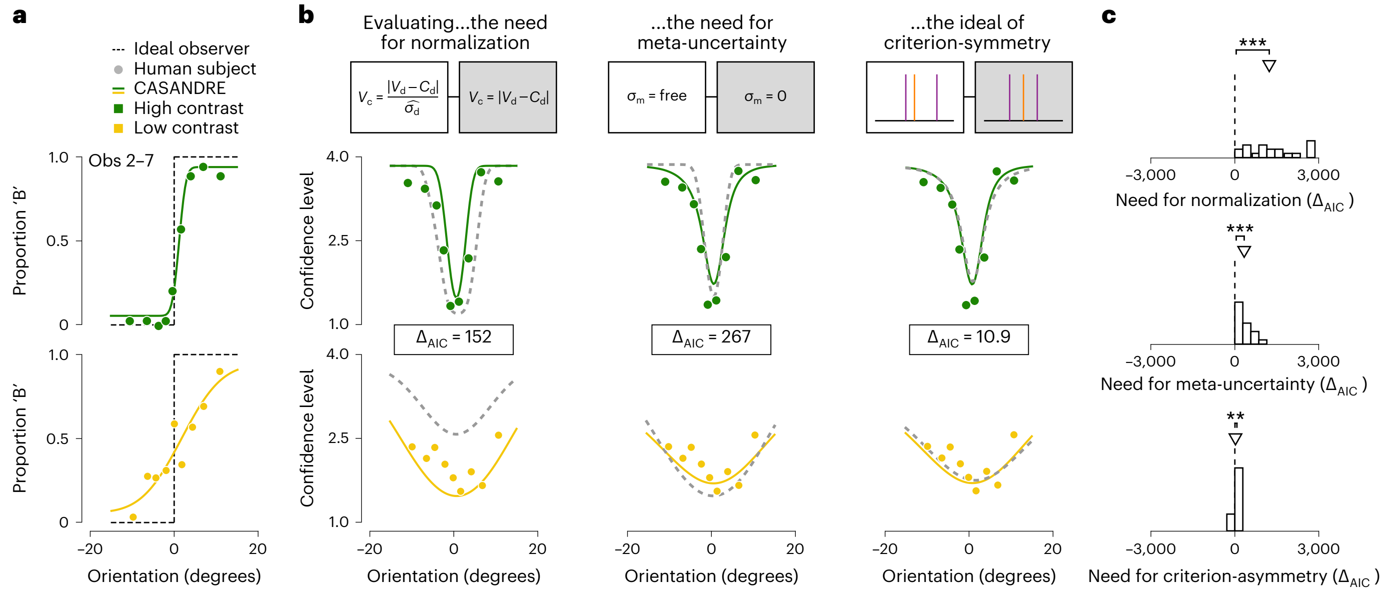


Fig. 3 | Comparison of different model architectures. **a**, Proportion of 'category B' choices is plotted against stimulus orientation for high and low contrast stimuli (top versus bottom). Symbols summarize observed choice behaviour of an example subject (observer 7 in experiment 2 from ref. [22]), the dotted line illustrates the theoretical optimum and the full lines show the fit of the first stage which is shared across all model variants examined in this analysis. As previously, the model was fit to all data simultaneously. **b**, Mean confidence level is plotted as a function of stimulus orientation for the example subject. Left, fits of two model variants in which confidence either reflects an estimate of decision reliability (full lines) or of stimulus strength (dashed lines). Middle, fits of two model variants in which confidence either reflects a noisy (full lines) or noiseless (dashed lines) estimate of decision reliability. Right, fits of two model variants in which confidence criteria either depend (full lines) or not (dashed lines) on the primary decision. **c**, Distribution of the difference in AIC value for each model comparison across 19 subjects. Positive values indicate evidence for the more complex model variant. Arrows indicate median of the distribution. $^{***}P < 0.001$, $^{**}P < 0.01$, Wilcoxon signed rank test. (Top, $z = 3.82$, $P < 0.001$, effect size = 0.88; Middle, $z = 3.82$, $P < 0.001$, effect size = 0.88; Bottom, $z = 3.06$, $P = 0.002$, effect size = 0.70).

Evaluating the model architecture

We have motivated our framework on the basis of a qualitative observation (the lawful confidence–consistency relationship) and first principles (the inherent noisiness of perceptual and cognitive processes). To further test the central tenets of the CASANDRE model, we quantitatively examined the choice-confidence data collected by Adler and Ma[22]. We conducted several model comparisons designed to interrogate the framework's second-stage operations. For this reason, we began by fitting the first-stage parameters to each subject's choice data and then kept these parameters constant across all model variants (see example in Fig. 3a). We first asked whether a simpler computation can account for confidence reports. We compared a model variant in which confidence reflects a subject's estimate of stimulus strength[14,35–39] with one in which it reflects an estimate of decision reliability (that is, stimulus strength normalized by stimulus uncertainty; Fig. 3b, left). To quantify model quality, we computed each model's Akaike Information Criterion (AIC) value (Methods). For all 19 subjects, the more complex model outperformed the simpler variant (median difference in AIC = 1,179.5; Fig. 3c, top). We then asked whether meta-uncertainty is a necessary model component, and found this to be the case (Fig. 3b, middle). Including meta-uncertainty improved model quality for all 19 subjects (median difference in AIC = 285.2; Fig. 3c, middle). Both these simpler variants of the CASANDRE model correspond to commonly used models for confidence[1,14,22,25,27]. These model comparisons thus support the hypothesis that confidence reflects a subject's noisy estimate of the reliability of their decision.

Further attempts to improve the model architecture yielded comparatively weak and inconsistent results. In particular, we wondered whether model performance would benefit from allowing criterion-asymmetry (meaning that the confidence criteria depend on the primary decision) and adopting a different second-stage noise distribution (the Gamma distribution). Allowing criterion-asymmetry improved model performance for 16 out of 19 subjects (median difference in AIC = 27.9; Fig. 3b, right; Fig. 3c, bottom; different example subject shown in Supplementary Fig. 2), while the log-normal distribution was preferred over the Gamma distribution for 16 out of 19 subjects (median difference in AIC = 17.7). For simplicity, we chose to use symmetric confidence criteria for all further analyses. Finally, we compared the CASANDRE model with a model recently proposed by Shekhar and Rahnev[25] (the 'log-normal meta-noise model'). In this model, confidence reflects a subject's estimate of evidence strength and metacognitive ability is limited by simulating instability in the placement of confidence criteria[25]. As this model is tailored to experiments that employ only two levels of stimulus strength, we examined the choice-confidence data collected by Shekhar and Rahnev and found that the CASANDRE model explained these data equally well. (Supplementary Fig. 8a; see Supplementary information for further discussion).

Estimating meta-uncertainty from sparse data

We seek to quantify a subject's ability to introspect about the reliability of a decision. Our method consists of interpreting human choice-confidence data through the lens of a principled two-stage process model. What kind of measurements are required to obtain robust and reliable estimates of meta-uncertainty, the model's parameter that governs metacognitive ability? We verified that Adler and Ma's experimental design affords solid parameter recovery (see Supplementary Fig. 3). However, their design is exceptional for its large number of stimulus conditions[22]. Many studies use as little as two conditions[31]. To test whether our approach generalizes to such experiments, we performed a recovery analysis. We used the CASANDRE model to generate synthetic datasets for five model subjects performing a 2-AFC discrimination task with binary confidence report options (Methods). The model subjects only differed in their level of meta-uncertainty,

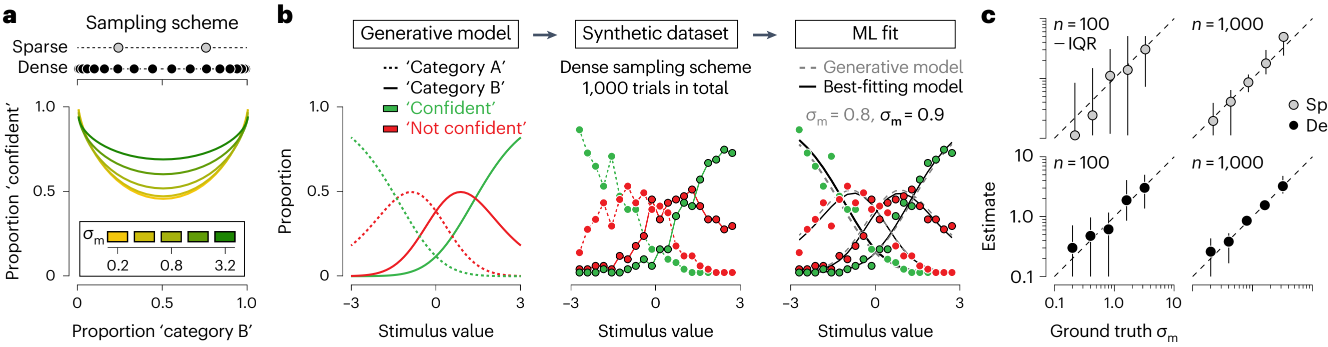


Fig. 4 | Model recovery analysis. a, We simulated choice-confidence data for five model subjects who differed in their level of meta-uncertainty (coloured lines) using experimental designs that varied in the number of trials (100 versus 1,000) and in the number of conditions (2 versus 20, grey and black 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. Meta-uncertainty was limited to a minimum value of 0.1. Error bars illustrate the interquartile range (IQR) computed from 100 simulated datasets.

which ranged from negligible to considerable (Fig. 4a, coloured lines). We simulated data for each model subject using experimental designs that varied in the number of trials (100 versus 1,000) and in the number of conditions (2 versus 20; Fig. 4a, top). Figure 4b 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 (Fig. 4b, left). We used these probabilities to simulate a synthetic dataset of 1,000 trials distributed across 20 conditions (Fig. 4b, middle). We then identified the set of parameter values that best describes these data (Fig. 4b, 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. 4c, symbols). The reliability of these estimates is higher for more trials and somewhat higher for denser stimulus sampling (Fig. 4c, error bars). Estimation error in σ_m covaried with estimation error in C_c (Supplementary Fig. 7). We conclude that the CASANDRE model typically can be identified in sparse experimental designs.

Construct reliability and validity of meta-uncertainty

So far, we have presented evidence that confidence is well described as reflecting a subject's decision reliability estimate. In the CASANDRE model, the quality of this estimate is limited by meta-uncertainty. This naturally raises the question of whether meta-uncertainty is a 'real' thing. In other words, do meta-uncertainty estimates isolate a stable property of human subjects that captures their metacognitive ability?

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. conducted a perceptual confidence experiment in which 14 subjects performed the same task twice with approximately 1 month in between sessions[32]. We used the CASANDRE model to analyse their data (see Methods and Supplementary Fig. 4). 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 $r(170) = 0.96$, $P < 0.001$; condition-specific mean confidence level: $r(170) = 0.99$, $P < 0.001$). Critically, we found meta-uncertainty estimates to be strongly correlated across both sessions as well ($r(12) = 0.78$, $P = 0.002$; Fig. 5a). This suggests that meta-uncertainty measures a stable characteristic of human confidence reporting behaviour.

Under the CASANDRE model, meta-uncertainty provides a measure of metacognitive ability, not of confidence reporting strategy. To investigate whether this holds true in human choice-confidence data, we analysed 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[32]. We artificially biased these confidence reports by mapping them onto a liberal and a conservative four-point rating scale (Methods)[46]. 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 analyse both perturbed datasets (Methods). Meta-uncertainty estimates were strongly correlated ($r(41) = 0.84$, $P < 0.001$; Fig. 5b), though note that they were on average higher for the conservatively biased version of the data (mean increase: 47%, median increase: 0%, $z = 3.10$, $P = 0.002$, effect size = 0.47, Wilcoxon signed rank test). This suggests that meta-uncertainty estimates are largely, but not fully, independent of subjects' confidence reporting strategy.

We wondered whether meta-uncertainty depends on the absolute level of stimulus uncertainty[23]. We analysed 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[32]. We used the CASANDRE model to analyse these data and estimated meta-uncertainty separately for the two lowest and the two highest levels of stimulus variance (Methods). The former conditions resulted in much higher task performance than the latter (average proportion correct decisions: 87% versus 70%). 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%, $z = -2.87$, $P = 0.004$, effect size = −0.44, Wilcoxon signed rank test). Moreover, meta-uncertainty estimates were strongly correlated across both sets of conditions ($r(41) = 0.70$, $P < 0.001$; Fig. 5c). This suggests 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 debated[8,47–49]. We analysed data from 20 subjects who performed a perceptual and cognitive confidence task with 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[32]. Subjects' performance level was correlated across both tasks (condition-specific proportion correct choices: $r(18) = 0.69$, $P < 0.001$), and so were their reported confidence levels, albeit to a lesser degree ($r(18) = 0.53$, $P < 0.001$). We used the CASANDRE model to analyse both datasets (Methods). Meta-uncertainty estimates were strongly correlated

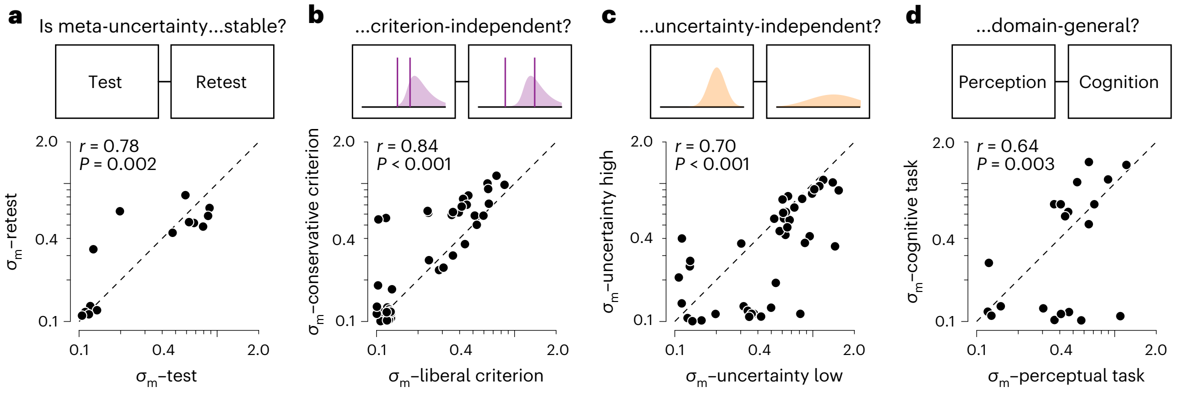


Fig. 5 | Evaluating meta-uncertainty as a psychological construct. **a**, Comparison of meta-uncertainty estimates for 14 subjects who performed the same perceptual confidence task on two different occasions, separated by 1 month. We added a small amount of jitter to get a better view of overlapping data points in the lower left region of the plot. Meta-uncertainty was limited to a minimum value of 0.1. $r(12) = 0.78$, $P = 0.002$, Spearman correlation. **b**, Comparison of meta-uncertainty estimates for 43 sessions (performed by 32 subjects, see Methods) whose six-point confidence ratings were mapped onto a liberal and conservative 4-point rating scale. $r(41) = 0.84$, $P < 0.001$. **c**, Comparison of meta-uncertainty estimates for 43 sessions where subjects performed a confidence task involving low and high levels of stimulus uncertainty. $r(41) = 0.70$, $P < 0.001$. **d**, Comparison of meta-uncertainty estimates for 20 subjects who performed a perceptual and cognitive confidence task. $r(18) = 0.64$, $P = 0.003$.

($r(18) = 0.64$, $P = 0.003$; Fig. 5d). Thus, meta-uncertainty appears to capture an aspect of confidence reporting behaviour that generalizes across at least some domains.

We found that other commonly used metrics for metacognitive ability do not isolate the factors determining meta-uncertainty, and that meta-uncertainty compared favourably on this set of four benchmark tests (Supplementary Fig. 9; see Supplementary information for further discussion).

Manipulating meta-uncertainty

Can metacognitive ability be manipulated experimentally? Key to our framework is that confidence judgements require a subject to estimate uncertainty on a trial-by-trial basis. This becomes more difficult when experiments involve more confusable levels of stimulus uncertainty. We therefore expect that meta-uncertainty will grow with the number of stimulus uncertainty levels. To appreciate our logic, consider the ideal Bayesian uncertainty estimation strategy which consists of combining information obtained from ambiguous sensory measurements with prior task-specific knowledge. Specifically, the sensory measurement informs the uncertainty likelihood function, while knowledge of task statistics (that is, the distribution of stimulus uncertainty levels) is summarized in a prior uncertainty belief function (Fig. 6a). The combination of both yields a posterior uncertainty belief function, the maximum of which is the 'best possible' uncertainty estimate (Fig. 6a). Due to noise, repeated presentations of the same condition will yield different likelihood functions (Fig. 6a and Methods). If the task involves only one level of stimulus uncertainty, the prior is a fixed delta function, and so is the posterior. Consequently, the maximum posterior estimate will not vary across trials and the ideal estimation strategy results in zero meta-uncertainty. However, when a task involves multiple levels of stimulus uncertainty, the prior will be more dispersed, causing the resulting maximum posterior estimate to be more variable across trials. Under an ideal Bayesian estimation strategy, meta-uncertainty thus initially grows with the number of uncertainty levels (Fig. 6b). We wondered whether this normative prediction affords insight into human metacognition. To test this hypothesis, we used the CASANDRE model to analyse six confidence experiments that varied in the number of randomly interleaved uncertainty levels (Methods). These experiments utilized different stimuli and employed different experimental designs[22,25,32–34]. Yet, as expected, meta-uncertainty appears to grow lawfully with the number of uncertainty levels (Fig. 6c).

Discussion

It has long been known that humans and other animals can meaningfully introspect about the quality of their decisions and actions[5–7,37,50]. Quantifying this ability has remained a challenge, even for simple binary decision-making tasks[12,13,15,23,25,46]. 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 behavioural 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[14,27,51]. 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. Meta-uncertainty appears to be mostly independent of task difficulty and confidence reporting strategy. Widely used metrics for metacognitive ability are poor proxies for meta-uncertainty. As such, the CASANDRE model represents a notable advance toward realizing crucial medium and long-term goals in the field of metacognition[52].

The mental operations underlying confidence in a decision have long intrigued psychologists. Two key unresolved issues are the structure and nature of the confidence computation[52]. At stake are two intertwined questions: (1) Does confidence arise from a single, dual or hierarchical process? and (2) What exactly does confidence reflect? Some authors have proposed that decision outcome and confidence both arise from a single stimulus strength estimation process[37,53–55]. Such models can explain the effects of stimulus strength, but not of stimulus reliability. Others have argued in favour of a dual process in which decision outcome and confidence are based on different stimulus strength estimates[21,56–58]. This may be the appropriate framework for cases in which subjects acquire additional task-relevant information after reporting their choice[21,24,59,60]. For all other cases, it appears overly complex. Instead, we have modelled confidence judgements as arising

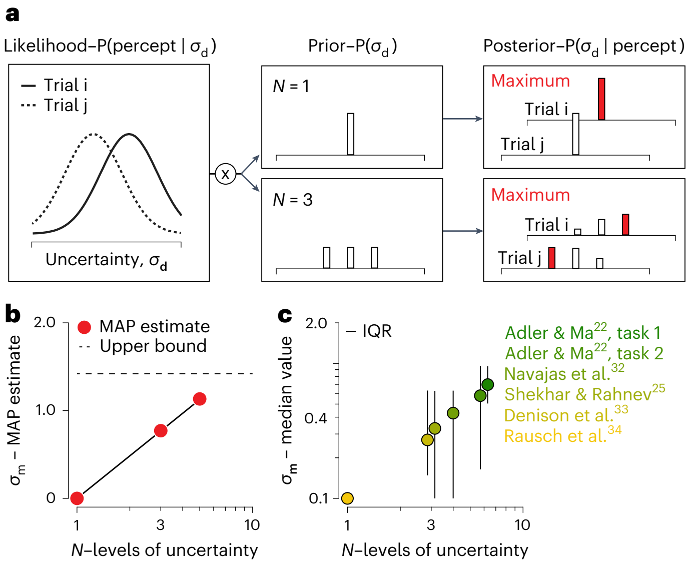


Fig. 6 | Meta-uncertainty depends on task structure. **a**, We studied how meta-uncertainty depends on the number of uncertainty levels under an ideal Bayesian uncertainty estimation strategy. The likelihood of each uncertainty value is computed from a sensory measurement (left) while a prior belief function specifies task-specific knowledge of possible uncertainty values (middle). The product of the prior and likelihood gives the posterior (right). Due to noise, the likelihood function will differ across repeated trials (left, full versus dotted line). The impact of this variability on the posterior depends on the dispersion of the prior (right, top versus bottom panel). **b**, Meta-uncertainty plotted against number of uncertainty levels for the ideal Bayesian estimator. The upper bound (dotted line) is set by the cross-trial variability of the maximum of the likelihood function and is reached when the prior is a uniform distribution. **c**, Median level of meta-uncertainty plotted against number of uncertainty levels for six confidence experiments. Meta-uncertainty was limited to a minimum value of 0.1. Error bars illustrate the interquartile range (IQR) across subjects (Adler and Ma[22] task 1: 19 subjects; task 2: 34 subjects; Navajas et al.[32]: 50 subjects; Shekhar and Rahnev[25]: 20 subjects; Denison et al.[33]: 12 subjects; Rausch et al.[34]: 25 subjects).

from a hierarchical process[61]. The first stage determines the choice, the second stage determines confidence (Fig. 1e). We found that this model structure systematically outperforms a single stage alternative (Fig. 3c, top). The structure of the computation clarifies its nature. Many previous studies are built on the premise that confidence reflects a subject's assessment of decision accuracy ('What is the probability that my choice is correct?'). This premise directly motivates Bayesian models of confidence[1,22,37,62–68] and tacitly underlies popular metrics of metacognitive ability[13,27]. However, when experimental manipulations bias perceptual choices, aggregated confidence reports do not track choice accuracy but choice consistency[27,42,43]. At the single trial level, this suggests that confidence reflects a subject's assessment of decision reliability ('What is the probability that I would make the same choice again?', see equation 1). For an unbiased subject who is choosing between two alternatives, decision accuracy and decision reliability are indistinguishable[27,67]. Yet, the distinction matters greatly, as it implies that the same computation that underlies confidence in decisions with a well-defined correct and incorrect option may generalize to subjective domains that lack this feature (for example, 'Which ice cream flavour should I have?')[69].

Key to our proposal is that assessing the reliability of a decision requires the use of additional information (stimulus uncertainty)[28] that in most tasks has no relevance for the choice as such. The notion that subjects can incorporate a stimulus uncertainty estimate when making perceptual inferences is well established[22,70–72]. And 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[68,73–78]. Our proposed confidence computation yielded a new prediction: the more levels of stimulus uncertainty a task involves, the more variable uncertainty estimates will be. We validated this prediction by analysing data from six different confidence experiments in which 160 subjects completed a total of 243,000 trials (Fig. 6c). This finding is arguably the strongest piece of empirical evidence that meta-uncertainty is the critical factor that limits human metacognitive ability. It was enabled by the use of modern computational tools to quickly compute the approximate ratio of two distributions (that is, the confidence variable distribution) and by the availability of the confidence database[31]. This phenomenon also raises the question to what degree metacognitive ability estimates are influenced by experimental design. We speculate that many previously proposed sources of metacognitive limitations (such as arousal or fatigue[79,80], sequential dependencies of confidence reports or 'confidence leak'[81], and disruptions to frontal cortical function[82]) could be mediated through their effect on the fidelity of uncertainty estimation. An important future direction will be to investigate the Bayesian uncertainty estimation framework introduced here (in Fig. 6a,b), in particular by examining how meta-uncertainty is affected by manipulations of both the likelihood and prior distributions of stimulus 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[83,84]. For this reason, some authors have suggested that decision time directly informs confidence[59,85]. This proposal enjoys strong empirical support[44,63,85,86]. It remains to be seen whether choice outcome, reaction time and metacognitive ability can all be modelled simultaneously.

Process models are powerful tools to study cognition and perception. Here we leveraged a process model to interrogate the computations underlying our sense of confidence, to determine the effectiveness of various experimental designs and to examine model recoverability. However, the usefulness of process models far exceeds our current application. Specifically, when coupled to an explicit goal such as maximizing choice accuracy, process models can be used to derive the optimal task strategy. The resulting predictions offer a critical point of reference for human behaviour[87]. This approach has revealed that humans improve the quality of uncertain decisions by accumulating evidence over time[83], by combining information acquired through different sensory modalities[70], and by exploiting knowledge of statistical regularities in the environment[88]. Might the same hold true for uncertain confidence judgements? 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 (for example, refs. [37,50,55,57,63,89–91]). 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. We took a first step in this direction and validated a new prediction: meta-uncertainty changes with task structure as expected under an ideal Bayesian uncertainty estimation strategy.

Methods

Modeling the 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[29] and describes 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 V_c' with a fixed criterion magnitude, C_c, mirrored across 0 to create two signed criteria (or two sets 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 (that is, 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%. We also note that in our Matlab implementation of the model, we fix σ_d to 1 and scale the relation between stimulus value and decision variable mean. This is equivalent to fixing the relation between stimulus value and decision variable mean while varying σ_d. 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, just like the absolute level of σ_d cannot be uniquely estimated from choice data).

Model parameterization, simulations and fitting

We analysed 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 (that is, the model simulations shown in Figs. 2 and 4). 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. We limited σ_m to a minimum value of 0.1, as values below this had indistinguishable effects on model behaviour. Figure 2 shows how each of the parameters affects the model's behaviour. Finally, when fitting data, we use one additional parameter, λ, to account for stimulus-independent lapses[92], 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 4b shows an example model fit to a synthetic dataset whereby we used five 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 (that is, refs. [22,25,32,34]). 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 modelled the data from ref. [34] using seven free parameters: λ, σ_d, C_d, σ_m and C_c (four-point confidence rating scale, thus three in total) (Fig. 6c and Supplementary Fig. 5a). We modelled some data from ref. [22] (task 1) using 17 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 (four-point confidence rating scale, thus three in total). Example fits are shown in Fig. 1b,c and in Supplementary Fig. 1 (also see Fig. 6c, task 1 and Supplementary Fig. 5e). We modelled the data from ref. [32] using 12 free parameters: λ, σ_d (one per stimulus variance level, four in total), C_d, σ_m and C_c (six-point confidence rating scale, thus five in total). Example fits are shown in Supplementary Fig. 4 (also see Fig. 6c, Supplementary Fig. 9b–d and Supplementary Fig. 5d). We modelled the data from ref. [25] using 10 free parameters: σ_d (one per stimulus reliability level, three in total), C_d, σ_m and C_c (continuous confidence rating scale, discretized into six-point confidence rating scale, thus five in total; see Fig. 6c and Supplementary Fig. 5b).

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 (that is, refs. [22,33]). 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 modelled some data from ref. [22] (task 2) using 23 free parameters: λ, σ_d (one per contrast level, six in total), C_d (two per contrast level, 12 in total), σ_m and C_c (four-point confidence rating scale, thus three in total; see Fig. 6c, task 2). Example fits are shown in Supplementary Fig. 6 (also see Supplementary Fig. 5e). We modelled data from ref. [33] using 22 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 (four-point confidence rating scale, one set per attention level, thus nine in total; see Fig. 6c and Supplementary Fig. 5c).

Some studies used a task design that combined a 2-IFC categorization decision with a confidence report (that is, Arbuzova and Filevich, unpublished but available in the Confidence Database[31]). 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[29]. 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 (for example, a preference for 'interval 1' choices; see example fit in Fig. 1d).

Model comparison

We evaluated CASANDRE's assumed confidence computation and overall model architecture by fitting different model variants to an experiment that involved joint manipulations of stimulus strength and stimulus reliability (ref. [22], task 1, 19 subjects). For each model comparison, we computed the AIC, given by:

$$\mathrm{AIC} = -2\ln(L) + 2k,$$

where L is the maximum value of a model's likelihood function and k is the number of fitted parameters. To focus this analysis on the model's second-stage operations, we began by fitting 13 first-stage parameters to each subject's choice data: λ, σ_d (one per contrast level, six in total) and C_d (one per contrast level, six in total). These parameters were kept constant across all model variants. The head-to-head model comparisons consisted of (1) confidence as a noiseless stimulus strength estimate versus confidence as a noiseless decision reliability estimate,

(2) confidence as a noiseless decision reliability estimate versus confidence as a noisy decision reliability estimate, (3) symmetric confidence criteria versus asymmetric confidence criteria and (4) a log-normal versus Gamma second-stage noise distribution.

Datasets

The majority of our analyses focus on two studies[22,32]. To test the effect of task structure on meta-uncertainty, we additionally analysed data from three other studies[25,33,34]. The data from Navajas et al.[32] were provided by an author[32]. All other datasets were obtained from the Confidence Database[31] (available at: https://osf.io/s46pr/). Given that the CASANDRE model yields more reliable parameter estimates for longer experiments with more stimulus conditions (error bars in Fig. 4c), we included all experiments from the database that involved a large number of subjects, several hundred trials per subject, and multiple levels of stimulus strength and/or stimulus reliability. All detailed experimental designs and procedures are available in the original publications or in abbreviated form in the Confidence Database. We briefly describe each dataset below.

We analysed data from all three experiments in ref. [22]. 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'). As stimulus orientations were drawn from a continuous distribution, to plot the data we grouped nearby orientations into nine 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. 6c (task 1) and Supplementary Fig. 5f. 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. 6. Fitted parameters from all 34 subjects who performed task 2 in experiments 1, 2 and 3 are included in Fig. 6c (task 2) and Supplementary Fig. 5e.

We analysed data from all three experiments in ref. [32]. Thirty subjects performed experiment 1. Fourteen 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 analysed each of these 84 different experimental sessions independently. Data and model fits from an example subject are shown in Supplementary Fig. 5. Fitted parameters and alternative metacognitive metrics from all 14 subjects who performed both experiments 1 and 2 are included in Fig. 5a and Supplementary Fig. 9d (test–retest stability). Fitted parameters and alternative metacognitive metrics from all 20 subjects who performed experiment 3 are included in Fig. 5d and Supplementary Fig. 9d (domain generality). Fitted parameters from 50 subjects performing experiment 1 and the perceptual task of experiment 3 (experiment 3A) are included in Supplementary Figs. 6c and 5d. Further analyses using these data to test the independence of meta-uncertainty from confidence reporting strategy and uncertainty are explained in the next section.

We analysed unpublished data from Arbuzova and Filevich (available in the Confidence Database under the name Arbuzova_unpub_1)[31]. 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. [34] and from all 20 subjects from ref. [25] are included in Fig. 6c. We analysed data from 12 subjects performing a version of task 2 in ref. [22] with an added attention manipulation from ref. [33]. To get the single estimate of meta-uncertainty included in Fig. 6c for each subject, we averaged the values estimated from all three attention conditions, as these were not notably different.

Construct validity analyses

To test the independence between confidence reporting strategy and measures of metacognitive ability, we manipulated the confidence reporting behaviour of subjects across all sessions from ref. [32] (following an analysis developed by ref. [46]). 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] (that is, for the conservative mapping, ratings 2, 3 and 4 were combined, and for the liberal mapping, ratings 3, 4 and 5 were combined.) 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 (and 50 subjects to 32) (Fig. 5b).

To test the independence between stimulus uncertainty and measures of metacognitive ability, we split experimental data from each session in half[32]. 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. 5c.

Bayesian uncertainty estimation

We examined a simple model of Bayesian uncertainty estimation (Fig. 6a,b.). We modelled the uncertainty likelihood function as a log Gaussian function with a geometric mean value, μ_u, that varied from trial to trial. Each trial, μ_u was randomly drawn from a log Gaussian distribution whose geometric average matched the true level of stimulus uncertainty, S_u, and with spread σ_u. As is typical for a well-calibrated model, the spread of the likelihood function equalled σ_u. We assumed three different experimental designs that yielded a prior uncertainty belief function composed of a single delta function ($N = 1$), three delta functions ($N = 3$) and five delta functions ($N = 5$). We simulated 1,000 trials per design. In this simulation, we computed the posterior on a single trial basis and selected its maximum as the MAP uncertainty estimate. Figure 6b summarizes a simulation in which $S_u = 2.5$, $\sigma_u = 1.5$ and the prior belief function peaked at 2.5 for $N = 1$, at 1.67, 2.5 and 3.33 for $N = 3$, and at 0.83, 1.67, 2.5, 3.33 and 4.17 for $N = 5$.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

This study generated no new data. The data used in this study are available from the Confidence Database (available at: https://osf.io/s46pr/).

Code availability

The code supporting the findings of this study and a software package implementing the CASANDRE model is publicly available (https://github.com/gorislab/CASANDRE.git).

References

1. Meyniel, F., Sigman, M. & Mainen, Z. F. Confidence as Bayesian probability: from neural origins to behavior. *Neuron* **88**, 78–92 (2015).
2. Drugowitsch, J., Mendonça, A. G., Mainen, Z. F. & Pouget, A. Learning optimal decisions with confidence. *Proc. Natl Acad. Sci. USA* **116**, 24872–24880 (2019).

3. Purcell, B. A. & Kiani, R. Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proc. Natl Acad. Sci. USA* **113**, E4531–E4540 (2016).
4. Bahador, B. et al. Optimally interacting minds. *Science* **329**, 1081–1085 (2010).
5. Peirce, C. S. & Jastrow, J. On small differences in sensation. *Memoirs of the National Academy of Sciences*, **3**, 75–83 (1884).
6. Ratcliff, R. A theory of memory retrieval. *Psychol. Rev.* **85**, 59–108 (1978).
7. Vickers, D. *Decision processes in visual perception*. (Academic, 1979).
8. de Gardelle, V., Le Corre, F. & Mamassian, P. Confidence as a common currency between vision and audition. *PLoS ONE* **11**, e0147901 (2016).
9. Fleming, S. M., Weil, R. S., Nagy, Z. Dolan, R. J. & Rees, G. Relating introspective accuracy to individual differences in brain structure. *Science* **329**, 1541–1543 (2010).
10. Rouault, M., Seow, T., Gillan, C. M. & Fleming, S. M. Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biol. Psychiatry* **84**, 443–451 (2018).
11. Kuhn, D. in *Children's Reasoning and the Mind* (eds Mitchell, P. & Riggs, K. J.) 301–326 (Psychology Press, 2000).
12. Nelson, T. O. A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychol. Bull.* **95**, 109–133 (1984).
13. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Front. Hum. Neurosci.* **8**, 443 (2014).
14. Mamassian, P. Visual confidence. *Annu. Rev. Vis. Sci.* **2**, 459–481 (2016).
15. Guggenmos, M. Measuring metacognitive performance: type 1 performance dependence and test-retest reliability. *Neurosci. Conscious.* **2021**, niab040 (2021).
16. Festinger, L. Studies in decision: I. Decision-time, relative frequency of judgment and subjective confidence as related to physical stimulus difference. *J. Exp. Psychol.* **32**, 291–306 (1943).
17. Hosseini, J. & Ferrell, W. R. Detectability of correctness: a measure of knowing that one knows. *Instructional Sci.* **11**, 113–127 (1982).
18. Critchfield, T. S. Signal-detection properties of verbal self-reports. *J. Exp. Anal. Behav.* **60**, 495–514 (1993).
19. Galvin, S. J., Podd, J. V., Drga, V. & Whitmore, J. Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychon. Bull. Rev.* **10**, 843–876 (2003).
20. Maniscalco, B. & Lau, H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn.* **21**, 422–430 (2012).
21. Fleming, S. M. & Daw, N. D. Self-evaluation of decision-making: a general Bayesian framework for metacognitive computation. *Psychol. Rev.* **124**, 91–114 (2017).
22. Adler, W. T. & Ma, W. J. Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLoS Comput. Biol.* **14**, e1006572 (2018).
23. Bang, J. W., Shekhar, M. & Rahnev, D. Sensory noise increases metacognitive efficiency. *J. Exp. Psychol. Gen.* **148**, 437–452 (2019).
24. Khalvati, K., Kiani, R. & Rao, R. P. N. Bayesian inference with incomplete knowledge explains perceptual confidence and its deviations from accuracy. *Nat. Commun.* **12**, 5704 (2021).
25. Shekhar, M. & Rahnev, D. The nature of metacognitive inefficiency in perceptual decision making. *Psychol. Rev.* **128**, 45–70 (2021).
26. Mamassian, P. & de Gardelle, V. Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychol. Rev*. https://doi.org/10.1037/rev0000312 (2021).
27. Caziot, B. & Mamassian, P. Perceptual confidence judgments reflect self-consistency. *J. Vis.* **21**, 8 (2021).
28. Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19**, 366–374 (2016).
29. Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics*, Vol. 1 (Wiley, 1966).
30. Koriat, A. The self-consistency model of subjective confidence. *Psychol. Rev.* **119**, 80–113 (2012).
31. Rahnev, D. et al. The confidence database. *Nat. Hum. Behav.* **4**, 317–325 (2020).
32. Navajas, J. et al. The idiosyncratic nature of confidence. *Nat. Hum. Behav.* **1**, 810–818 (2017).
33. Denison, R. N., Adler, W. T., Carrasco, M. & Ma, W. J. Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proc. Natl Acad. Sci. USA* **115**, 11090–11095 (2018).
34. Rausch, M., Zehetleitner, M., Steinhauser, M. & Maier, M. E. Cognitive modelling reveals distinct electrophysiological markers of decision confidence and error monitoring. *NeuroImage* **218**, 116963 (2020).
35. Balakrishnan, J. D. & Ratcliff, R. Testing models of decision making using confidence ratings in classification. *J. Exp. Psychol. Hum. Percept. Perform.* **22**, 615–633 (1996).
36. Ferrell, W. R. A model for realism of confidence judgments: implications for underconfidence in sensory discrimination. *Percept. Psychophys.* **57**, 246–254 (1995).
37. Kepecs, A., Uchida, N., Zariwala, H. A. & Zachary, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227–231 (2008).
38. Treisman, M. & Faulkner, A. The setting and maintenance of criteria representing levels of confidence. *J. Exp. Psychol. Hum. Percept. Perform.* **10**, 119–139 (1984).
39. Wallsten, T. S. & González-Vallejo, C. Statement verification: a stochastic model of judgment and response. *Psychol. Rev.* **101**, 490–504 (1994).
40. Jaynes, E. T. Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630 (1957).
41. Mahajan, S. *Street-Fighting Mathematics: The Art of Educated Guessing and Opportunistic Problem Solving* (MIT Press, 2010).
42. Locke, S. M., Gaffin-Cahn, E., Hosseinizaveh, N., Mamassian, P. & Landy, M. S. Priors and payoffs in confidence judgments. *Atten. Percept. Psychophys.* **82**, 3158–3175 (2020).
43. Mihali, A., Broeker, M. & Horga, G. Insightful inference compensates for distorted perception. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.13.468497 (2021).
44. Fetsch, C. R., Kiani, R., Newsome, W. T. & Shadlen, M. N. Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron* **83**, 797–804 (2014).
45. Fetsch, C. R. et al. Focal optogenetic suppression in macaque area MT biases direction discrimination and decision confidence, but only transiently. *eLife* **7**(July), e36523 (2018).
46. Xue, K., Shekhar, M. & Rahnev, D. Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. *Conscious. Cogn.* **95**, 103196 (2021).
47. McCurdy, L. Y. et al. Anatomical coupling between distinct metacognitive systems for memory and visual perception. *J. Neurosci.* **33**, 1897–1906 (2013).
48. Baird, B., Cieslak, M., Smallwood, J., Grafton, S. T. & Schooler, J. W. Regional white matter variation associated with domain-specific metacognitive accuracy. *J. Cogn. Neurosci.* **27**, 440–452 (2015).
49. Lee, A. L. F., Ruby, E., Giles, N. & Lau, H. Cross-domain association in metacognitive efficiency depends on first-order task types. *Front. Psychol.* **9**, 2464 (2018).
50. Shields, W. E., Smith, J. D., Guttmannova, K. & Washburn, D. A. Confidence judgments by humans and rhesus monkeys. *J. Gen. Psychol.* **132**, 165–186 (2005).

51. Locke, S. M., Landy, M. S. & Mamassian, P. Suprathreshold perceptual decisions constrain models of confidence. *PLoS Comput. Biol.* **18**, e1010318 (2022).
52. Rahnev, D. et al. Consensus goals in the field of visual metacognition. *Perspect. Psychol. Sci.* https://doi.org/10.1177/17456916221075615 (2022).
53. Ko, Y. & Lau, H. A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 1401–1411 (2012).
54. Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T. & Miyamoto, A. Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nat. Neurosci.* **16**, 749–755 (2013).
55. Massoni, S., Gajdos, T. & Vergnaud, J.-C. Confidence measurement in the light of signal detection theory. *Front. Psychol.* **5**, 1455 (2014).
56. Zylberberg, A., Barttfeld, P. & Sigman, M. The construction of confidence in a perceptual decision. *Front. Integr. Neurosci.* **6**, 79 (2012).
57. Maniscalco, B., Peters, M. A. K. & Lau, H. Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Atten. Percept. Psychophys.* **78**, 923–937 (2016).
58. Peters, M. A. K. et al. Perceptual confidence neglects decision-incongruent evidence in the brain. *Nat. Hum. Behav.* **1**, 1–8 (2017).
59. Fetsch, C. R., Kiani, R. & Shadlen, M. N. Predicting the accuracy of a decision: a neural mechanism of confidence. *Cold Spring Harb. Symp. Quant. Biol.* **79**, 185–197 (2014).
60. Murphy, P. R., Robertson, I. H., Harty, S. & O'Connell, R. G. Neural evidence accumulation persists after choice to inform metacognitive judgments. *eLife* **4**, e11946 (2015).
61. Maniscalco, B. & Lau, H. The signal processing architecture underlying subjective reports of sensory awareness. *Neurosci. Conscious.* **2016**, niw002 (2016).
62. Lak, A. et al. Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* **84**, 190–201 (2014).
63. Kiani, R. & Shadlen, M. N. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759–764 (2009).
64. Sanders, J. I., Hangya, B. & Kepecs, A. Signatures of a statistical computation in the human sense of confidence. *Neuron* **90**, 499–506 (2016).
65. Hangya, B., Sanders, J. I. & Kepecs, A. A mathematical framework for statistical decision confidence. *Neural Comput.* **28**, 1840–1858 (2016).
66. Adler, W. T. & Ma, W. J. Limitations of proposed signatures of Bayesian confidence. *Neural Comput.* **30**, 3327–3354 (2018).
67. Li, H.-H. & Ma, W. J. Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. *Nat. Commun.* **11**, 2004 (2020).
68. Geurts, L. S., Cooke, J. R. H., van Bergen, R. S. & Jehee, J. F. M. Subjective confidence reflects representation of Bayesian probability in cortex. *Nat. Hum. Behav.* **6**, 294–305 (2022).
69. De Martino, B., Fleming, S. M., Garrett, N. & Dolan, R. Confidence in value-based choice. *Nat. Neurosci.* **16**, 105–110 (2013).
70. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429 (2002).
71. Fetsch, C. R., Pouget, A., DeAngelis, G. C. & Angelaki, D. E. Neural correlates of reliability-based cue weighting during multisensory integration. *Nat. Neurosci.* **15**, 146 (2012).
72. Qamar, A. T. et al. Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proc. Natl Acad. Sci. USA* **110**, 20332–20337 (2013).
73. Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9**, 1432 (2006).
74. Orban, G., Berkes, P., Fiser, J. & Lengyel, M. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron* **92**, 530–543 (2016).
75. van Bergen, R. S., Ji Ma, W., Pratte, M. S. & Jehee, J. F. M. Sensory uncertainty decoded from visual cortex predicts behavior. *Nat. Neurosci.* **18**, 1728–1730 (2015).
76. Henaff, O. J., Boundy-Singer, Z. M., Meding, K., Ziemba, C. M. & Goris, R. L. T. Representation of visual uncertainty through neural gain variability. *Nat. Commun.* **11**, 2513 (2020).
77. Walker, E. Y., Cotton, R. J., Ma, W. J. & Tolias, A. S. A neural basis of probabilistic computation in visual cortex. *Nat. Neurosci.* **23**, 122–129 (2020).
78. Festa, D., Aschner, A., Davila, A., Kohn, A. & Coen-Cagli, R. Neuronal variability reflects probabilistic inference tuned to natural image statistics. *Nat. Commun.* **12**, 3635 (2021).
79. Allen, M. et al. Unexpected arousal modulates the influence of sensory noise on confidence. *eLife* **5**, e18103 (2016).
80. Maniscalco, B., McCurdy, L. Y., Odegaard, B. & Lau, H. Limited cognitive resources explain a trade-off between perceptual and metacognitive vigilance. *J. Neurosci.* **37**, 1213–1224 (2017).
81. Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M. & Lau, H. Confidence leak in perceptual decision making. *Psychol. Sci.* **26**, 1664–1680 (2015).
82. Fleming, S. M. Action-specific disruption of perceptual confidence. *Psychol. Sci.* **26**, 89–98 (2015).
83. Palmer, J., Huk, A. C. & Shadlen, M. N. The effect of stimulus strength on the speed and accuracy of a perceptual decision. *J. Vis.* **5**, 1 (2005).
84. Hanks, T. D., Mazurek, M. E., Kiani, R., Hopp, E. & Shadlen, M. N. Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *J. Neurosci.* **31**, 6339–6352 (2011).
85. Kiani, R., Corthell, L. & Shadlen, M. N. Choice certainty is informed by both evidence and decision time. *Neuron* **84**, 1329–1342 (2014).
86. Zylberberg, A., Fetsch, C. R. & Shadlen, M. N. The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *eLife* **5**, e17688 (2016).
87. Geisler, W. S. in *The Visual Neurosciences*, Vol. 10 (eds Chalupa, L. & Werne, J.) 825–837 (MIT Press, 2003).
88. Weiss, Y., Simoncelli, E. P. & Adelson, E. H. Motion illusions as optimal percepts. *Nat. Neurosci.* **5**, 598 (2002).
89. Persaud, N., McLeod, P. & Cowey, A. Post-decision wagering objectively measures awareness. *Nat. Neurosci.* **10**, 257–261 (2007).
90. Dienes, Z. & Seth, A. Gambling on the unconscious: a comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Conscious. Cogn.* **19**, 674–681 (2010).
91. Murad, Z., Sefton, M. & Starmer, C. How do risk attitudes affect measured confidence? *J. Risk Uncertain.* **52**, 21–46 (2016).
92. Wichmann, F. A. & Hill, N. J. The psychometric function: I. Fitting, sampling, and goodness of fit. *Percept. Psychophys.* **63**, 1293–1313 (2001).

Acknowledgements

We thank the creators and contributors to the Confidence Database and J. Navajas for making their data available. This work was supported by the US National Science Foundation (Graduate Research Fellowship to Z.M.B.-S.), the US National Institutes of Health (grant nos. T32 EY021462 and K99 EY032102 to C.M.Z., and EY032999 to R.L.T.G.), and the Whitehall Foundation (grant no. UTA19-000535 to R.L.T.G.). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

Z.M.B.-S., C.M.Z. and R.L.T.G. conceived the study, developed the theory, performed the simulations, analysed the data, and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-022-01464-x.

Correspondence and requests for materials should be addressed to Robbe L. T. Goris.

Peer review information *Nature Human Behaviour* thanks Dobromir Rahnev, Rachel Denison, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

nature portfolio

Corresponding author(s): Robbe Goris

Last updated by author(s): Aug 30, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used by the authors of this study to collect data.
Data analysis	We used MATLAB 2019B to analyze all data. The code supporting the findings of this study and a software package implementing the CASANDRE model is publicly available (https://github.com/gorislab/CASANDRE.git). In addition, in supplementary section 8 we used MATLAB analysis code from Shekhar & Rahnev 2021 (https://osf.io/s8fnb/). This code implements their grid-search algorithm.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

This study generated no new data. The data used in this study are available from the Confidence Database (available at: https://osf.io/s46pr/).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	All studies analyzed in this manuscript were previously published or are publicly available in a repository. Data are quantitative and include information about decision choice, decision confidence, reaction time, and stimulus parameters.
Research sample	Each dataset has a different sample of adult participants. The details of each dataset used in this paper can be found in the original publications associated with each dataset.
Sampling strategy	Details of the sampling strategy of individual datasets can be found in the original publications associated with each dataset. Generally the main selection criterion for subjects was that the participant was over 18 years old and had normal or corrected to normal vision.
Data collection	Details of the data collection strategy of individual datasets can be found in the original publications associated with each dataset. No study employed blinding. The apparatus used to collect data varied based on the nature of the study.
Timing	Information about when data were collected is present in read-me files associated with each dataset available on the Confidence Database OSF website.
Data exclusions	Detailed information about data exclusion are reported in the manuscript's methods section. In two analyses (see Methods: Construct validity analyses) we excluded datasets if the subject did not use the highest confidence reporting level at least 10 times during the session. This reduced the number of possible sessions to analyze from 84 to 43.
Non-participation	Details about non-participation for each individual dataset can be found in the original publication associated with each dataset
Randomization	Details about randomization for each individual dataset can be found in the original publication associated with each dataset. In general, studies analyzed in this manuscript did not involve multiple groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
□	☒ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Each dataset has a different sample of adult participants. The details of each dataset used in this paper can be found in the original publications associated with each dataset or read-me file available on the Confidence Database.
Recruitment	Information about the recruitment of study participants can be found in the original publications associated with each dataset or read-me file available on the Confidence Database.
Ethics oversight	Each dataset was approved by a corresponding IRB committee that is identified in the orignal publication assosiated with each dataset.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

nature human behaviour

Article

https://doi.org/10.1038/s41562-022-01464-x

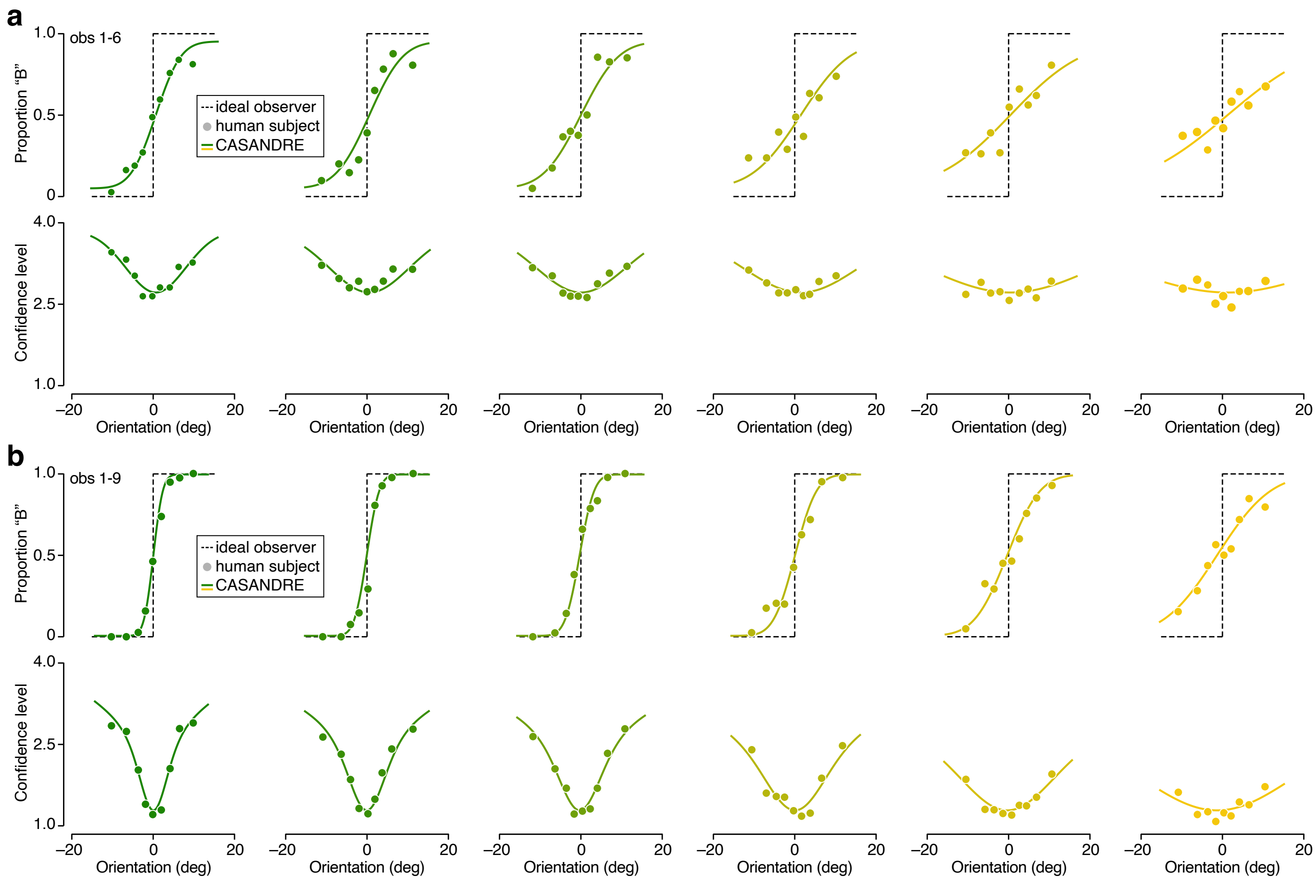
Confidence reflects a noisy decision reliability estimate

Supplementary Information

Table of Contents

1 Adler and Ma (2018), task 1

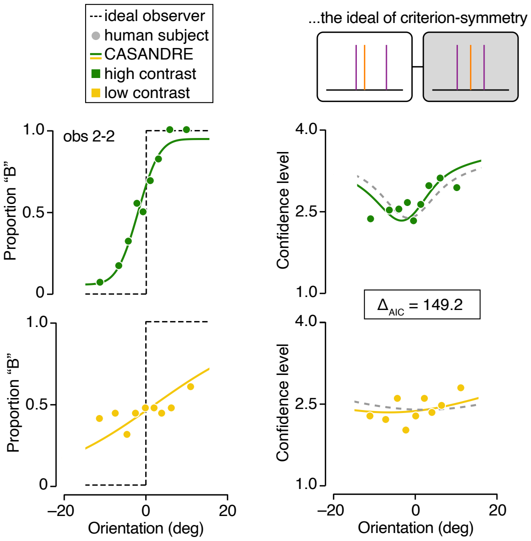
Figure 1b,c shows data from two subjects who performed a perceptual 2-AFC categorization task and additionally reported their confidence using a four-point rating scale. Data were collected by Adler and Ma (2018). Supplementary Figure 1a,b illustrates model fits by plotting the psychometric function (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 by stimulus contrast (columns, contrast decreases from left to right), for one example subject (observer 6 in experiment 1 from ref.[22]). 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 (observer 9 in experiment 1 from ref.[22]).

2 Evaluating the ideal of criterion-symmetry

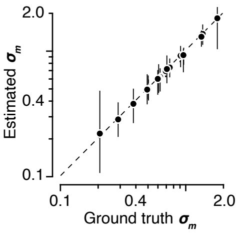
Figure 3a,b shows data for one example subject (2-7) from Adler and Ma (2018) fit with different model variants. This example subject's data were marginally better fit with a model with asymmetrical rather than symmetric confidence criteria (AIC difference = 10.9). Supplementary Figure 2 illustrates data from another example subject (2-2) for whom the difference in model performance is more substantial (AIC difference = 149.2).



Supplementary Figure 2 Following the same conventions as Figure 3. Left: proportion of "Category B" choices is plotted against stimulus orientation for high contrast (top, green) and low contrast (bottom, yellow) for example subject (observer 2 in experiment 2 from ref.[22]). Right: Mean confidence level is plotted as a function of stimulus orientation for the same example observer. Symbols indicate data; lines indicate model fits. Solid lines indicate CASANDRE model fit with asymmetric confidence criteria. Dashed lines indicate CASANDRE model fit with symmetric confidence criteria.

3 Meta-uncertainty recovery, Adler and Ma (2018) task 1

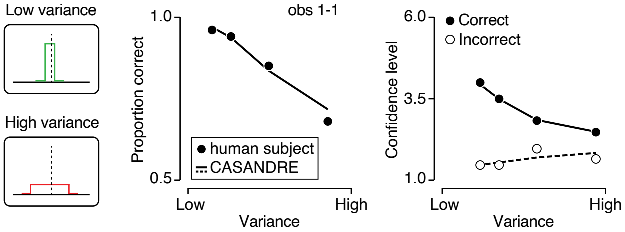
We verified that meta-uncertainty can be reliably recovered for the datasets used to evaluate the model's architecture. These datasets came from the 19 subjects who participated in ref.[22]'s experiment 1 and 2. For each subject, we simulated 100 choice-confidence datasets using the CASANDRE model and the best-fitting parameter values. Each simulated experiment exactly matched the set of trials completed by the subject. We then analyzed the synthetic data in the same manner as the real data. As can be seen in Supplementary Figure 3, under this experimental design, meta-uncertainty is recoverable.



Supplementary Figure 3 Model recovery analysis for Adler and Ma (2018) task 1 data. The median estimate of meta-uncertainty is plotted against the ground truth value for each subject. Error bars illustrate the interquartile range (IQR) computed from 100 simulated data sets.

4 Navajas et al. (2017), experiment 1

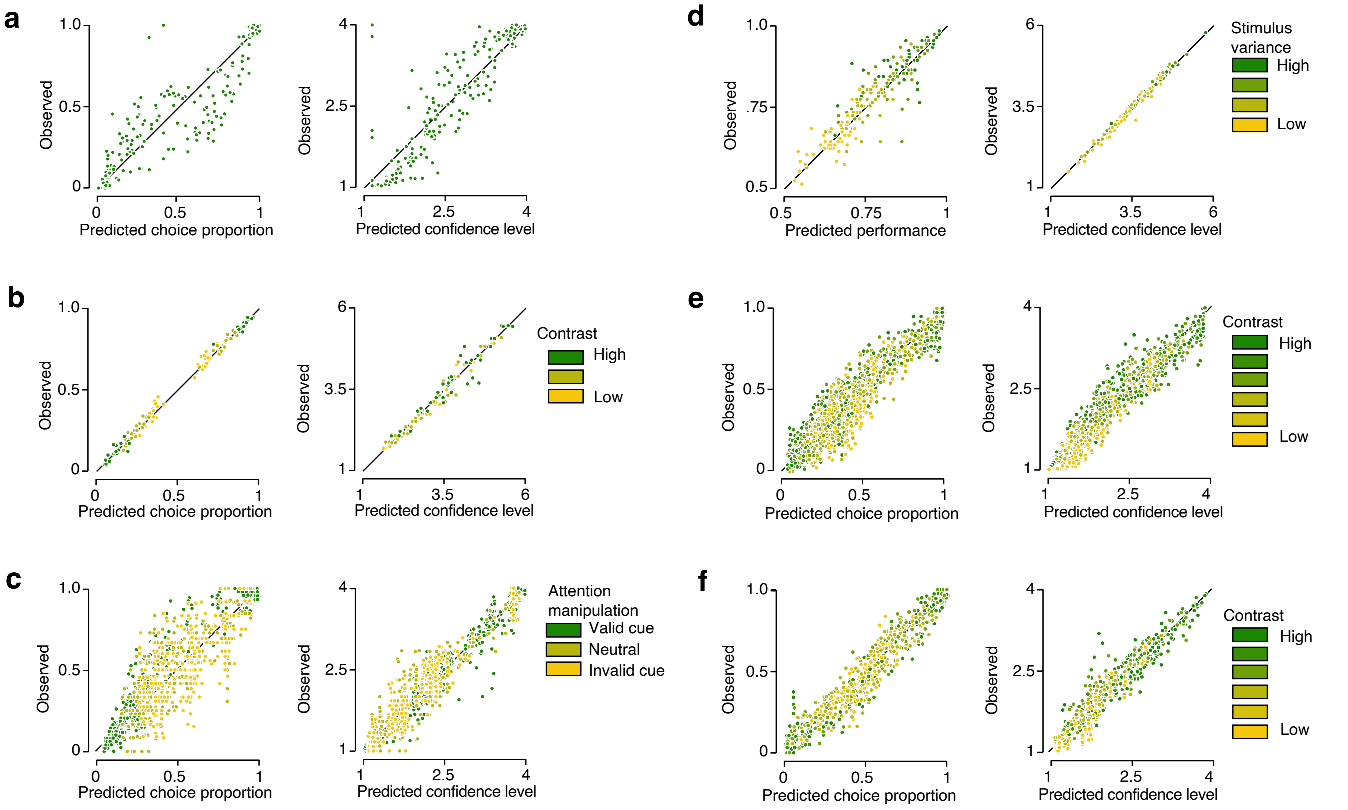
Figure 5a-d shows an analysis of data from subjects who performed either a perceptual or cognitive 2-AFC categorization task and additionally reported their confidence using a six-point rating scale. Data were collected by Navajas et al. (2017). Supplementary Figure 4 illustrates the model fit for an example subject by plotting the data in the format used in the original publication[32]. Proportion correct and confidence level are plotted against stimulus variance. The confidence reports are split 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 4 Model fit for an example subject from Navajas et al. (2017) (observer 1 in experiment 1). 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 by decision accuracy. Symbols summarize observed choice behavior, the full line shows the fit of a two-stage process model of decision-making.

5 Goodness-of-fit across datasets

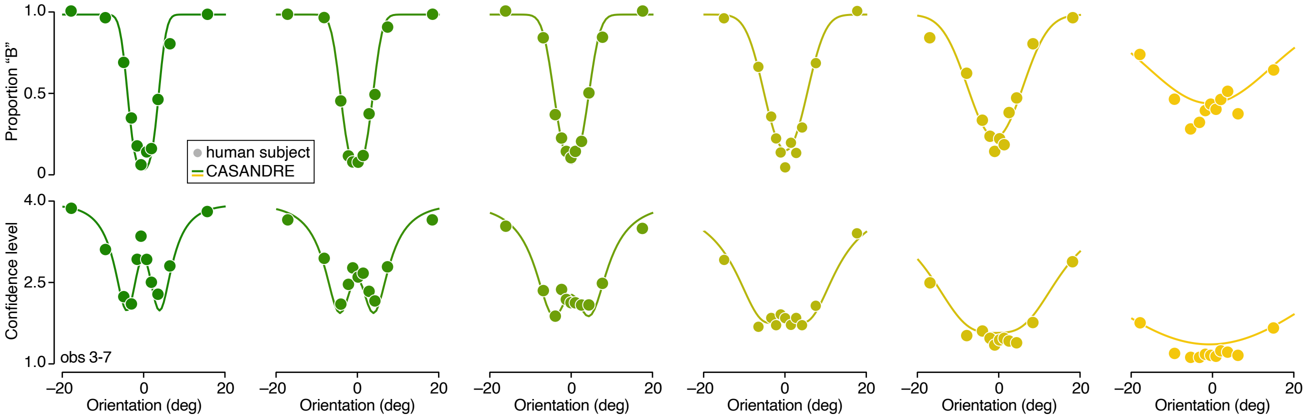
Supplementary Figure 5 shows a comparison of the model predicted and observed choice behavior and confidence reports for the six tasks included in figure 6c.



Supplementary Figure 5 Each symbol summarizes a single stimulus condition for a single subject. Color indicates stimulus reliability. Left: Observed versus predicted choice behavior. Right: Observed versus predicted confidence level. (**a**) Data from Rausch et al. (2020): 25 subjects. (**b**) Shekhar and Rahnev (2021): 20 subjects. **c**) Denison et al. (2018): 12 subjects. **d**) Navajas et al. (2017): 50 subjects; **e**) Adler and Ma (2018) task 2: 34 subjects; **f**) Adler and Ma (2018) task 1: 19 subjects;

6 Adler and Ma (2018), task 2

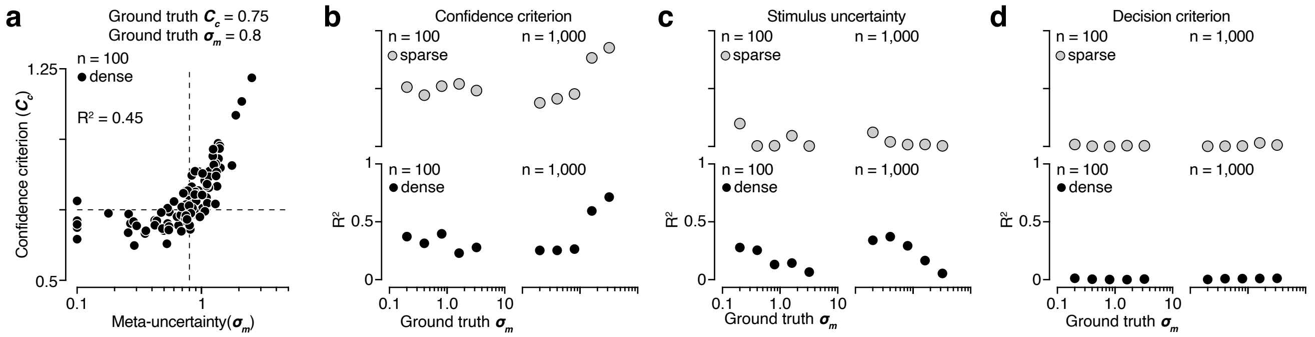
Figure 6c includes a data-point for task 2 from Adler and Ma (2018) and one for Denison et al. (2018). Both studies employed a task design in which subjects discriminated two categories of orientation distributions with the same mean but different standard deviations. Supplementary Figure 6 illustrates an example model fit for this task 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. (Note: while the model fit illustrated here employs asymmetric confidence criteria, all fits in Figure 6c were with symmetric confidence criteria for consistency.)



Supplementary Figure 6 Model fits for an example subject from Adler and Ma (2018) (observer 7 in experiment 3). 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 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 the CASANDRE model.

7 Parameter trade-offs

Figure 4 illustrates a recovery analysis for the meta-uncertainty parameter of the CASANDRE model. Supplementary Figure 7 illustrates an additional analysis of the trade-off between meta-uncertainty and the other parameters of the CASANDRE model using the same generated data and model fits as in Figure 4c. Although the variance in meta-uncertainty explained by trade-offs with confidence criterion can reach high levels, this is somewhat mitigated by denser stimulus sampling (Supplementary Fig. 7b, bottom) and is reasonable for datasets with a larger number of trials (Supplementary Fig. 7b, right) and for values of meta-uncertainty that are empirically observed more often (less than 1).



Supplementary Figure 7 Trade-off between meta-uncertainty and other CASANDRE model parameters. (**a**) Parameter correlation for an example condition. Recovered meta-uncertainty and confidence criterion are plotted against each other for 100 model-generated datasets. The dashed lines represent the ground truth values for confidence criterion ($C_c = 0.75$) and meta-uncertainty ($\sigma_m = 0.8$). Each symbol represents one dataset generated with 100 trials and a dense stimulus sampling regime. (**b**) Trade-off between meta-uncertainty and confidence criterion. (**c**) Trade-off between meta-uncertainty and stimulus uncertainty. (**d** Trade-off between meta-uncertainty and decision criterion.)

8 Comparison with the "log-normal meta-noise" model

Shekhar and Rahnev recently described a hierarchical process model of confidence with desirable properties that dissociate a parameter capturing metacognitive ability from stimulus sensitivity and confidence reporting strategy[25]. They refer to the parameter capturing metacognitive ability as "meta-noise" and find that log-normally distributed meta-noise provides a better quantitative and qualitative match to empirical data than normally distributed noise. In the CASANDRE model, the standard deviation of a log-normal distribution also serves as a metric for the metacognitive ability of an observer, however these two uses of log-normal noise, like the models themselves, are not equivalent. In the CASANDRE model, the confidence variable is distributed according to the ratio of a normally and log-normally distributed variable, whereas in the model of Shekhar and Rahnev the confidence variable has a normal distribution identical to the decision variable but the positions of the confidence criteria are subject to log-normally distributed noise.

We quantitatively compared the "log-normal meta-noise" model of Shekhar and Rahnev with the CASANDRE model. Because the log-normal meta-noise model is currently limited to experiments with two stimulus strengths, we did not apply it to data from ref.[22] (as we did for comparing other model variants in Fig. 3), but instead fit the CASANDRE model to the data reported in their original paper[25]. For purposes of quantitative comparison to the log-normal meta-noise model, we fit the CASANDRE model with asymmetric confidence criteria (yielding 15 total parameters, see Methods). First, we compared the CASANDRE model to the log-normal meta-noise model as described by Shekhar and Rahnev in the main paper, with a different set of confidence criteria for each of three contrast values (yielding 35 total parameters). The CASANDRE model marginally outperformed this model variant (median difference in AIC = 7.8; Supplementary Fig. 8a, top), likely indicating that multiple sets of confidence criteria are an unnecessary feature of the log-normal meta-noise model. Second, we compared the CASANDRE model to a simpler variant of the log-normal meta-noise model (described in the supplement of ref.[25]) with only one set of confidence criteria (yielding 15 total parameters). There was no difference in performance between the CASANDRE model and this variant of the log-normal meta-noise model (median difference in AIC = -0.1; Supplementary Fig. 8a, bottom). Note: we discovered an incorrect scaling of likelihood values in the original code accompanying ref.[25]. We fixed this scaling and thus the AIC values used in this model comparison for the log-normal meta-noise model differ from those reported in ref.[25].

Shekhar and Rahnev demonstrated that the level of meta-noise can serve as a measure of metacognitive ability uncontaminated by stimulus sensitivity or confidence reporting strategy[25]. For comparison, we performed the same analysis using the CASANDRE model. Following their procedure, we mapped each subject's continuous confidence reports into five different binary confidence rating scales, biasing confident reports to be more liberal or conservative. For each subject, we fit the CASANDRE model independently to each of these five remapped datasets across the three contrast levels. We removed one subject that in some conditions did not generate a single response to one of the four possible response options. Meta-uncertainty was largely insensitive to both confidence reporting strategy and stimulus sensitivity (Supplementary Fig. 8b; compare to Fig. 11a in ref.[25]). A two-way ANOVA revealed no main effect of confidence criterion (F(4, 18) = 1.97, P = 0.11) or stimulus contrast (F(2, 18) = 0.04, P = 0.96) on meta-uncertainty. Further, the interaction between confidence reporting strategy and stimulus sensitivity was not significant (F(2, 4) = 1.57, P = 0.14). These results along with the model comparison (Supplementary Fig. 8a) demonstrate that the CASANDRE model performs quantitatively as well as the log-normal meta-noise in explaining the data reported in ref.[25]. We now turn to several more qualitative considerations that differentiate the CASANDRE model and the log-normal meta-noise model.

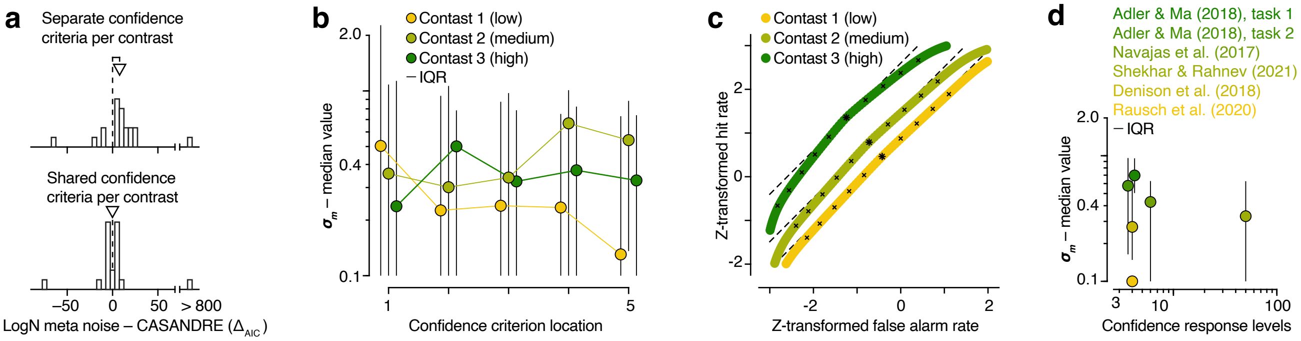
First, Shekhar and Rahnev focus on a particular qualitative metric and show that empirical, averaged zROC functions have significant curvature compared with the straight zROC functions predicted by signal detection theory (their Fig. 4 and 5b). The log-normal meta-noise model shows curved zROC functions but, as the authors note, they resemble piecewise linear functions rather than the smoothly curving zROC functions of the empirical data (see their Fig. 11, bottom left). The CASANDRE model generates smoothly curving averaged zROC functions that more closely resemble the empirically estimated zROC curves (Supplementary Fig. 8c; compare with Fig. 4 and 5b in ref.[25]).

Second, the CASANDRE model is more general and can be applied to experiments that vary stimulus strength in addition to stimulus uncertainty (such as ref.[22]), whereas the log-normal meta-noise model is limited to experiments with two stimulus strengths. This is because the log-normal meta-noise model makes the assumption that confidence is measured in units of d', but does not specify the computation that transforms units of stimulus to units of d'.

Third, the CASANDRE model specifies this confidence computation and posits that it is exactly noise in this transformation that can lead to limited metacognitive ability. Analogous to stimulus discrimination ability being limited by variation in the estimation of the stimulus, the CASANDRE model posits that metacognition is limited by variation in the estimation of the uncertainty required to compute confidence. In contrast, the log-normal meta-noise model uses trial-to-trial variation in confidence criteria. Stochastic confidence criteria cause problems for model tractability, allowing for them to cross both the decision criterion and each other. By casting criteria meta-noise as log-normally distributed, Shekhar and Rahnev avoid the problem of crossovers with the decision criterion, but to solve the problem of crossovers between confidence criteria they make the assumption that noise is perfectly correlated across criteria. Perfect correlation among noisy processes seems an implausible cognitive mechanism, and instead suggests a more parsimonious explanation would be a computation that shifts the space that confidence criteria are defined in. The CASANDRE model naturally avoids both issues of criteria noise through proposing the mechanism of noisy uncertainty normalization.

Fourth, the process captured by the CASANDRE model leads to new predictions about how metacognitive ability can be experimentally manipulated. Figure 6 illustrates how increasing the number of uncertainty levels in a task increases meta-uncertainty. If instead the mechanism maintaining stable confidence criteria is noisy, increasing the number of confidence response levels on the rating scale used by subjects should increase the demands on this mechanism and lower metacognitive ability (thus increasing meta-uncertainty estimated from the CASANDRE model). We see no evidence for this prediction when rearranging the estimated meta-uncertainty across tasks according to the number confidence response levels (Supplementary Fig. 8d).

Finally, the CASANDRE model is easier to fit to data given that its parameters can be optimized using standard maximum likelihood estimation procedures, rather than the purpose-built, two-stage parameter search algorithm developed by Shekhar and Rahnev. While ease of use does not imply one model is more correct than another, we hope that a simpler and more generic optimization procedure will allow the CASANDRE model to be more readily adopted and applied to new contexts.



Supplementary Figure 8 Comparison with Shekhar and Rahnev (2021). (**a**) Distribution of the difference in AIC values across 20 subjects for the CASANDRE model compared with the log-normal meta-noise model with a different set of confidence criteria for each of three contrast levels (top) or with one shared set of confidence criteria across all three contrasts (bottom). Positive values indicate evidence favoring the CASANDRE model. Arrows indicate the median of the distribution. Top: z = 1.90, P = 0.056, effect size = 0.43; Bottom: z = 0, P = 1, effect size = 0; Wilcoxon signed-rank test. (**b**) Median meta-uncertainty across 20 subjects estimated independently for each contrast and confidence criterion location. Error bars illustrate the interquartile range (IQR) across subjects. Compare to Fig. 11a in ref.[25]. (**c**) Averaged zROC functions across 20 subjects generated from fits of the CASANDRE model. The location of the decision criterion is indicated by an asterisk, and the location of each confidence criterion is indicated by an x. The dashed lines illustrate the linear zROC functions predicted from signal detection theory. Compare to Fig. 4, 5b, and 10 in ref.[25]. (**d**) Median level of meta-uncertainty plotted against number of confidence response levels for six confidence experiments. Error bars illustrate the interquartile range (IQR) across subjects. Note the symbol representing ref.[25] is plotted at 50 although subjects rated their confidence on a continuous scale ranging from 50-100.

9 Comparison with other metrics for metacognitive ability

Our method to analyze choice-confidence data is built on the hypothesis that metacognitive ability is determined by meta-uncertainty. It is natural to ask how this metric of metacognitive ability relates to alternatives. We approach this question in two ways. First, by investigating how much meta-uncertainty and other metrics differ 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. 5a-d).

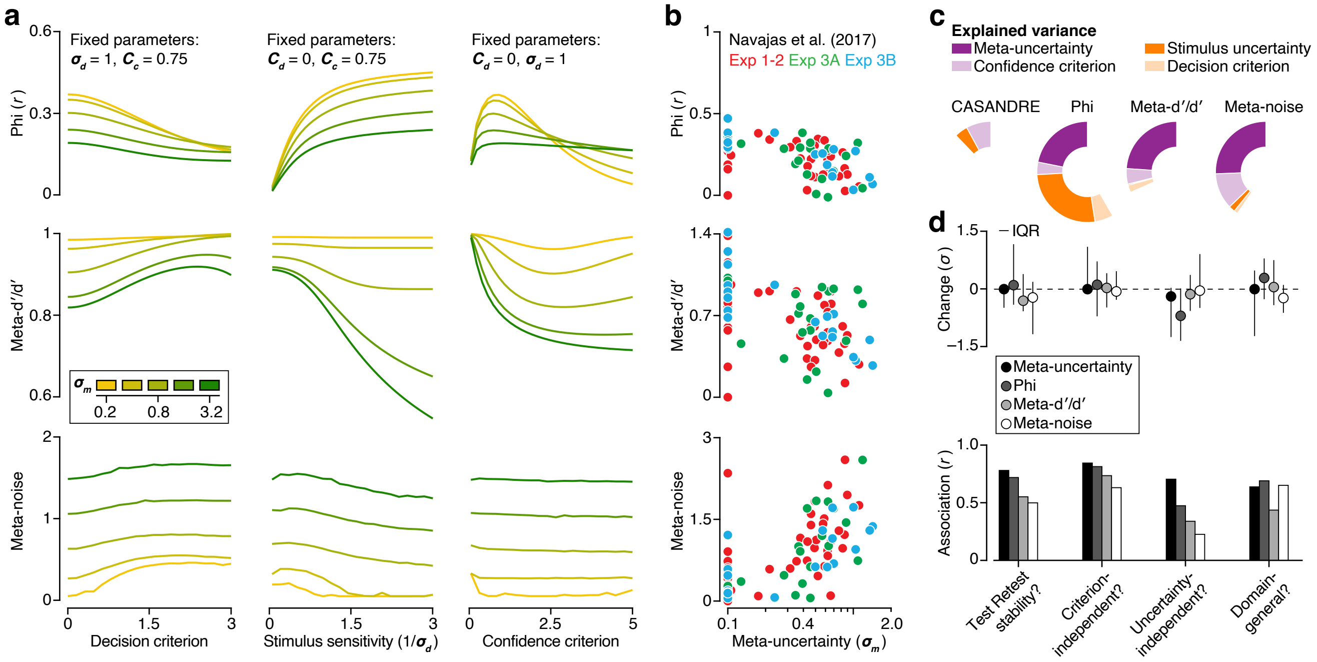
One historically popular approach to quantify metacognitive ability consists of measuring the trial-by-trial correlation between choice accuracy and the confidence report (this metric is sometimes termed "phi")[12]. Consider an analysis of the choice-confidence reports of five model subjects who differed 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 (Supplementary Fig. 9a, 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[13,20]. 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[20]. Under the CASANDRE model, the meta-d'/d' metric does not provide a direct measurement of meta-uncertainty, but instead reflects a complex mixture of model components (Supplementary Fig. 9a, middle).

Finally, a recently introduced model of confidence judgments models metacognitive inefficiencies as perfectly correlated cross-trial variability in the confidence criteria[25]. For experiments involving only two levels of stimulus strength, this criteria noise (hereafter termed "meta-noise") can be estimated by fitting this model to choice-confidence data[25]. In a practical sense, meta-noise resembles meta-uncertainty in that it solely impacts confidence reports. However, different assumptions result in metrics that behave somewhat differently (Supplementary Fig. 9a, bottom).

Now consider the relationship between these metrics and meta-uncertainty for the three experiments performed by Navajas et al. (2017)[32]. Meta-uncertainty estimates and phi are well correlated across experimental sessions ($r(82)$ = –0.60, P < 0.001, Spearman correlation, Supplementary Fig. 9b, top). But the correlation of two other model components with phi also reaches statistical significance: stimulus uncertainty ($r(82)$ = –0.59, P < 0.001) and the confidence criterion ($r(82)$ = –0.24, P = 0.028). Likewise, meta-uncertainty and meta-d'/d' are well correlated ($r(82)$ = –0.52, P < 0.001, Supplementary Fig. 9b, middle). But the confidence criterion is also correlated with meta-d'/d' ($r(82)$ = –0.29, P = 0.008). Finally, meta-uncertainty and meta-noise are well correlated ($r(82)$ = 0.64, P < 0.001, Supplementary Fig. 9b, bottom). But the confidence criterion is also correlated with meta-noise ($r(82)$ = 0.37, P < 0.001). Across experimental sessions and subjects, variability in each of these metrics of metacognitive ability thus in part reflects variability in meta-uncertainty, and in part variability in other components of the CASANDRE model.

To identify the relative importance of the different model components, we again used the experiments from Navajas et al. (2017) and decomposed the variance of these metrics using the averaging-over-orderings technique (see Methods)[93,94]. We first asked whether variability in meta-uncertainty could be explained by other model components, but found this not be be the case (fraction of explained variance: 13%, Supplementary Fig. 9c). In contrast, variability in phi is predominantly explained by stimulus uncertainty (27%), followed by meta-uncertainty (22%). For meta-d'/d' and meta-noise, most variance is explained by meta-uncertainty (24% and 26%) while the contribution of the other model components is rather small (Supplementary Fig. 9c). In summary, for all three alternative metrics, about three quarters of the variance arises from factors other than meta-uncertainty.

Our analysis suggest that phi, meta-d'/d', and, to a lesser extent, meta-noise do not isolate the factors that determine meta-uncertainty but instead measure a complex mixture of factors underlying choice-confidence data. We wondered how the performance of these metrics compares to that of meta-uncertainty in bench-marking experiments. We computed phi, meta-d'/d', and meta-noise for the data sets shown in Fig. 5a-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 four metrics (Supplementary Fig. 9d, top). We then asked for each test whether the different metrics were correlated across both test conditions. Correlations ranged from weak to strong levels (Supplementary Fig. 9d, bottom), with three tests failing to reach statistical significance (uncertainty independence of meta-noise: $r(41)$ = 0.23, P = 0.145; test-retest reliability of meta-noise: $r(12)$ = 0.50, P = 0.072; and domain generality of meta-d'/d': $r(18)$ = 0.44, P = 0.056). Overall, meta-uncertainty compared favorably to the alternative metrics. The mean correlation value across the four tests was 0.74 for meta-uncertainty, 0.67 for phi, 0.52 for meta-d'/d', and 0.50 for meta-noise (meta-uncertainty compared with phi, P = 0.21; meta-uncertainty compared with meta-d'/d', P = 0.02; meta-uncertainty compared with meta-noise, P = 0.01; bootstrap test, see methods).



Supplementary Figure 9 Comparing meta-uncertainty with three existing metrics of 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 estimated phi (top), meta-*d*′/*d*′ (middle), and meta-noise (bottom) for each simulated experiment. (**b**) Phi (top), meta-*d*′/*d*′ (middle), and meta-noise (bottom) plotted against meta-uncertainty estimates for three confidence experiments. Each symbol summarizes data from a single session (84 total sessions across 50 subjects, see methods). Meta-uncertainty was limited to a minimum value of 0.1. (**c**) Wedges indicate the proportion of variance in meta-uncertainty (left), phi, meta-*d*′/*d*′*, and meta-noise]* explained by each model component. (**d**) Comparison of the performance of four metrics of metacognitive ability in four bench-marking tests. Top: analysis of estimation bias. Bottom: analysis of estimation robustness. Error bars illustrate the interquartile range (IQR) and center the median across subjects.

Supplementary methods: calculating alternative metacognitive metrics

We probed the relation between meta-uncertainty and three alternative metrics of metacognitive 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 ref.[20]. 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 middle panels of Supplementary Fig. 9a. 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 (Supplementary Fig 9a, top panels). We used these same simulated trials to fit the "log-normal meta-noise model" of Shekhar and Rahnev[25]. We downloaded their parameter optimization code and modified it as appropriate to fit our simulated data (available at https://osf.io/s8fnb/). In their procedure, a nested two-step, coarse-to-fine search algorithm is used to optimize the estimated confidence criteria and the meta-noise level. The resulting meta-noise estimates are plotted in the bottom panels of Supplementary Fig 9a. The non-smooth appearance of the curves is a consequence of instabilities in the fitting procedure.

We also computed these alternative metrics for each session from ref.[32] (see Supplementary Fig. 9b-d). As is conventional, we estimated d' for each stimulus condition from the observed hit and false alarm rates[29]. 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. We again used the fitting procedure of Shekhar and Rahnev[25], estimating decision criterion and four values of d' and optimizing four sets of confidence criteria and the value of meta-noise across the four stimulus conditions (Supplementary Fig. 9b-d).

To compute the proportion of variance in each alternative metric across 84 sessions[32] explained by different components of the CASANDRE model, we used the averaging-over-orderings technique[93,94]. 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 Supplementary Fig. 9c.

To statistically test the differences in correlation of alternative metacognitive ability metrics and meta-uncertainty across our four benchmark tests (Supplementary Fig. 9d) we performed a bootstrap resampling procedure. For each test, we resampled the subjects 10,000 times with replacement and recomputed the correlations. We then computed the mean correlation value across the four tests. We report the proportion of resampled mean correlations that were larger for the alternative metric compared with meta-uncertainty.

Supplementary references

93. William Kruskal. Relative Importance by Averaging over Orderings. *The American Statistician*, 41(1):6–10, February 1987.

94. Ulrike Grömping. Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. *The American Statistician*, 61(2):139–147, May 2007.