PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Comparing Bayesian and non-Bayesian accounts of human confidence reports

William T. Adler[1]*, **Wei Ji Ma**[1,2]

1 Center for Neural Science, New York University, New York, NY, United States of America, **2** Department of Psychology, New York University, New York, NY, United States of America

* will@wtadler.com

OPEN ACCESS

Citation: Adler WT, Ma WJ (2018) Comparing Bayesian and non-Bayesian accounts of human confidence reports. PLoS Comput Biol 14(11): e1006572. https://doi.org/10.1371/journal.pcbi.1006572

Editor: Samuel J. Gershman, Harvard University, UNITED STATES

Received: March 16, 2018

Accepted: October 11, 2018

Published: November 13, 2018

Data Availability Statement: All data and code used for running experiments, model fitting, and plotting is available on a GitHub repository at https://github.com/wtadler/confidence. We have also used Zenodo to assign a DOI to the repository: 10.5281/zenodo.1422804.

Funding: This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1342536 (WTA), https://www.nsfgrfp.org/. The funders had no role in study design, data

Abstract

Humans can meaningfully report their confidence in a perceptual or cognitive decision. It is widely believed that these reports reflect the Bayesian probability that the decision is correct, but this hypothesis has not been rigorously tested against non-Bayesian alternatives. We use two perceptual categorization tasks in which Bayesian confidence reporting requires subjects to take sensory uncertainty into account in a specific way. We find that subjects do take sensory uncertainty into account when reporting confidence, suggesting that brain areas involved in reporting confidence can access low-level representations of sensory uncertainty, a prerequisite of Bayesian inference. However, behavior is not fully consistent with the Bayesian hypothesis and is better described by simple heuristic models that use uncertainty in a non-Bayesian way. Both conclusions are robust to changes in the uncertainty manipulation, task, response modality, model comparison metric, and additional flexibility in the Bayesian model. Our results suggest that adhering to a rational account of confidence behavior may require incorporating implementational constraints.

Author summary

Humans are able to report a sense of confidence in decisions that we make. It is widely hypothesized that confidence reflects the computed probability that a decision is accurate; however, this hypothesis has not been fully explored. We use several human behavioral experiments to test a variety of models that may be considered to be distinct hypotheses about the computational underpinnings of confidence. We find that reported confidence does not appear to reflect the probability that a decision is correct, but instead emerges from a heuristic approximation of this probability.

Introduction

People often have a sense of a level of confidence about their decisions. Such a "feeling of knowing" [1, 2] may serve to improve performance in subsequent decisions [3], learning [1], and group decision-making [4]. Much recent work has focused on identifying brain regions and neural mechanisms responsible for the computation of confidence in humans [5–7],

collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

nonhuman primates [8–10], and rodents [11]. In the search for the neural correlates of confidence, the leading premise has been that confidence is Bayesian, i.e., the observer's estimated probability that a choice is correct [1, 12–14]. In human studies, however, naïve subjects can give a meaningful answer when you ask them to rate their confidence about a decision [15]; thus, "confidence" intrinsically means something to people, and it is not a foregone conclusion that this intrinsic sense corresponds to the Bayesian definition. Therefore, we regard the above "definition" as a testable hypothesis about the way the brain computes explicit confidence reports; we use Bayesian decision theory to formalize this hypothesis.

Bayesian decision theory provides a general and often quantitatively accurate account of perceptual decisions in a wide variety of tasks [16–18]. According to this theory, the decision-maker combines knowledge about the statistical structure of the world with the present sensory input to compute a posterior probability distribution over possible states of the world. In principle, a confidence report might be derived from the same posterior distribution; this is the hypothesis described above, which we will call the Bayesian confidence hypothesis (BCH). The main goal of this paper is to test that hypothesis. Recent studies have attempted to test the BCH [19, 20] but, because of their experimental designs, are unable to meaningfully distinguish the Bayesian model from any other model of confidence.

Recent work has proposed possible qualitative signatures of Bayesian confidence [21]. However, the observation (or lack thereof) of these signatures provides an uncertain amount of evidence in favor of (or against) the Bayesian model, and the signatures are therefore not useful for determining which computations underlie confidence reports [22]. To objectively and quantitatively determine whether confidence ratings appear to be Bayesian, we use a formal model comparison approach. We test the predictions of the BCH as we vary the quality of the sensory evidence and the task structure within individuals. We compare Bayesian models against a variety of alternative models, something that is important for the epistemological standing of Bayesian claims [23, 24]. We find that the BCH qualitatively describes human behavior but that quantitatively, even the most flexible Bayesian model is outperformed by models that take uncertainty into account in a non-Bayesian way.

Results

Experiment 1

During each session, each subject completed two orientation categorization tasks, Tasks A and B. On each trial, a category C was selected randomly (both categories were equally probable), and a stimulus s was drawn from the corresponding stimulus distribution and displayed. The subject categorized the stimulus and simultaneously reported their confidence on a 4-point scale, with a single button press (Fig 1a). Using a single button press for choice and confidence may prevent post-choice influences on the confidence judgment ([25], but see [26]) and emphasized that confidence should reflect the observer's perception rather than a preceding motor response. The categories were defined by normal distributions on orientation, which differed by task (Fig 1b). In Task A, the distributions had different means ($\pm\mu_C$) and the same standard deviation (σ_C); leftward-tilting stimuli were more likely to be from category 1. Variants of Task A are common in decision-making studies [27]. In Task B, the distributions had the same mean (0˚) and different standard deviations (σ_1, σ_2); stimuli around the horizontal were more likely to be from category 1. Variants of Task B are less common [28–30] but have some properties of perceptual organization tasks; for example, a subject may have to detect when a stimulus belongs to a narrow category (e.g., in which two line segments are collinear) that is embedded in a a broader category (e.g., in which two line segments are unrelated).

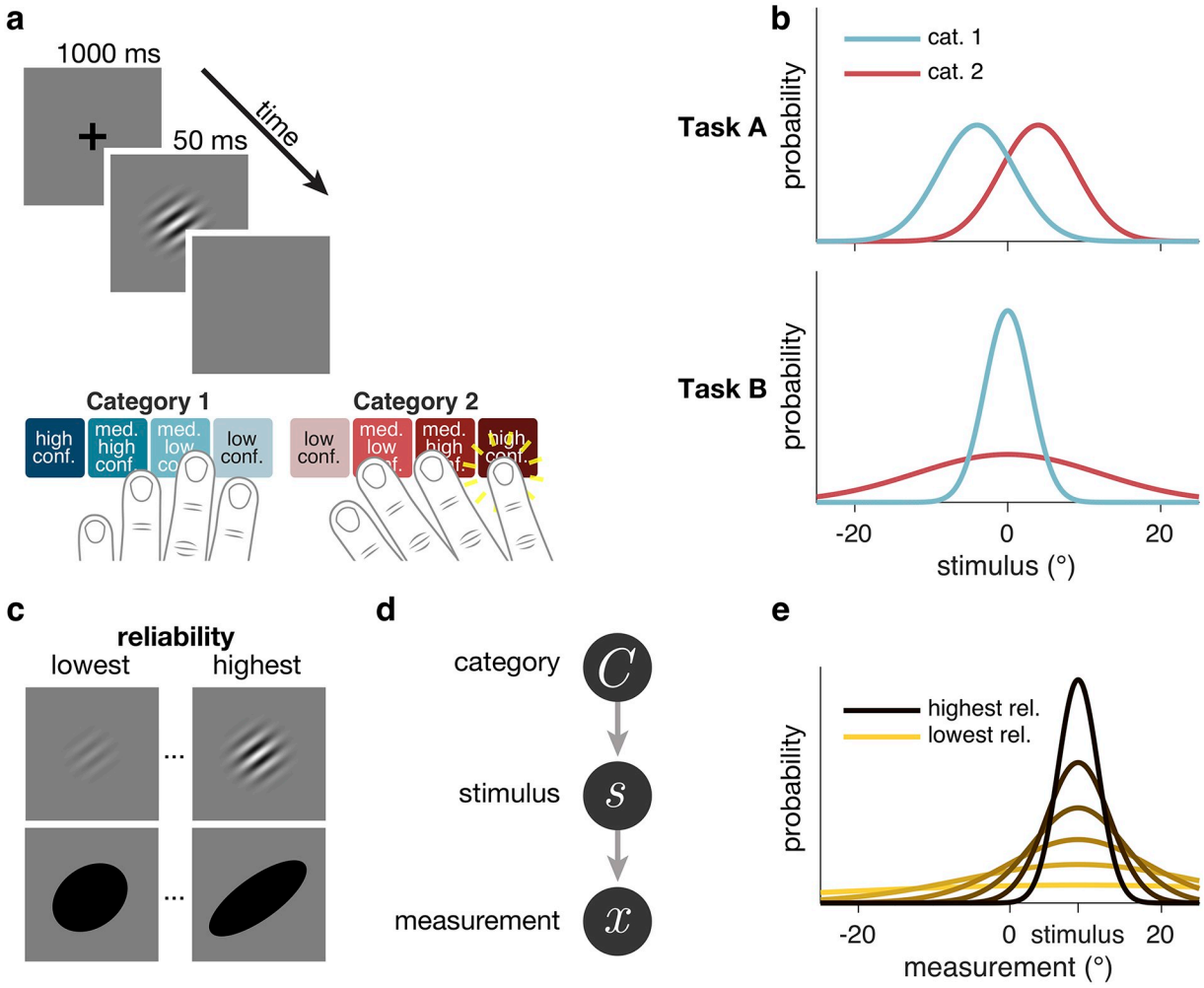

Fig 1. Task design. (**a**) Schematic of a test block trial. After stimulus offset, subjects reported category and confidence level with a single button press. (**b**) Stimulus distributions for Tasks A and B. (**c**) Examples of low and high reliability stimuli. Six (out of eleven) subjects saw drifting Gabors, and five subjects saw ellipses. (**d**) Generative model. (**e**) Example measurement distributions at different reliability levels. In all models (except Linear Neural), the measurement is assumed to be drawn from a Gaussian distribution centered on the true stimulus, with s.d. dependent on reliability.

https://doi.org/10.1371/journal.pcbi.1006572.g001

Subjects were highly trained on the categories; during training, we only used highest-reliability stimuli, and we provided trial-to-trial category correctness feedback. Subjects were then tested with 6 different reliability levels, which were chosen randomly on each trial. During testing, correctness feedback was withheld to avoid the possibility that confidence simply reflects a learned mapping between stimulus orientation and reliability and the probability of being correct [30–33].

Because we are interested in subjects' intrinsic computation of confidence, we did not instruct or incentivize them to assign probability ranges to each button (e.g., by using a scoring rule [34–36]). If we had, we would have essentially been training subjects to use a specific model of confidence.

To ensure that our results were independent of stimulus type, we used two kinds of stimuli. Some subjects saw oriented drifting Gabors; for these subjects, stimulus reliability was manipulated through contrast. Other subjects saw oriented ellipses; for these subjects, stimulus

reliability was manipulated through ellipse elongation (Fig 1c). We found no major differences in model rankings between Gabor and ellipse subjects, therefore we will make no distinctions between the groups.

For modeling purposes, we assume that the observer's internal representation of the stimulus is a noisy measurement x, drawn from a Gaussian distribution with mean s and s.d. σ (Fig 1d and 1e). In the model, σ (i.e., uncertainty) is a fitted function of stimulus reliability.

Bayesian model

A Bayes-optimal observer uses knowledge of the generative model to make a decision that maximizes the probability of being correct. Here, when the measurement on a given trial is x, this strategy amounts to choosing the category C for which the posterior probability $p(C \mid x)$ is highest. This is equivalent to reporting category 1 when the log posterior ratio, $d = \log\frac{p(C=1|x)}{p(C=2|x)}$, is positive.

In Task A, d is $d_A = \frac{2x\mu_C}{\sigma^2+\sigma_C^2}$. Therefore, the ideal observer reports category 1 when x is positive; this is the structure of many psychophysical tasks [37]. In Task B, however, d is $d_B = \frac{1}{2}\log\frac{\sigma^2+\sigma_2^2}{\sigma^2+\sigma_1^2} - \frac{\sigma_2^2-\sigma_1^2}{2(\sigma^2+\sigma_1^2)(\sigma^2+\sigma_2^2)}x^2$; the observer needs both x and σ in order to make an optimal decision.

From the point of view of the observer, σ is the trial-to-trial level of sensory uncertainty associated with the measurement [38]. In a minor variation of the optimal observer, we allow for the possibility that the observer's prior belief over category, $p(C)$, is different from the true value of (0.5, 0.5); this adds a constant to d_A and d_B.

We introduce the Bayesian confidence hypothesis (BCH), stating that confidence reports depend on the internal representation of the stimulus (here x) only via d. In the BCH, the observer chooses a response by comparing d to a set of category and confidence boundaries. For example, whenever d falls within a certain range, the observer presses the "medium-low confidence, category 2" button. The BCH is thus an extension of the choice model described above, wherein the value of d is used to compute confidence as well as chosen category. There is another way of thinking about this. Bayesian models assume that subjects compute d in order to make an optimal choice. Assuming people compute d at all, are they able to use it to report confidence as well? We refer to the Bayesian model here as simply "Bayes." We also tested several more constrained versions of this model.

The observer's decision can be summarized as a mapping from a combination of a measurement and an uncertainty level (x, σ) to a response that indicates both category and confidence. We can visualize this mapping as in Fig 2, first column. It is clear that the pattern of decision boundaries in the BCH is qualitatively very different between Task A and Task B. In Task A, the decision boundaries are quadratic functions of uncertainty; confidence decreases monotonically with uncertainty and increases with the distance of the measurement from 0. In Task B, the decision boundaries are neither linear nor quadratic.

Alternative models

At first glance, it seems obvious that sensory uncertainty is relevant to the computation of confidence. However, this is by no means a given; in fact, a prominent proposal is that confidence is based on the distance between the measurement and the decision boundary, without any role for sensory uncertainty [10, 11, 39]. Therefore, we tested a model (Fixed) in which the response is a function of the measurement alone (equivalent to a maximum likelihood estimate of the stimulus orientation), and not of the uncertainty of that measurement (Fig 2, second column).

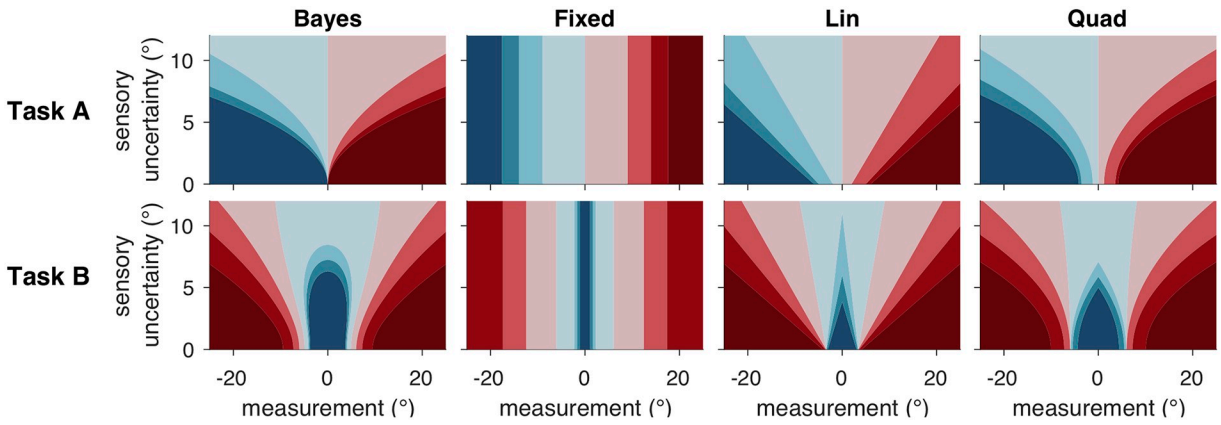

Fig 2. Decision rules/mappings in four models. Each model corresponds to a different mapping from a measurement and uncertainty level to a category and confidence response. Colors correspond to category and confidence response, as in Fig 1a. Plots were generated using parameter values that were roughly similar to those found after fitting subject data but were chosen primarily to illustrate the different features of the models.

https://doi.org/10.1371/journal.pcbi.1006572.g002

We also tested heuristic models in which the subject uses their knowledge of their sensory uncertainty but does not compute a posterior distribution over category. We have previously classified such models as *probabilistic non-Bayesian* [40]. In the Orientation Estimation model, subjects base their response on a maximum a posteriori estimate of orientation (rather than category), using the mixture of the two stimulus distributions as a prior distribution. In the Linear Neural model, subjects base their response on a linear function of the output of a hypothetical population of neurons.

We derived two additional probabilistic non-Bayesian models, Lin and Quad, from the observation that the Bayesian decision criteria are an approximately linear function of uncertainty in some measurement regimes and approximately quadratic in others. These models are able to produce approximately Bayesian behavior without actually performing any computation of the posterior. In Lin and Quad, subjects base their response on a linear or a quadratic function of x and σ, respectively. A comparison of the Lin and Quad columns to the Bayes column in Fig 2 demonstrates that Lin and Quad can approximate the Bayesian mapping from (x, σ) to response despite not being based on the Bayesian decision variable. All of the models we tested were variants of the six models described so far (Bayes, Fixed, Orientation Estimation, Linear Neural, Lin, Quad).

Each trial consists of the experimentally determined orientation and reliability level and the subject's category and confidence response (an integer between 1 and 8). This is a very rich data set, which we summarize in Fig 3. We find the following effects: performance and confidence increase as a function of reliability (Fig 3a, 3b, 3h and 3i), and high-confidence reports are less frequent than low-confidence reports (Fig 3e and 3f). Note Fig 3c and 3d especially; this is the projection of the data that we will use to demonstrate model fits for the rest of this paper. We use this projection because the vertical axis (mean button press) most closely approximates the form of the raw data. Additionally, because our models are differentiated by how they use uncertainty, it is informative to plot how response changes as a function of reliability, in addition to category and task.

Recently, a measure of the degree of association between accuracy and confidence, meta-d', has been developed [41, 42]. While it can be useful for characterizing individual differences, we do not include it in our analyses or display it in Fig 3. That is because one strength of our experimental design is that we parametrically vary stimulus strength and stimulus reliability;

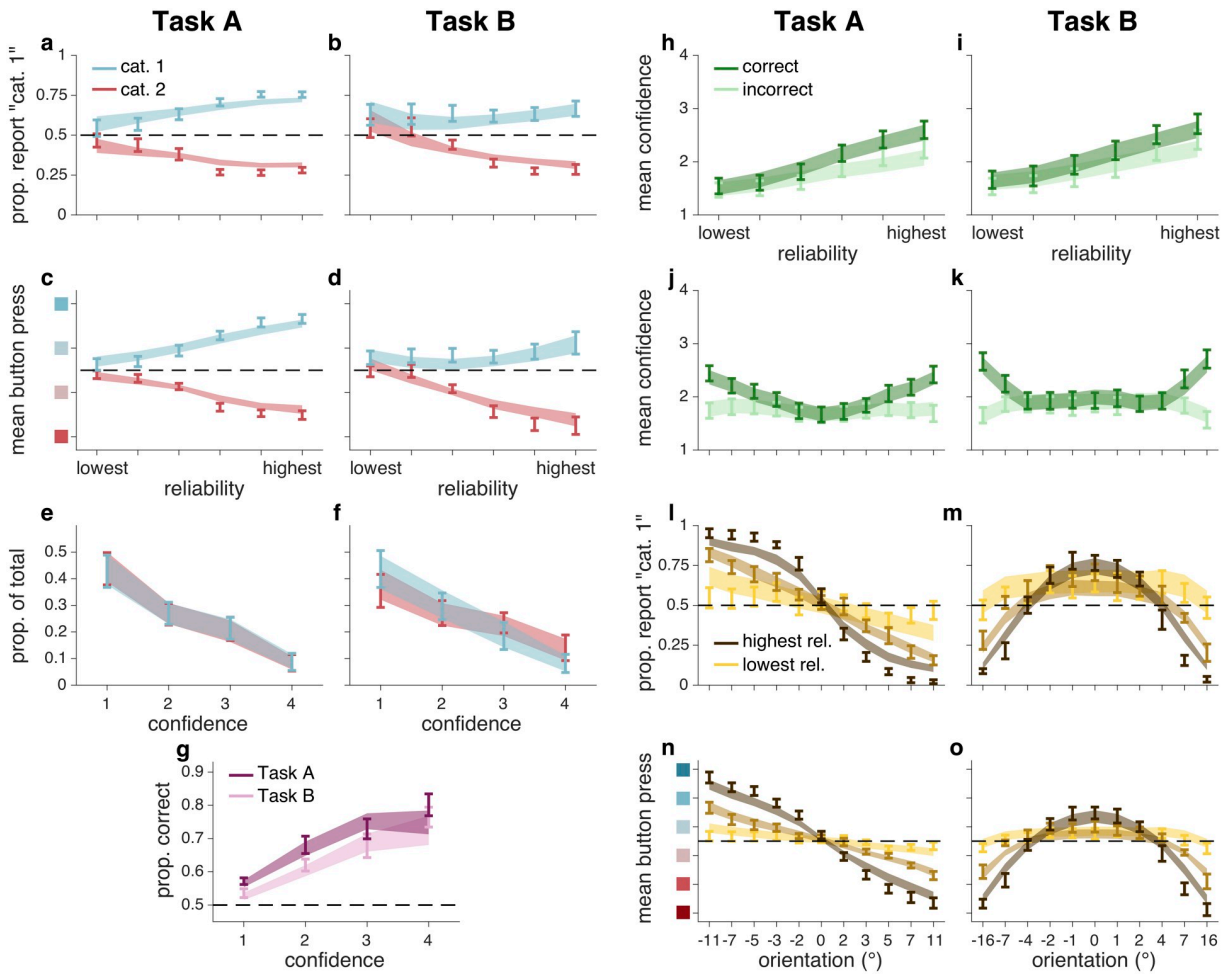

Fig 3. Behavioral data and fits from best model (Quad), experiment 1. Error bars represent ±1 s.e.m. across 11 subjects. Shaded regions represent ±1 s.e.m. on model fits. (**a,b**) Proportion "category 1" reports as a function of stimulus reliability and true category. (**c,d**) Mean button press as a function of stimulus reliability and true category. (**e,f**) Normalized histogram of confidence reports for both true categories. (**g**) Proportion correct category reports as a function of confidence report and task. (**h,i**) Mean confidence as a function of stimulus reliability and correctness. (**j,k**) Mean confidence as a function of stimulus orientation and reliability. (**l,m**) Proportion "category 1" reports as a function of stimulus orientation and reliability. (**n,o**) Mean button press as a function of stimulus orientation and reliability. (**c,d,n,o**) Vertical axis label colors correspond to button presses, as in Fig 1a. (**l–o**) For clarity, only 3 of 6 reliability levels are shown, although models were fit to all reliability levels.

https://doi.org/10.1371/journal.pcbi.1006572.g003

this differs from papers in which meta-d' plays a central role because, in those papers, the stimulus is often only a binary category.

Model comparison

We used Markov Chain Monte Carlo (MCMC) sampling to fit models to raw individual-subject data. To account for overfitting, we compared models using leave-one-out cross-validated log likelihood scores (LOO) computed with the full posteriors obtained through MCMC [43]. A model recovery analysis ensured that our models are meaningfully distinguishable (Methods). Unless otherwise noted, models were fit jointly to Task A and B category and confidence responses.

Use of sensory uncertainty. We first compared Bayes to the Fixed model, in which the observer does not take trial-to-trial sensory uncertainty into account (Fig 4). Fixed provides a

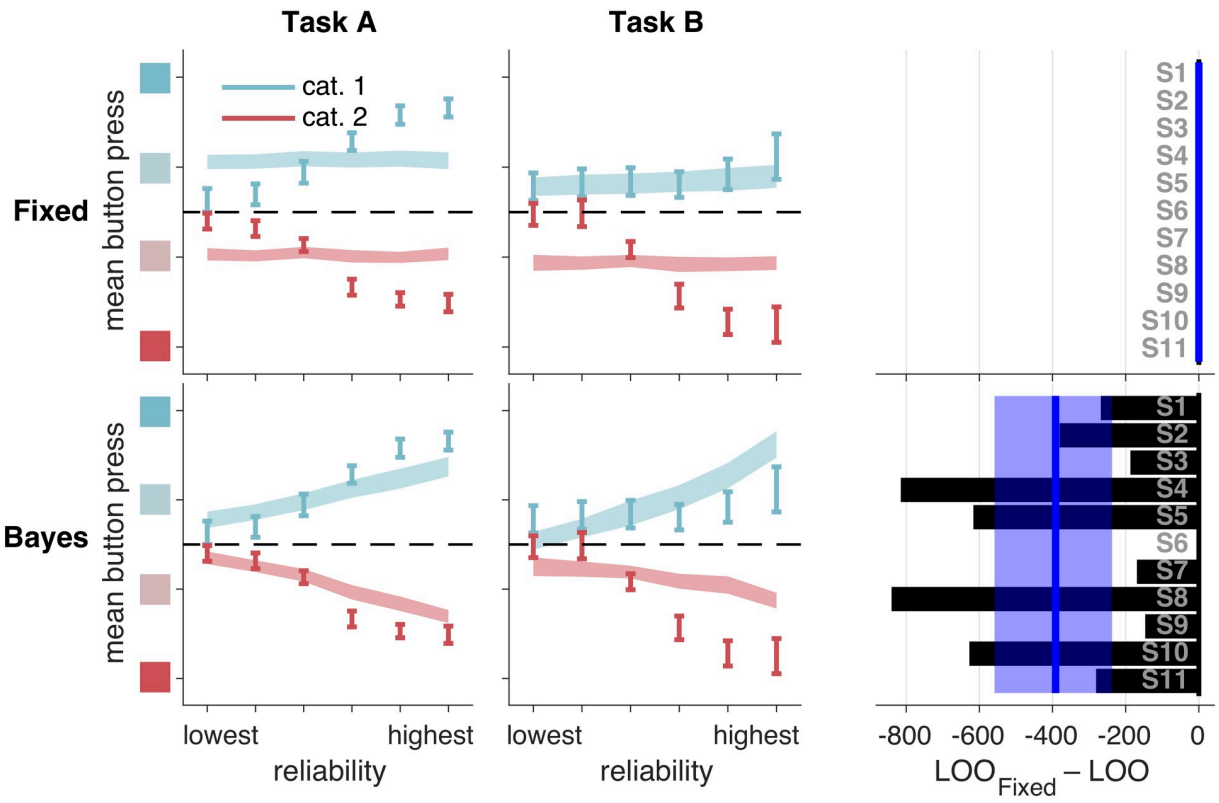

Fig 4. Model fits and model comparison for models Fixed and Bayes. Bayes provides a better fit, but both models have large deviations from the data. Left and middle columns: model fits to mean button press as a function of reliability, true category, and task. Error bars represent ±1 s.e.m. across 11 subjects. Shaded regions represent ±1 s.e.m. on model fits, with each model on a separate row. Right column: LOO model comparison. Bars represent individual subject LOO scores for Bayes, relative to Fixed. Negative (leftward) values indicate that, for that subject, Bayes had a higher (better) LOO score than Fixed. Blue lines and shaded regions represent, respectively, medians and 95% CI of bootstrapped mean LOO differences across subjects. These values are equal to the summed LOO differences reported in the text divided by the number of subjects. Although we plot data as a function of the true category here, the model only takes in measurement and reliability as an input; it is not free to treat individual trials from each true category differently.

https://doi.org/10.1371/journal.pcbi.1006572.g004

poor fit to the data, indicating that observers use not only a point estimate of their measurement, but also their uncertainty about that measurement. Bayes outperforms Fixed by a summed LOO difference (median and 95% CI of bootstrapped sums across subjects) of 2265 [498, 4253]. For the rest of this paper, we will report model comparison results using this format.

Although Bayes fits better than Fixed, it still shows systematic deviations from the data, especially at high reliabilities. (Because we fit our models to all of the raw data and because boundary parameters are shared across all reliability levels, the fit to high-reliability trials is constrained by the fit to low-reliability trials).

Noisy log posterior ratio. To see if we could improve Bayes's fit, we tried a version that included decision noise, i.e. noise on the log posterior ratio d. We assumed that this noise takes the form of additive zero-mean Gaussian noise with s.d. σ_d. This is almost equivalent to the probability of a response being a logistic (softmax) function of d [44]. Adding d noise improves the Bayesian model fit by 804 [510, 1134] (S1 Table).

For the rest of the reported fits to behavior, we will only consider this version of Bayes with d noise, and will refer to this model as Bayes-dN. We will refer to Bayes-dN, Fixed, Orientation

Estimation, Linear Neural, Lin, and Quad, when fitted jointly to category and confidence data from Tasks A and B, as our core models.

Heuristic models. Orientation Estimation performs worse than Bayes-*d*N by 2041 [385, 3623] (Fig 5, second row). The intuition for one way that this model fails is as follows: at low levels of reliability, the MAP estimate is heavily influenced by the prior and tends to be very close to the prior mean (0˚). This explains why, in Task B, there is a bias towards reporting "high confidence, category 1" at low reliability. Linear Neural performs about as well as Bayes-*d*N, with summed LOO differences of 1188 [-588, 2704], and the fits to the summary statistics are qualitatively poor (Fig 5, third row).

Finally, Lin and Quad outperform Bayes-*d*N by 1398 [571, 2644] and 1167 [858, 2698], respectively. Both models provide qualitatively better fits, especially at high reliabilities (compare Fig 5, first row, to Fig 5, fourth and fifth rows), and strongly tilted orientations (compare S10n and S10o Fig to S14n and S14o Fig and Fig 3n and 3o).

We summarize the performance of our core models in Fig 6. Noting that a LOO difference of more than 5 is considered to be very strong evidence [45], the heuristic models Lin and Quad perform much better than Bayes-*d*N. Furthermore, we can decisively rule out Fixed. We will now describe variants of our core models.

Non-parametric relationship between reliability and σ. One potential criticism of our fitting procedure is that we assumed a parameterized relationship between reliability and σ. To see if our results were dependent on that assumption, we modified the models such that σ was non-parametric (i.e., there was a free parameter for σ at each level of reliability). With this feature added to our core models, Quad still fits better than Bayes-*d*N by 1676 [839, 2730] and it fits better than Fixed by 6097 [4323, 7901] (S1 Table). This feature improved Quad's performance by 325 [141, 535]. For the rest of this paper, we will only report the fits of Bayes-*d*N, the best-fitting non-Bayesian model, and Fixed. See supplementary figures and tables for all other model fits.

Incorrect assumptions about the generative model. Suboptimal behavior can be produced by optimal inference using incorrect generative models, a phenomenon known as "model mismatch" [46–48]. Up to now, Bayes-*d*N has assumed that observers have accurate knowledge of the parameters of the generative model. To test whether this assumption prevents Bayes-*d*N from fitting the data well, we tested a series of Bayesian models in which the observer has inaccurate knowledge of the generative model.

Bayes-*d*N assumed that, because subjects were well trained, they knew the true values of σ_C, σ_1, and σ_2, the standard deviations of the stimulus distributions. We tested a model in which these values were free parameters, rather than fixed to the true value. We would expect these free parameters to improve the fit of Bayes-*d*N in the case where subjects were not trained enough to sufficiently learn the stimulus distributions. This feature improves Bayes-*d*N's fit by 908 [318, 1661], but it still underperforms Quad by 768 [399, 1144] (S1 Table).

Previous models also assumed that subjects had accurate knowledge of their own measurement noise; their perceptual uncertainty, used in the computation of d, was identical to their measurement noise, used to generate measurements x, with both uncertainty and measurement noise equal to σ. We tested models in which we fit $\sigma_{\text{measurement}}$ and $\sigma_{\text{inference}}$ as two independent functions of reliability [46]. This feature improves Bayes-*d*N's fit by 1310 [580, 2175], but it still underperforms Quad by 362 [162, 602] (S1 Table).

Weighted average of precision and perceived probability of being correct. A recent paper ([49]; although see [22]) proposed that confidence is a weighted average of a function of variance, such as $\frac{1}{\sigma^2}$, and the perceived probability of being correct (incidentally, under a non-Bayesian decision rule). We tested such a model (using a Bayesian decision rule), which fits

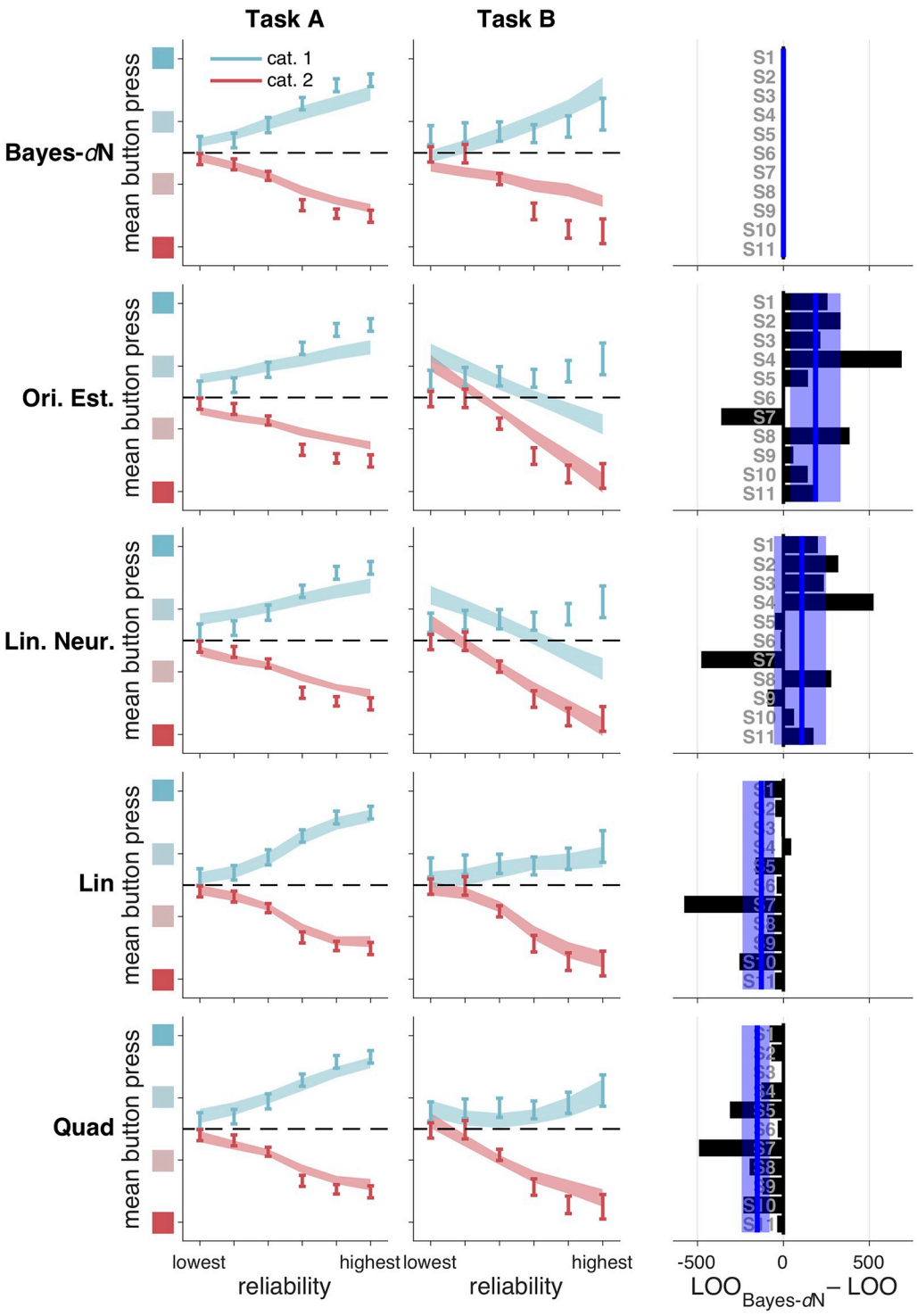

Fig 5. Model fits and model comparison for Bayes-*d*N and heuristic models. In both tasks, Bayes-*d*N fails to describe the data at high reliabilities; Lin and Quad provides a good fit at most reliabilities. Left and middle columns: as in Fig 4. Right column: bars represent individual subject LOO scores for each model, relative to Bayes-*d*N. Negative (leftward) values indicate that, for that subject, the model in the corresponding row had a higher (better) LOO score than Bayes-*d*N. Blue lines and shaded regions: as in Fig 4.

https://doi.org/10.1371/journal.pcbi.1006572.g005

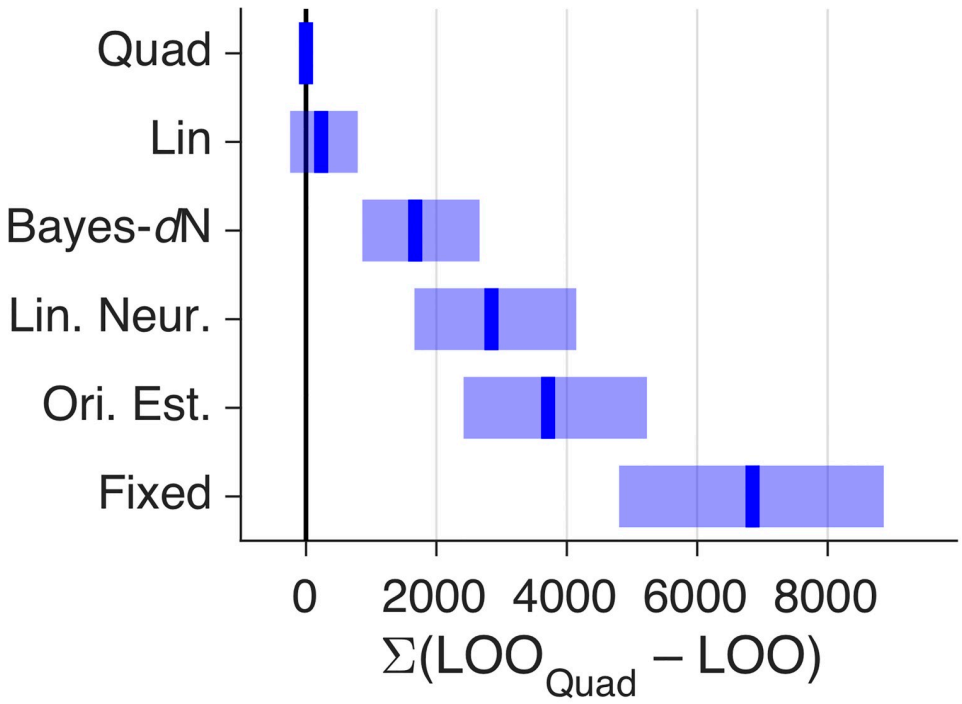

Fig 6. Comparison of core models, experiment 1. Models were fit jointly to Task A and B category and confidence responses. Blue lines and shaded regions represent, respectively, medians and 95% CI of bootstrapped summed LOO differences across subjects. LOO differences for these and other models are shown in S1a Fig.

https://doi.org/10.1371/journal.pcbi.1006572.g006

better than Fixed by 3059 [758, 5528] but still underperforms Lin by 3478 [2211, 5020] (S1 Table).

Separate fits to Tasks A and B. In order to determine whether model rankings were primarily due to differences in one of the two tasks, we fit our models to each task individually. In Task A, Quad fits better than Bayes-*d*N by 581 [278, 938], and better than Fixed by 3534 [2529, 4552] (S2 Fig and S2 Table). In Task B, Quad fits better than Bayes-*d*N by 978 [406, 1756] and fits better than Fixed by 3234 [2099, 4390] (S3 Fig and S3 Table). [41]

Fits to category choice data only. In order to see whether our results were peculiar to combined category and confidence responses, we fit our models to the category choices only. Lin fits better than Bayes-*d*N by 595 [311, 927] and fits better than Fixed by 1690 [976, 2534] (S4 Fig and S4 Table).

Fits to Task B only, with noise parameters fitted from Task A. To confirm that the fitted values of sensory uncertainty in the probabilistic models are meaningful, we treated Task A as an independent experiment to measure subjects' sensory noise. The category choice data from Task A can be used to determine the four uncertainty parameters. We fit Fixed with a decision boundary of 0˚ (equivalent to a Bayesian choice model with no prior), using maximum likelihood estimation. We fixed these parameters and used them to fit our models to Task B category and confidence responses. Lin fits better than Bayes-*d*N by 1773 [451, 2845] and fits better than Fixed by 5016 [3090, 6727] (S5 Fig and S5 Table).

Separate category and confidence responses (experiment 2). There has been some recent debate as to whether it is more appropriate to collect choice and confidence with a single motor response (as described above) or with separate responses [20, 25, 50, 51]. Aitchison et al. [19] found that confidence appears more Bayesian when subjects use separate responses.

To confirm this, we ran a second experiment in which subjects chose a category by pressing one of two buttons, then reported confidence by pressing one of four buttons. Aitchison et al. [19] also provided correctness feedback on every trial; in order to ensure that we could compare our results to theirs, we also provided correctness feedback in this experiment, even though this manipulation was not of primary interest. After fitting our core models, our results did not differ substantially from experiment 1: Lin fits better than Bayes-*d*N by 396 [186, 622] and fits better than Fixed by 2095 [1344, 2889] (S6 Fig and S6 Table).

Task B only (experiment 3). It is possible that subjects behave suboptimally when they have to do multiple tasks in a session; in other words, perhaps one task "corrupts" the other. To explore this possibility, we ran an an experiment in which subjects completed Task B only. We chose Task B over Task A for this experiment because Task B has the desirable characteristic that uncertainty is required for optimal categorization. Quad fits better than Bayes-*d*N by 1361 [777, 2022] and fits better than Fixed by 7326 [4905, 9955] (S7 Fig and S7 Table). In experiments 2 and 3, subjects only saw drifting Gabors; we did not use ellipses.

We also fit only the choice data, and found that Lin fits about as well as Bayes-*d*N, with summed LOO differences of 117 [-76, 436] and fits better than Fixed by 1084 [619, 1675] (S8 Fig and S8 Table). This approximately replicates our previously published results [30].

Model comparison metric. None of our model comparison results depend on our choice of metric: in all three experiments, model rankings changed negligibly if we used AIC, BIC, AICc, or WAIC instead of LOO.

Discussion

Although people can report subjective feelings of confidence, the computations that produce this feeling are not well understood. It has been proposed that confidence is the observer's computed posterior probability that a decision is correct [1, 12–14]. However, this hypothesis has not been fully tested. We carried out a strong test of human confidence reports, using overlapping categories [52], withholding feedback on testing trials, and varying experimental components such as task, stimulus type, and stimulus reliability [32]. We used model comparison to investigate the computational underpinnings of confidence, fitting a total of 75 models from 6 distinct model families.

Our first finding is that, like the optimal observer, subjects use knowledge of their sensory uncertainty when reporting confidence in a categorical decision; models in which the observer ignores their sensory uncertainty provide a poor fit to the data (Fig 4). Our second finding is that subjects do not appear to use knowledge of their sensory uncertainty in a way that is fully consistent with the Bayesian confidence hypothesis. Instead, heuristic models that approximate Bayesian computation—but do not compute a posterior probability over category—outperform the Bayesian models in two tasks (Fig 5, compare top row to bottom two rows). This result continued to hold after we relaxed assumptions about the relationship between reliability and noise, and about the subject's knowledge of the generating model. We accounted for the fact that our models had different amounts of flexibility by using a wide array of model comparison metrics and by showing that our models are meaningfully distinguishable.

Limitations

Our study has several limitations. For instance, because of our short presentation time, we cannot say much about how our results generalize to tasks that require integration of evidence over time [8, 53–55]. Additionally, because our stimuli are very low-level, we cannot say much about high-level stimuli like faces [56]. Also, we only considered explicit confidence ratings, which differ from the implicit confidence that can be gathered from humans

(e.g., by presenting two tasks and asking the subject to choose which one they feel more confident about completing correctly [57, 58]) or from nonhuman animals [13] (e.g., by measuring how frequently they decline to make a difficult choice [8], or how long they will wait for a reward [11]). It is possible that implicit confidence might be more Bayesian than explicit confidence; Barthelmé and Mamassian [58] conduct an implicit confidence experiment and rule out some heuristic models. However, their experimental task is substantially different from the one presented here. In their experiment, the stimulus feature of interest (orientation) only takes on two values rather than varying parametrically, so it requires a different class of heuristic models. Future studies of the difference between implicit and explicit confidence should use experiments that are able to distinguish the models presented here, which has not been done.

Other investigations of deviations from Bayesian confidence

Like the present study, Aitchison et al. [19] found evidence that confidence reports may emerge from heuristic computations. However, they sampled stimuli from only a small region of their two-dimensional space, where model predictions may not vary greatly. Therefore, their stimulus set did not allow for the models to be strongly distinguished. Furthermore, although they tested for *Bayesian* computation, they did not test for *probabilistic* computation (whether observers take sensory uncertainty into account on a trial-to-trial basis [40]) as we do here. Such a test requires that the experimenter vary the reliability, not only the value, of the stimulus feature of interest.

Navajas et al. [49] suggested that confidence reports are best described as a weighted average of precision and the probability of being correct. However, their model uses the estimated probability of being correct under a non-Bayesian decision rule [22]. They did not show the fit of a Bayesian model, and therefore their study does not constitute a true test of whether confidence is Bayesian. Here, we tested and rejected the hypothesis that confidence is a weighted average of precision and the posterior probability of being correct under a Bayesian decision rule.

Sanders et al. [20] reported that confidence has a "statistical" nature. However, their experiment was unable to determine whether confidence is Bayesian or not [17], because the stimuli varied along only one dimension. Aitchison et al. [19] note that, to distinguish models of confidence, the experimenter must use stimuli that are characterized by two dimensions (e.g., contrast and orientation as in this experiment, or contrast and crowding as in Barthelmé and Mamassian [58]). This is because, when fitting models that map from an internal variable to an integer confidence rating, it is impossible to distinguish between two internal variables that are monotonically related (in the case of Sanders et al. [20], the measurement and the posterior probability of being correct). Therefore, the only alternative model proposed by Sanders et al. [20] is based on reaction time, rather than on the presented stimuli.

In detection and coarse discrimination tasks, Lau, Rahnev, and colleagues report that subjects overestimate their confidence in the periphery and for unattended stimuli. The authors have proposed a signal detection theory model in which high eccentricity or lower attention induces higher noise, and the confidence criterion may not change at all [39, 59–63]. As a result, more probability mass will "spill over" the criterion to the high-confidence regime. How do these findings relate to ours? At a qualitative level, they are consistent in that confidence does not seem Bayesian. However, in detection and coarse discrimination tasks, it is not possible to distinguish between fixed-criterion and probabilistic models [64], and their data cannot be used to infer that the criterion is fixed. The paradigms in the present paper are able to distinguish such models because of the parametric manipulation of orientation, the

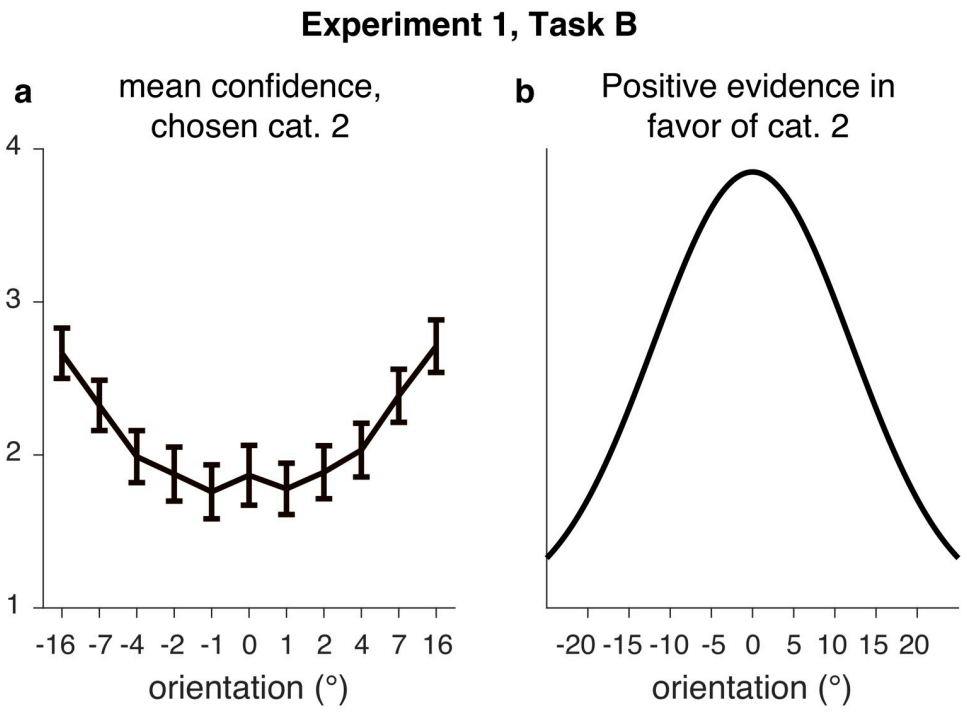

Fig 7. (**a**) In experiment 1, Task B, on trials in which the subject chose category 2, mean confidence increases with the absolute value of stimulus orientation. (**b**) The "positive evidence" in favor of category 2, however, decreases with the absolute value of stimulus orientation. This plot depicts the category-conditioned stimulus distribution $p(s \mid C = 2)$; positive evidence in this experiment is equivalent to the likelihood $p(x \mid C = 2)$, which is just $p(s \mid C = 2)$ convolved with the subject's measurement noise.

https://doi.org/10.1371/journal.pcbi.1006572.g007

stimulus feature of interest; indeed, we find strong evidence against the Fixed criterion model. It remains to be seen whether claims that confidence can be systematically dissociated from perceptual performance [1, 51, 65–69] are consistent with the account presented here, in which the brain adjusts confidence criteria based on uncertainty but in a non-Bayesian manner.

Another form of non-Bayesian confidence ratings is the recent proposal that, in confidence judgments, only the "positive evidence" in favor of the chosen option matters, instead of the "balance of evidence" between two options [31, 53, 56, 70, 71]. In our tasks, this form of suboptimality would entail that confidence is derived from the (log) likelihood of the chosen category, instead of from the (log) likelihood ratio. This does not seem consistent with our data; for example, the likelihood of category 2 decreases as the absolute value of the stimulus and, correspondingly, the measurement, increases (Fig 7b). However, confidence for a category 2 decision steadily increases with the absolute value of the stimulus (Fig 7a). More work is needed to understand whether alternative models could explain the "positive evidence" data, and if not, what causes the difference with our results.

Status of Bayesian models

What do our findings tell us about the neural basis of confidence? Previous studies have found that neural activity in some brain areas (e.g., human medial temporal lobe [7] and prefrontal cortex [72], monkey lateral intraparietal cortex [8] and pulvinar [10], rodent orbitofrontal cortex [11]) is associated with behavioral indicators of confidence, and/or with the distance of a

stimulus to a decision boundary. However, such studies mostly used stimuli that vary along a single dimension (e.g., net retinal dot motion energy, mixture of two odors). Because measurement is indistinguishable from the probability of being correct in these classes of tasks, neural activity associated with confidence may represent either the measurement or the probability of being correct [19]. In addition to the recommendation of Aitchison et al. [19] to distinguish between these possibilities by varying stimuli along two dimensions, we recommend fitting both Bayesian and non-Bayesian probabilistic models to behavior. In view of the relatively poor performance of the Bayesian models in the present study, the proposal [12] to correlate behavior and neural activity with predictions of the Bayesian confidence model should be viewed with skepticism.

Our results raise general issues about the status of Bayesian models as descriptions of behavior. First, because it is impossible to exhaustively test all models that might be considered "Bayesian," we cannot rule out the entire class of models. However, we have tried to alleviate this issue as much as possible by testing a large number of Bayesian models—far more than the number of Bayesian and non-Bayesian models tested in other studies of confidence. Second, Bayesian models are often held in favor for their generalizability; one can determine the performance-maximizing strategy for any task. Although generalizability indeed makes Bayesian models attractive and powerful, we do not believe that this property should override a bad fit.

One could take two different views of our heuristic model results. The first view is that the heuristics should be taken seriously as principled models [73]; here, the challenge is to demonstrate that they describe behavior in a variety of tasks and can be motivated based on underlying principles. The second view is that these are descriptive models simply meant to demonstrate that a simple model can provide a good fit to the data; here, the heuristics are benchmarks for more principled models, and the challenge is to find a principled model that fits the data as well as the heuristics. We lean towards the second view and interpret our results as demonstrating that the purest form of the Bayesian confidence hypothesis does not describe human confidence reports particularly well.

However, one might still conclude, after examining the fits of the Bayesian model, that the behavior is "approximately Bayesian" rather than "non-Bayesian." As written, this is a semantic distinction because it relies on one's definition of "approximate." However, it can be turned into a more meaningful question: Are the differences between human behavior and Bayesian models accounted for by an unknown principle, such as an ecologically relevant objective function that includes both task performance and biological constraints?

Although there are benefits associated with veridical explicit representations of confidence [74–76], there are also neural constraints that may give rise to non-Bayesian behavior [23, 24]. Such constraints include the kinds of operations that neurons can perform, the high energy cost of spiking [77, 78], and the cost of neural wiring length [79, 80]. A search for ecologically rational constraints on Bayesian computation benefits from a positive characterization of the deviations from Bayesian computation, in the form of heuristic models such as Lin and Quad. Specifically, one could define neural networks with various combinations of constraints, and train them as if they were psychophysical subjects in our tasks. After training, one could fit behavioral models to them; this approach has already shown that the output from such neural networks is sometimes best described by heuristic models [81]. Using model ranking as a measure of similarity, one could determine which network architecture and training procedure produces confidence behavior that is most similar to that of humans. This could reveal which constraints are responsible for the specific deviations from Bayesian computation that we have observed.

Methods

Ethics statement

The experiments were approved by the University Committee on Activities Involving Human Subjects of New York University. Informed consent was given by each subject before the experiment.

Experiment 1

Subjects. 11 subjects (2 male), aged 20–42, participated in the experiment. Subjects received $10 per 40-60 minute session, plus a completion bonus of $15. All subjects were naïve to the purpose of the experiment. No subjects were fellow scientists.

Apparatus and stimuli. *Apparatus*. Subjects were seated in a dark room, at a viewing distance of 32 cm from the screen, with their chin in a chinrest. Stimuli were presented on a gamma-corrected 60 Hz 9.7-inch 2048-by-1536 display. The display (LG LP097QX1-SPA2) was the same as that used in the 2013 iPad Air (Apple); we chose it for its high pixel density (264 pixels/inch). The display was connected to a Windows desktop PC using the Psychophysics Toolbox extensions [82, 83] for MATLAB (Mathworks).

Stimuli. The background was mid-level gray (199 cd/m^2). The stimulus was either a drifting Gabor (Subjects 3, 6, 8, 9, 10, and 11) or an ellipse (Subjects 1, 2, 4, 5, and 7). The Gabor had a peak luminance of 398 cd/m^2 at 100% contrast, a spatial frequency of 0.5 cycles per degrees of visual angle (dva), a speed of 6 cycles per second, a Gaussian envelope with a standard deviation of 1.2 dva, and a randomized starting phase. Each ellipse had a total area of 2.4 dva^2, and was black (0.01 cd/m^2). We varied the contrast of the Gabor and the elongation (eccentricity) of the ellipse.

Categories. In Task A, stimulus orientations were drawn from Gaussian distributions with means $\mu_1 = -4^\circ$ (category 1) and $\mu_2 = 4^\circ$ (category 2) and standard deviations $\sigma_1 = \sigma_2 = 5^\circ$. In Task B, stimulus orientations were drawn from Gaussian distributions with means $\mu_1 = \mu_2 = 0^\circ$, and standard deviations $\sigma_1 = 3^\circ$ (category 1) and $\sigma_2 = 12^\circ$ (category 2) (Fig 1b). We chose these category means and standard deviations such that the accuracy of an optimal observer would be around 80%.

Procedure. Each subject completed 5 sessions. Each session consisted of two parts; the subject did Task A in the first part, followed by Task B in the second part, or vice versa (chosen randomly each session). Each part started with instruction and was followed by alternating blocks of 96 category training trials and 144 testing trials, for a total of three blocks of each type, with a block of 24 confidence training trials immediately after the first category training block. Combining all sessions and both tasks, each subject completed 2880 category training trials, 240 confidence training trials, and 4320 testing trials; we did not analyze category training or confidence training trials.

Instruction. At the start of each part of a session, subjects were shown 30 (72 in the first session) exemplar stimuli from each category. Additionally, we provided them with a printed graphic similar to Fig 1b, and explained how the stimuli were generated from distributions. We answered any questions.

Category training. To ensure that subjects knew the stimulus distributions well, we gave them extensive category training. Each trial proceeded as follows (Fig 1a): Subjects fixated on a central cross for 1 s. Category 1 or category 2 was selected with equal probability. The stimulus orientation was drawn from the corresponding stimulus distribution (Fig 1b). Gabors had 100% contrast, and ellipses had 0.95 eccentricity (elongation). The stimulus appeared at fixation for 300 ms, replacing the fixation cross. Subjects were asked to report category 1 or

category 2 by pressing a button with their left or right index finger, respectively. Subjects were able to respond immediately after the offset of the stimulus, at which point verbal correctness feedback was displayed for 1.1 s. The fixation cross then reappeared.

Confidence training. To familiarize subjects with the button mappings, they completed a short confidence training block at the start of every task. We told subjects that in this block, it would be harder to tell what the stimulus orientation was, there would be no correctness feedback, and they would be reporting their confidence on each trial in addition to their category choice. We provided them with a printed graphic similar to the buttons pictured in Fig 1a, indicating that they had to press one of eight buttons to indicate both category choice and confidence level, the latter on a 4-point scale. The confidence levels were labeled as "very high," "somewhat high," "somewhat low," and "very low." Gabors had 0.4%, 0.8%, 1.7%, 3.3%, 6.7%, or 13.5% contrast, and ellipses had 0.15, 0.28, 0.41, 0.54, 0.67, or 0.8 eccentricity, chosen randomly with equal probability on each trial (Fig 1c). Stimuli were only displayed for 50 ms. Trial-to-trial feedback consisted only of a message telling them which category and confidence level they had reported. Other than these changes, the trial procedure was the same as in category training.

Subjects were not instructed to use the full range of confidence reports [20], as that might have biased them away from reporting what felt most natural. Instead, they were simply asked to be "as accurate as possible in reporting their confidence" on each trial.

Testing. The trial procedure in testing blocks was the same as in confidence training blocks, except that trial-to-trial feedback was completely withheld. At the end of each block, subjects were required to take at least a 30 s break. During the break, they were shown the percentage of trials that they had correctly categorized. Subjects were also shown a list of the top 10 block scores (across all subjects, indicated by initials) for the task they had just done. This was intended to motivate subjects to score highly, and to reassure them that their scores were normal, since it is rare to score above 80% on a block.

Descriptive statistics. Since our models do not include any learning effects, we wanted to ensure that task performance was stable. For all tasks and experiments, we found no evidence that performance changed significantly as a function of the number of trials. For each experiment and task (the 5 lines in Fig 8), we fit a logistic regression to the binary correctness data for each subject, obtaining a set of slope coefficients. We then used a t-test to determine whether these sets of coefficients differed significantly from zero. In no group did the slopes differ significantly from zero; across all 5 groupings the minimum p-value was 0.077 (Task

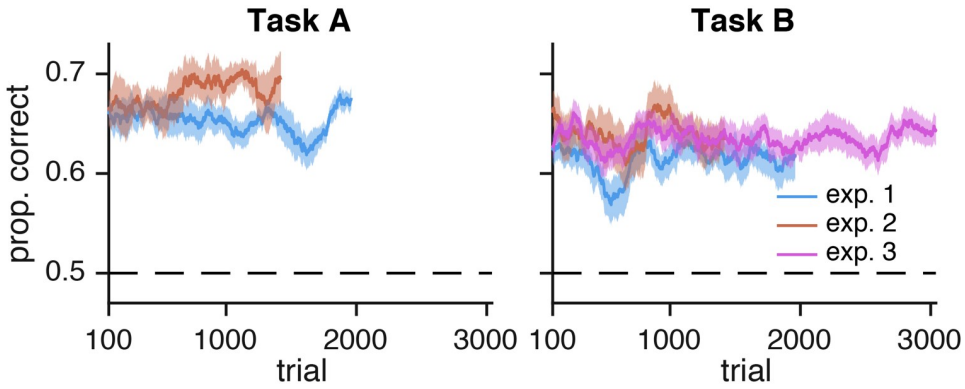

Fig 8. Performance as a function of number of trials, for both tasks and for all experiments. Performance was computed as a moving average over test trials (200 trials wide). Shaded regions represent ±1 s.e.m. over subjects. Performance did not change significantly over the course of each experiment.

https://doi.org/10.1371/journal.pcbi.1006572.g008

A, experiment 2), which would not be significant even before correcting for multiple comparisons.

Experiment 1. The following statistical differences were assessed using repeated-measures ANOVA.

In Task A, there was a significant effect of true category on category choice ($F_{1,10} = 285$, $p < 10^{-7}$). There was no main effect of reliability, which took 6 levels of contrast or ellipse elongation, on category choice ($F_{5,50} = 0.27$, $p = 0.88$). In other words, subjects were not significantly biased to respond with a particular category at low reliabilities. There was a significant interaction between reliability and true category, which is to be expected ($F_{5,50} = 59.6$, $p < 10^{-15}$) (Fig 3a).

In Task B, there was again a significant effect of true category on category choice ($F_{1,10} = 78.3$, $p < 10^{-5}$). There was no main effect of reliability ($F_{5,50} = 2.93$, $p = 0.051$). There was again a significant interaction between reliability and true category ($F_{5,50} = 28$, $p < 10^{-12}$) (Fig 3b).

In Task A, there was a significant effect of true category on response ($F_{1,10} = 136$, $p < 10^{-6}$). There was no main effect of reliability ($F_{5,50} = 0.61$, $p = 0.642$). There was a significant interaction between reliability and true category ($F_{5,50} = 58.7$, $p < 10^{-13}$) (Fig 3c).

In Task B, there was a significant effect of true category on response ($F_{1,10} = 54.2$, $p < 10^{-6}$). There was a significant effect of reliability ($F_{5,50} = 4.84$, $p = 0.0128$). There was a significant interaction between reliability and true category ($F_{5,50} = 29.2$, $p < 10^{-8}$) (Fig 3d).

In Task A, there was a main effect of confidence on the proportion of reports ($F_{3,30} = 7.75$, $p < 10^{-3}$); low-confidence reports were more frequent than high-confidence reports. There was no significant effect of true category ($F_{1,10} = 0.784$, $p = 0.397$) and no interaction between confidence and category on proportion of responses ($F_{3,30} = 1.45$, $p = 0.25$) (Fig 3e).

In Task B, there was a main effect of confidence on the proportion of reports ($F_{3,30} = 4.36$, $p = 0.012$). There was no significant effect of category ($F_{1,10} = 0.22$, $p = 0.64$), although there was an interaction between confidence and category ($F_{3,30} = 8.37$, $p = 0.003$). This is likely because for task B, category 2 has a higher proportion of "easy" stimuli (Fig 3f).

In both tasks, reported confidence had a significant effect on performance ($F_{3,30} = 36.9$, $p < 10^{-3}$). Task also had a significant effect on performance ($F_{1,10} = 20.1$, $p = 0.001$); although we chose the category parameters such that the performance of the optimal observer is matched, subjects were significantly better at Task A. There was no interaction between task and confidence ($F_{3,30} = 0.878$, $p = 0.436$) (Fig 3g).

Fig 3l and 3m shows psychometric choice curves for both tasks, at all 6 levels of reliability. Each point represents roughly the same number of trials.

Fig 3n and 3o shows a similar set of psychometric curves. These curves differ from Fig 3l and 3m in that they represent the mean button press rather than mean category choice.

In Task A (Fig 3l and 3n), mean category choice and mean button press depend monotonically on orientation, with a slope that increases with reliability. In Task B (Fig 3m and 3o), the mean category choice and mean button press tends towards category 1 when stimulus orientation is near horizontal, and tends towards category 2 when orientation is strongly tilted; this reflects the stimulus distributions.

Effect of stimulus type on results: Gabor vs. ellipse. Since some subjects only saw Gabors and some only saw ellipses, we used Spearman's rank correlation coefficient to measure the similarity of the two groups' model rankings. Spearman's rank correlation coefficient between Gabor and ellipse subjects for the summed LOO scores of the model groupings in Fig 6 and S1 Fig was 0.952 and 0.944, respectively (a value of 1 would indicate identical rankings). In both model groupings, the identities of the lowest- and highest-ranked models were the same for both Gabor and ellipse subjects. This indicates that the choice of stimulus type did not have a systematic effect on model rankings.

Experiment 2: Separate category and confidence responses and testing feedback

This control experiment was identical to experiment 1 except for the following modifications:

- Subjects first reported choice by pressing one of two buttons with their left hand, and then reported confidence by pressing one of four buttons with their right hand.
- Subjects reported confidence in category training blocks, and received correctness feedback after reporting confidence.
- There were no confidence training blocks.
- In testing blocks, subjects received correctness feedback after each trial.
- Subjects completed a total of 3240 testing trials.
- 8 subjects (0 male), aged 19–23, participated. None were participants in experiment 1, and again, none were fellow scientists.
- Drifting Gabors were used; no subjects saw ellipses.

Experiment 3: Task B only

This experiment was identical to experiment 1 except for the following modifications:

- Subjects completed blocks of Task B only.
- Subjects completed a total of 3240 testing trials.
- 15 subjects (8 male), aged 19–30, participated. None were participants in experiments 1 or 2.
- Drifting Gabors were used; no subjects saw ellipses.

Modeling

Measurement noise. For models (such as our core models) where the relationship between reliability (i.e., contrast or ellipse eccentricity) and noise was parametric, we assumed a power law relationship between reliability c and measurement noise variance σ^2: $\sigma^2(c) = \gamma + \alpha c^{-\beta}$. We have previously [30] used this power law relationship because it encompasses a large family of monotonically decreasing relationships using only three parameters. The relationship is also consistent with a form of the Naka-Rushton function [84, 85] commonly used to describe the mapping from reliability to neural gain g: $g = \frac{\gamma c^{\beta}}{c^{\beta}+\alpha}$. The power law relationship then holds under the assumption that measurement noise variance is inversely proportional to gain [86].

For all models except the Bayesian model with additive precision, we assumed additive orientation-dependent noise in the form of a rectified 2-cycle sinusoid, accounting for the finding that measurement noise is higher at non-cardinal orientations [87]. The measurement noise s.d. comes out to

$$\sigma(c, s) = \sqrt{\gamma + \alpha c^{-\beta}} + \psi \left| \sin \frac{\pi s}{90} \right|. \quad (1)$$

Response probability. We coded all responses as $r \in \{1, 2, \ldots, 8\}$, with each value indicating category and confidence. For all models except the Linear Neural model, the probability of a single trial i is equal to the probability mass of the measurement distribution $p(x \mid s_i) = \mathcal{N}(x; s_i, \sigma_i^2)$ (i.e., a normal distribution over x with mean s_i and variance σ_i^2) in a range corresponding to the subject's response r_i. Because we only use a small range of orientations, we can safely approximate measurement noise as a normal distribution rather than a Von Mises distribution. We find the boundaries $(b_{r_i-1}(\sigma_i), b_{r_i}(\sigma_i))$ in measurement space, as defined by the fitting model and parameters θ, and then compute the probability mass of the measurement distribution between the boundaries. For Task A, this quantity is

$$\int_{b_{r_i-1}}^{b_{r_i}} \mathcal{N}(x; s_i, \sigma_i^2)\mathrm{d}x, \tag{2}$$

where $b_0 = -\infty^\circ$ and $b_8 = \infty^\circ$. For Task B, this quantity is

$$p_{m,\theta}(r_i \mid s_i, \sigma_i) = \int_{-b_{r_i}}^{-b_{r_i-1}} \mathcal{N}(x; s_i, \sigma_i^2)\mathrm{d}x + \int_{b_{r_i-1}}^{b_{r_i}} \mathcal{N}(x; s_i, \sigma_i^2)\mathrm{d}x, \tag{3}$$

where $b_0 = 0^\circ$ and $b_8 = \infty^\circ$.

To obtain the log likelihood of the dataset, given a model with parameters θ, we compute the sum of the log probability for every trial i, where t is the total number of trials:

$$\log\ p(\text{data} \mid \theta) = \sum_{i=1}^{t} \log\ p(r_i \mid \theta) = \sum_{i=1}^{t} \log\ p_\theta(r_i \mid s_i, \sigma_i). \tag{4}$$

Model specification. *Bayesian*. Derivation of d_A and d_B: The log posterior ratio d is equivalent to the log likelihood ratio plus the log prior ratio:

$$d = \log \frac{p(C = 1 \mid x)}{p(C = 2 \mid x)} = \log \frac{p(x \mid C = 1)}{p(x \mid C = 2)} + \log \frac{p(C = 1)}{p(C = 2)}. \tag{5}$$

To get d_A and d_B, we need to find the task-specific expressions for $p(x \mid C)$. The observer knows that the measurement x is caused by the stimulus s, but has no knowledge of s. Therefore, the optimal observer marginalizes over s:

$$p(x \mid C) = \int p(x \mid s)p(s \mid C)\mathrm{d}s. \tag{6}$$

We substitute the expressions for the noise distribution and the stimulus distribution, and evaluate the integral:

$$p(x \mid C) = \int \mathcal{N}(s; x, \sigma^2)\mathcal{N}(s; \mu_C, \sigma_C^2)\mathrm{d}s = \mathcal{N}(x; \mu_C, \sigma^2 + \sigma_C^2). \tag{7}$$

Plugging the task- and category-specific μ_C and σ_C into Eq (7), and substituting the resulting

expression back into Eq (5), we get:

$$d_A = \frac{2x\mu_1}{\sigma^2 + \sigma_1^2} + \log \frac{p(C=1)}{p(C=2)} \tag{8}$$

$$d_B = \frac{1}{2} \log \frac{\sigma^2 + \sigma_2^2}{\sigma^2 + \sigma_1^2} - \frac{\sigma_2^2 - \sigma_1^2}{2(\sigma^2 + \sigma_1^2)(\sigma^2 + \sigma_2^2)} x^2 + \log \frac{p(C=1)}{p(C=2)}. \tag{9}$$

The 8 possible category and confidence responses are determined by comparing the log posterior ratio d to a set of decision boundaries $\mathbf{k} = (k_0, k_1, \ldots, k_8)$. k_4 is equal to the log prior ratio $\log \frac{p(C=1)}{p(C=2)}$, which functions as the boundary on d between the 4 category 1 responses and the 4 category 2 responses; k_4 is the only boundary parameter in models of category choice (and not confidence). k_0 is fixed at $-\infty$ and k_8 is fixed at ∞. In all models, the observer chooses category 1 when d is positive.

Because the decision boundaries are free parameters, our models effectively include a large family of possible cost functions. A different cost function would be equivalent to a rescaling of the confidence boundaries $\mathbf{k}$. To see this, it is probably easiest to consider category choice alone; there, asymmetric costs for getting either category wrong would translate into a different value of k_4, the category decision boundary (i.e., the observer's prior over category). For us, this boundary (like all other boundaries) is a free parameter.

The posterior probability of category 1 can be written as as $p(C = 1 \mid x) = \frac{1}{1+\exp(-d)}$.

Levels of strength: The Bayesian model is unique in that it is possible to formulate a principled version with relatively few boundary parameters. In principle, it is possible that such a model could perform better than more flexible models, if those models are overfitting. We formulated several levels of strength of the BCH, with weaker versions having fewer assumptions and more sets of mappings between the posterior probability of being correct and the confidence report (Fig 9). In the *ultrastrong BCH*, confidence is a function solely of the posterior

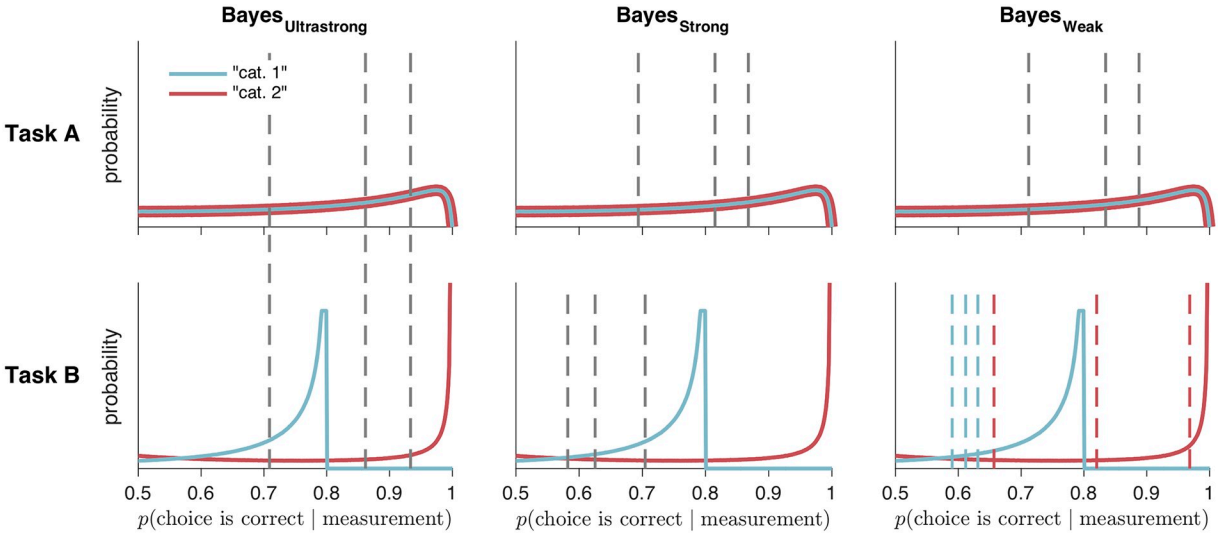

Fig 9. Distributions of posterior probabilities of being correct, with confidence criteria for Bayesian models with three different levels of strength. Solid lines represent the distributions of posterior probabilities for each category and task in the absence of measurement noise and sensory uncertainty. Dashed lines represent confidence criteria, generated from the mean of subject 4's posterior distribution over parameters. Each model has a different number of sets of mappings between posterior probability and confidence report. In $\text{Bayes}_{\text{Ultrastrong}}$, there is one set of mappings. In $\text{Bayes}_{\text{Strong}}$, there is one set for Task A, and another for Task B. In $\text{Bayes}_{\text{Weak}}$, as in the non-Bayesian models, there is one set for Task A, and one set for each reported category in Task B. Plots were generated from the mean of subject 4's posterior distribution over parameters as in Fig 2.

https://doi.org/10.1371/journal.pcbi.1006572.g009

probability of the chosen category. In the *strong BCH*, it is additionally a function of the current task.

Most studies cannot distinguish between the ultrastrong and strong BCH because they test subjects in only one task. Furthermore, the weak BCH is only justifiable in tasks where the categories have different distributions of the posterior probability of being correct; the subject may then rescale their mappings between the posterior and their confidence. Here, one can see that Task B has this feature by observing that, in the bottom row of Fig 9, the distributions of posterior probabilities are different for the two categories). Most experimental tasks are like Task A, where the distributions are identical. We compared Bayesian models ($\text{Bayes}_{\text{Ultrastrong}}$, $\text{Bayes}_{\text{Strong}}$) corresponding to each of these versions of the BCH.

In $\text{Bayes}_{\text{Ultrastrong}}$, $\mathbf{k}$ is symmetric across k_4: $k_{4+j} - k_4 = k_4 - k_{4-j}$ for $j \in \{1, 2, 3\}$. Furthermore, in $\text{Bayes}_{\text{Ultrastrong}}$, $\mathbf{k}_A = \mathbf{k}_B$. So $\text{Bayes}_{\text{Ultrastrong}}$ has a total of 4 free boundary parameters: k_1, k_2, k_3, k_4. $\text{Bayes}_{\text{Ultrastrong}}$ consists of the observer determining the response by comparing d_A and d_B to a single symmetric set of boundaries (Fig 9, left column).

$\text{Bayes}_{\text{Strong}}$ is identical to $\text{Bayes}_{\text{Ultrastrong}}$ except that $\mathbf{k_A}$ is allowed to differ from $\mathbf{k_B}$. So $\text{Bayes}_{\text{Strong}}$ has a total of 8 free boundary parameters: k_{1A}, k_{2A}, k_{3A}, k_{4A}, k_{1B}, k_{2B}, k_{3B}, k_{4B}. $\text{Bayes}_{\text{Strong}}$ consists of the observer determining the response by comparing d_A to a symmetric set of boundaries, and d_B to a different symmetric set of boundaries (Fig 9, middle column).

$\text{Bayes}_{\text{Weak}}$ is identical to $\text{Bayes}_{\text{Strong}}$ except that symmetry is not enforced for $\mathbf{k_B}$. So $\text{Bayes}_{\text{Weak}}$ has a total of 11 free boundary parameters: k_{1A}, k_{2A}, k_{3A}, k_{4A}, k_{1B}, k_{2B}, k_{3B}, k_{4B}, k_{5B}, k_{6B}, k_{7B}. $\text{Bayes}_{\text{Weak}}$ consists of the observer comparing d_A to a symmetric set of boundaries, and d_B to a different non-symmetric set of boundaries (Fig 9, right column).

We did not include $\text{Bayes}_{\text{Strong}}$ and $\text{Bayes}_{\text{Ultrastrong}}$ in the core models reported in the main text, because $\text{Bayes}_{\text{Weak}}$ provided a much better fit to the data. Because it was not necessary in the main text to distinguish the three strengths of Bayesian models, we refer to $\text{Bayes}_{\text{Weak}}$ there simply as Bayes. However, we do include $\text{Bayes}_{\text{Strong}}$ and $\text{Bayes}_{\text{Ultrastrong}}$ in our model recovery analysis (described below) and in our supplemental model comparison tables.

Decision boundaries: In the Bayesian models without d noise, we translate boundary parameters $\mathbf{k}$ to measurement boundaries $\mathbf{b}$ corresponding to fitted noise levels σ. To do this, we use the parameters $\mathbf{k}$ as the left-hand side of Eqs (8) and (9) and solve for x at the fitted levels of σ. These values were used as the measurement boundaries $\mathbf{b}(\sigma)$.

In the Bayesian models with d noise, we assume that, for each trial, there is an added Gaussian noise term on d, $\eta_d \sim p(\eta_d)$, where $p(\eta_d) = \mathcal{N}(0, \sigma_d^2)$, and σ_d is a free parameter. We pre-computed 101 evenly spaced draws of η_d and their corresponding probability densities $p(\eta_d)$. We used Eqs (8) and (9) to compute a lookup table containing the values of d as a function of x, σ, and η_d. We then used linear interpolation to find sets of measurement boundaries $\mathbf{b}(\sigma)$ corresponding to each draw of η_d [46]. We then computed 101 response probabilities for each trial, one for each draw of η_d, and computed the weighted average according to $p(\eta_d)$.

Probability correct with additive precision. We tested a model in which the decision variable was a weighted mixture of precision (equivalent in this case to the Fisher information of the measurement variable x) and the perceived probability of being correct [49]. In this model, the decision variable is $\frac{\omega}{\sigma^2} + \frac{1}{1+\exp(-|d|)}$, where ω is a free parameter. To find the measurement boundaries $\mathbf{b}(\sigma)$, we substituted Eqs (8) and (9) for d, and set the whole value equal to parameters $\mathbf{k}$, solving for x at the fitted levels of σ. This model can be considered a hybrid Bayesian-heuristic model. Like $\text{Bayes}_{\text{Ultrastrong}}$, it has 4 free boundary parameters. Although the model is a hybrid Bayesian-heuristic model, not a strictly Bayesian one, we refer to it as $\text{Bayes}_{\text{Ultrastrong}}$ + precision in S1 Fig and S1 Table.

Fixed. In Fixed, the observer compares the measurement to a set of boundaries that are not dependent on σ. We fit free parameters $\mathbf{k}$, and use measurement boundaries $b_r = k_r$.

Lin and Quad. In Lin and Quad, the observer compares the measurement to a set of boundaries that are linear or quadratic functions of σ. We fit free parameters $\mathbf{k}$ and $\mathbf{m}$, and use measurement boundaries $b_r(\sigma) = k_r + m_r\sigma$ (Lin) or $b_r(\sigma) = k_r + m_r\sigma^2$ (Quad).

Lin and Quad are each a supermodel of Fixed. In other words, there are parameter settings where Lin and Quad are equivalent to Fixed (although our model comparison methods ensure that the models are still distinguishable, see "Model recovery" section). Additionally, in Task A, Quad is a supermodel of the Bayesian models without d noise.

Orientation estimation. In Orientation Estimation, the observer uses the mixture of the two stimulus distributions as a prior distribution to compute a maximum a posteriori estimate of the stimulus:

$$\hat{s} = \underset{s}{\operatorname{argmax}}\, p(s \mid x) \tag{10}$$

$$= \underset{s}{\operatorname{argmax}}\, p(x \mid s)p(s) \tag{11}$$

$$= \underset{s}{\operatorname{argmax}}\, [\mathcal{N}(s; x, \sigma^2)(p(s \mid C = 1) + p(s \mid C = 2))]. \tag{12}$$

The observer then compares $\hat{s}$ to a set of boundaries $\mathbf{k}$ to determine category and confidence response.

Decision boundaries. To find the decision boundaries in measurement space, we used *gmm1max_n2_fast* from Luigi Acerbi's gmm1 (github.com/lacerbi/gmm1) 1-D Gaussian mixture model toolbox to solve Eq (12), computing a lookup table containing the value of $\hat{s}$ as a function of x and σ [46]. We then found, using linear interpolation, the values of x corresponding to σ and the free parameters $\mathbf{k}$. These values were used as the measurement boundaries $\mathbf{b}(\sigma)$.

Linear neural. In this section, $\mathbf{r}$ refers to neural activity, not button responses. This model is different from all other models in that the generative model does not include measurement x. The model can be derived as follows.

All neurons have Gaussian tuning curves with variance σ_{TC}^2 and gain $g = \frac{1}{\sigma^2}$. Tuning curve means are contained in the vector of preferred stimuli $\tilde{\mathbf{s}}$. The number of spikes in the population is $\mathbf{r} \sim \mathrm{Poisson}(g\mathcal{N}(s; \tilde{\mathbf{s}}, \sigma_{\mathrm{TC}}^2))$. Neural weights are a linear function of the preferred stimuli: $\mathbf{w} = a\tilde{\mathbf{s}}$.

On each trial, we get some quantity that is a weighted sum of each neuron's activity, $z = \mathbf{w} \cdot \mathbf{r}$. $\mathbb{E}[z \mid s] = \mathbf{w} \cdot \mathbb{E}[\mathbf{r} \mid s] = ag\sum_j \tilde{s}_j \exp\left(-\frac{(s-\tilde{s}_j)^2}{2\sigma_{\mathrm{TC}}^2}\right)$.

Rather than sum over all neurons, we assume an infinite number of neurons uniformly spanning all possible preferred stimuli $\tilde{s}$. This allows us to replace the sum with an integral. The expected value of z is $ag\int \tilde{s}\exp\left(-\frac{(s-\tilde{s}_j)^2}{2\sigma_{\mathrm{TC}}^2}\right)\mathrm{d}\tilde{s} = ags\sqrt{2\pi\sigma_{\mathrm{TC}}^2}$. The variance of z is $\sum_j w_j^2 f_j(s) = ag\int \tilde{s}^2\exp\left(-\frac{(s-\tilde{s})^2}{2\sigma_{\mathrm{TC}}^2}\right)\mathrm{d}\tilde{s} = ag\sqrt{2\pi\sigma_{\mathrm{TC}}^2}(\sigma_{\mathrm{TC}}^2 + s^2)$.

Now that we have the mean and variance of z, we assume that z is normally distributed. This is equivalent to assuming that there are a high number of spikes, because the Poisson distribution approximates the normal distribution as the rate parameter becomes high. To

compute response probability, we fit neural activity boundaries **k**, and replace Eqs (2) or (3) with

$$p_\theta(r_i \mid s_i, \sigma_i) = \int_{k_{r_i-1}}^{k_{r_i}} \mathcal{N}(z; ags_i\sqrt{2\pi\sigma_{\mathrm{TC}}^2}, ag\sqrt{2\pi\sigma_{\mathrm{TC}}^2}(\sigma_{\mathrm{TC}}^2 + s_i^2))\mathrm{d}z. \quad (13)$$

Lapse rates. In confidence and category models, we fit three different types of lapse rate. On each trial, there is some fitted probability of:

- A "full lapse" in which the category report is random, and confidence report is chosen from a distribution over the four levels defined by λ_1, the probability of a "very low confidence" response, and λ_4, the probability of a "very high confidence" response, with linear interpolation for the two intermediate levels.
- A "confidence lapse" $\lambda_{\mathrm{confidence}}$ in which the category report is chosen normally, but the confidence report is chosen from a uniform distribution over the four levels.
- A "repeat lapse" $\lambda_{\mathrm{repeat}}$ in which the category and confidence response is simply repeated from the previous trial.

In category choice models, we fit a standard category lapse rate λ, as well as the above "repeat lapse" $\lambda_{\mathrm{repeat}}$.

Parameterization. Because of tradeoffs when directly fitting parameters γ, α, β, we re-parameterized Eq (1) as

$$\sigma(c, s) = \sqrt{\sigma_{\mathrm{L}}^2 + \frac{(\sigma_{\mathrm{L}}^2 - \sigma_{\mathrm{H}}^2)(c^{-\beta} - c_{\mathrm{L}}^{-\beta})}{c_{\mathrm{L}}^{-\beta} - c_{\mathrm{H}}^{-\beta}}} + \psi\left|\sin\frac{\pi s}{90}\right|, \quad (14)$$

where c_{L} and c_{H} were the values of the lowest and highest reliabilities used. This way, σ_{L} and σ_{H} were free parameters that determined the s.d. of the measurement distributions for the lowest and highest reliabilities, and β was a free parameter determining the curvature of the function between the two reliabilities. For models where the relationship between reliability and noise was non-parametric, the first term in Eq (1) was replaced with free s.d. parameters ($\sigma_{\mathrm{rel.\ 1}}, \ldots, \sigma_{\mathrm{rel.\ 6}}$) corresponding to each of the six reliability levels.

For models where subjects had incorrect knowledge about their measurement noise, we fit two sets of uncertainty-related parameters. One set was for the generative measurement noise (used in Eqs (2) and (3)), and the other set was for the subject's belief about their noise, e.g., their sensory uncertainty (used in Eqs (8), (9) and (12)).

All parameters that defined the width of a distribution (e.g., σ_{L}, σ_{H}, σ_d, $\sigma_{\mathrm{rel.\ 1}}, \ldots$) were sampled in log-space and exponentiated during the computation of the log likelihood. See S9 Table for a complete list of each model's parameters.

Model fitting. Rather than find a maximum likelihood estimate of the parameters, we sampled from the posterior distribution over parameters, $p(\theta \mid \mathrm{data})$; this has the advantage of maintaining a measure of uncertainty about the parameters, which can be used both for model comparison and for plotting model fits. We used the log posterior

$$\log\, p(\theta \mid \mathrm{data}) = \log\, p(\mathrm{data} \mid \theta) + \log\, p(\theta) + \mathrm{constant}, \quad (15)$$

where $\log p(\mathrm{data} \mid \theta)$ is given in Eq (4). We assumed a factorized prior over each parameter j:

$$\log\, p(\theta) = \sum_{j=1}^{n} \log\, p(\theta_j), \quad (16)$$

where j is the parameter index and n is the number of parameters. We took uniform (or, for parameters that were standard deviations, log-uniform) priors over reasonable, sufficiently large ranges [46], which we chose before fitting any models.

We sampled from the probability distribution using a Markov Chain Monte Carlo (MCMC) method, slice sampling [88]. For each model and dataset combination, we ran between 4 and 7 parallel chains with random starting points. For each chain, we took 40,000 to 600,000 total samples (depending on model computational time) from the posterior distribution over parameters. We discarded the first third of the samples and kept 6,667 of the remaining samples, evenly spaced to reduce autocorrelation. All samples with log posteriors more than 40 below the maximum log posterior were discarded. Marginal probability distributions of the sample log likelihoods were visually checked for convergence across chains. In total we had 842 model and dataset combinations, with a median of 26,668 kept samples (IQR = 13,334).

After sampling, we conducted a visual check to confirm that our parameter ranges were sufficiently large. For each model, we plotted the posterior distribution over parameter values for each subject; an example plot is shown in Fig 10. Visual checks of these plots confirmed that the distributions are unimodal and roughly Gaussian. Visual checks also confirmed that the parameter distributions are well-contained within the chosen parameter ranges, except for the distributions of:

- Lapse rate parameters, which tend to mass around 0, where they are necessarily bounded.
- Log noise parameters, which have a large negative range where they are effectively at zero noise.
- Upper confidence boundary parameters, which become small for subjects who frequently report "high confidence," or large for subjects who frequently do.

Model comparison. *Model groupings.* We used 8 groupings of model-subject combinations where it made sense to consider the models as being on equal footing for the purpose of model comparison. The model-subject combinations were grouped by: experiment (which corresponded to subject population), data type (category response only vs. category and

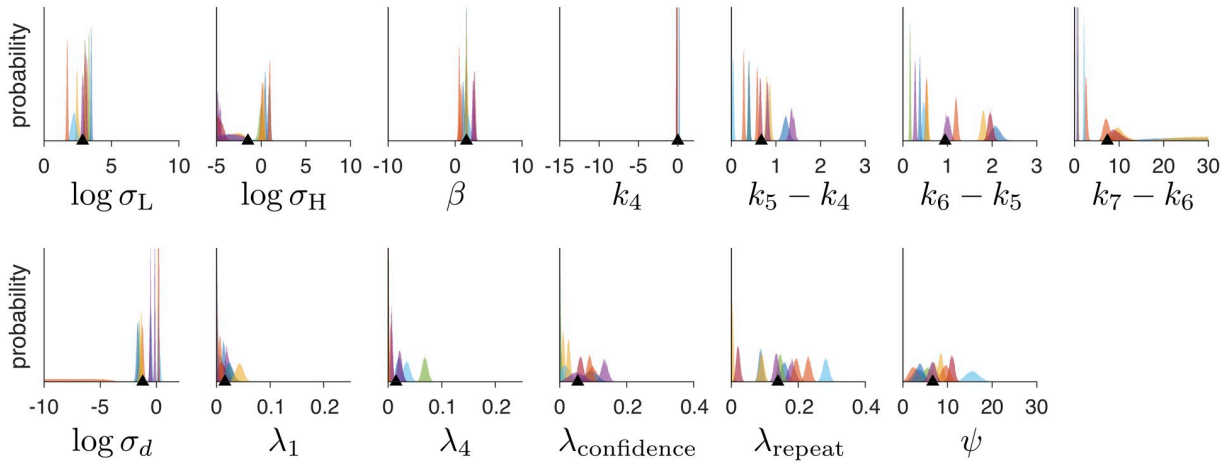

Fig 10. Posterior distributions over parameter values for an example model. Each subplot represents a parameter of the model. Each colored histogram represents the sampled posterior distribution for a parameter and a subject in experiment 1, with colors consistent for each subject. The limits of the x-axis indicates the allowable range for each parameter. Black triangles indicate the overall mean parameter value.

https://doi.org/10.1371/journal.pcbi.1006572.g010

confidence response), task type (Task A, B, or both fit jointly). The 8 groupings correspond to S1 to S8 Figs and S1 to S8 Tables.

Metric choice. A more complex model is likely to fit a dataset better than a simpler model, even if only by chance. Since we are interested in our models' predictive accuracy for unobserved data, it is important to choose a metric for model comparison that takes the complexity of the model into account, avoiding the problem of overfitting. Roughly speaking, there are two ways to compare models: information criteria and cross-validation.

Most information criteria (such as AIC, BIC, and AICc) are based on a point estimate for θ, typically θ_{MLE}, the θ that maximizes the log likelihood of the dataset (Eq (4)). For instance, AIC adds a correction for the number of parameters n to the log likelihood of the dataset: $\mathrm{AIC} = -2\sum_{i=1}^{t} \log\, p(r_i \mid \theta_{\mathrm{MLE}}) + 2n$.

WAIC is a more Bayesian approach to information criteria that adds a correction for the effective number of parameters [89]. Because WAIC is based on samples from the full posterior of θ (Eq (15), typically sampled via MCMC), it takes into account the model's uncertainty landscape.

Although information criteria are computationally convenient, they are based on asymptotic results and assumptions about the data that may not always hold [89]. An alternative way to estimate predictive accuracy for unobserved data is to cross-validate, fitting the model to training data and evaluating the fit on held out data. Leave-one-out cross-validation is the most thorough way to cross-validate, but is very computationally intensive; it requires that you fit your model t times, where t is the number of trials. Here we use a method (PSIS-LOO, referred to here simply as LOO) proposed by Vehtari et al. [43] for approximating leave-one-out cross-validation that, like WAIC, uses samples from the full posterior of θ:

$$\mathrm{LOO} = \sum_{i=1}^{t} \log \frac{\sum_u w_{i,u} p(r_i \mid \theta_u)}{\sum_u w_{i,u}}, \tag{17}$$

where θ_u is the u-th sampled set of parameters, and $w_{i,u}$ is the importance weight of trial i for sample u. Pareto smoothed importance sampling provides an accurate and reliable estimate of the weights. LOO is currently the most accurate approximation of leave-one-out cross-validation [90]. Conveniently, it has a natural diagnostic that allows the user to know when the metric may be inaccurate [43]; we used that diagnostic and confirmed that our use of the metric is justified.

We determined that our results were not dependent on our choice of metric. We computed AIC, BIC, AICc, WAIC, and LOO for all models in the 8 model groupings, multiplying the information criteria by $-\frac{1}{2}$ to match the scale of LOO. For AIC, BIC, and AICc, we used the parameter sample with the highest log likelihood as our estimate of θ_{MLE}. Then we computed Spearman's rank correlation coefficient for every possible pairwise comparison of model comparison metrics for all model and dataset combinations, producing 80 total values (8 model groupings × 10 possible pairwise comparisons of model comparison metrics). All values were greater than 0.998, indicating that, had we used an information criterion instead of LOO, we would not have changed our conclusions. Furthermore, there are no model groupings in which the identities of the lowest- and highest-ranked models are dependent on the choice of metric. The agreement of these metrics strengthens our confidence in our conclusions.

Metric aggregation. Summed LOO differences: In all figures where we present model comparison results (e.g., Fig 5, right column), we aggregate LOO scores by the following procedure. Choose a reference model (usually the one with the lowest mean LOO score across subjects). Subtract all LOO scores from the corresponding subject's score for the reference model; this converts all scores to a LOO "difference from reference" score, with higher scores

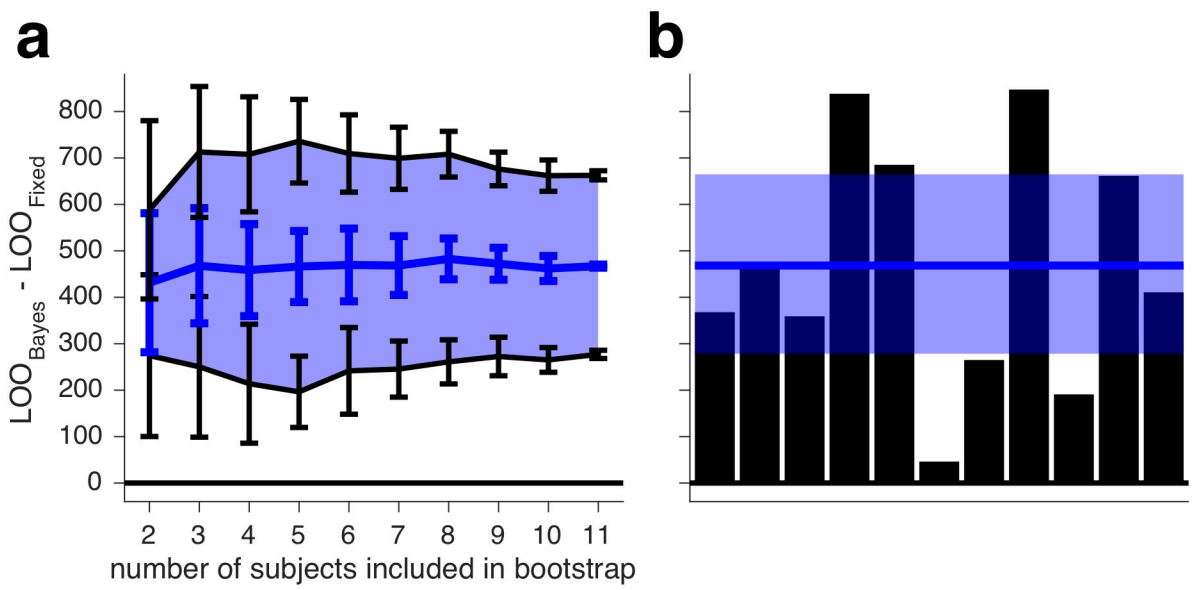

Fig 11. Example analysis of a bootstrapped confidence interval. (**a**) Uncertainty estimates for bootstrapped confidence intervals, as a function of the number of subjects included. Blue line represents the median bootstrapped mean of LOO differences, and black lines indicate the lower and upper bounds of the 95% CI. Error bars represent ±1 s.d. (**b**) For comparison to **a**, the standard style of plot used to show model comparison results (e.g., Fig 4).

https://doi.org/10.1371/journal.pcbi.1006572.g011

being worse. Repeat the following standard bootstrap procedure 10,000 times: Choose randomly, with replacement, a group of datasets equal to the total number of unique datasets, and take the sum over subjects of their "difference from reference" scores for each model. Plots indicate the median and 95% CI of these bootstrapped summed "difference from reference" scores. This approach implicitly assumes that all data was generated from the same model.

To confirm that our sample size was large enough to trust our bootstrapped confidence intervals, we bootstrapped our bootstrapping procedure to see how the confidence intervals were affected by the number of subjects *N*. For an example pair of models that we might be interested in comparing, we took the 11 LOO differences between the models, one for each subject in experiment 1. For each *N* between 2 and 11, we took 50 subsamples of our subject LOO differences with replacement; this is akin to running the experiment 50 times for each *N*. For each subsample, we conducted the above bootstrap procedure, which give us a median and 95% CI on the mean of differences. We then plot the mean of these values, with error bars indicating ±1 s.d., at each *N* (Fig 11a). A visual check indicates that the confidence interval appears to converge at about $N = 9$. This indicates that our bootstrapped confidence intervals are trustworthy.

Group level Bayesian model selection: We also used LOO scores to compute two metrics that allow for model heterogeneity across the group. The first metric was "protected exceedance probability," the posterior probability that one model occurs more frequently than any other model in the set [91], above and beyond chance (e.g., S1b Fig). The second was the expected posterior probability that a model generated the data of a randomly chosen dataset [92] (e.g., S1c Fig). The latter metric assumes a uniform prior over models, which is a function of the total number of datasets. We used the SPM12 (www.fil.ion.ucl.ac.uk/spm) software package to compute these metrics.

In all but one of the 8 model groupings, all three methods of metric aggregation identify the same overall best model. For example, in S1 Fig, one model (Quad + non-param. σ) has the

lowest summed LOO differences, the highest protected exceedance probability, and the highest expected posterior probability.

Visualization of model fits. Model fits were plotted by bootstrapping synthetic group datasets with the following procedure: For each task, model, and subject, we generated 20 synthetic datasets, each using a different set of parameters sampled, without replacement, from the posterior distribution of parameters. Each synthetic dataset was generated using the same stimuli as the ones presented to the real subject. We randomly selected a number of synthetic datasets equal to the number of subjects to create a synthetic group dataset. For each synthetic group dataset, we computed the mean output (e.g., button press, confidence, performance) per bin. We then repeated this 1,000 times and computed the mean and standard deviation of the mean output per bin across all 1,000 synthetic group datasets, which we then plotted as the shaded regions. Therefore, shaded regions represent the mean ±1 s.e.m. of synthetic group datasets.

For plots with orientation on the horizontal axis (e.g., Fig 3j–3o), stimulus orientation was binned according to quantiles of the task-dependent stimulus distributions so that each point consisted of roughly the same number of trials. For each task, we took the overall stimulus distribution $p(s) = \frac{1}{2}(p(s \mid C = 1) + p(s \mid C = 2))$ and found bin edges such that the probability mass of $p(s)$ was the same in each bin. We then plotted the binned data with linear spacing on the horizontal axis.

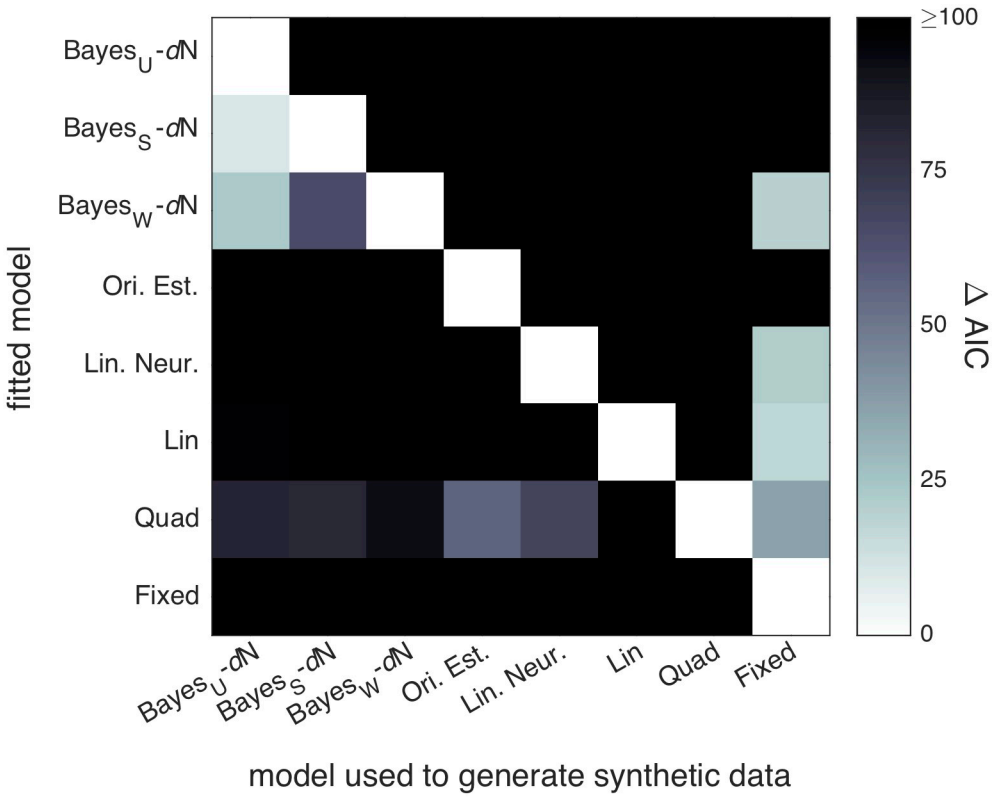

Fig 12. Model recovery analysis. Shade represents the difference between the mean AIC score (across datasets) for each fitted model and for the one with the lowest mean AIC score. White squares indicate the model that had the lowest mean AIC score when fitted to data generated from each model. The squares on the diagonal indicate that the true generating model was the best-fitting model, on average, in all cases.

https://doi.org/10.1371/journal.pcbi.1006572.g012

Model recovery. We performed a model recovery analysis [93] to test our ability to distinguish our 6 core models, as well as the 2 stronger versions of the Bayesian model. We generated synthetic datasets from each of the 8 models for both Tasks A and B, using the same sets of stimuli that were originally randomly generated for each of the 11 subjects. To ensure that the statistics of the generated responses were similar to those of the subjects, we generated responses to these stimuli from 4 of the randomly chosen parameter estimates obtained via MCMC sampling for each subject and model. In total, we generated 352 datasets (8 generating models × 11 subjects × 4 datasets). We then fit all 8 models to every dataset, using maximum likelihood estimation (MLE) of parameters by an interior-point constrained optimization (MATLAB's *fmincon*), and computed AIC scores from the resulting fits.

We found that the true generating model was the best-fitting model, on average, in all cases (Fig 12). Overall, AIC "selected" the correct model (i.e., AIC scores were lowest for the model that generated the data) for 86.6% of the datasets, indicating that our models are distinguishable.

Ideally, we would have evaluated our model recovery fits using LOO, as we evaluated the fits to human data. However, LOO can only be obtained when fitting with MCMC sampling, which takes orders of magnitudes longer than fitting with MLE. It would be impossible to fit all 352 synthetic datasets in a short amount of time using the procedure and sampling quality standards described above (i.e., a large number of samples, across multiple converged chains). Furthermore, we do not believe that our model recovery is dependent on how the models are fit and the fits are evaluated; we found that AIC and LOO scores gave us near-identical model rankings for data from real subjects.

Supporting information

S1 Fig. Model comparison, experiment 1. Models were fit jointly to Task A and B category and confidence responses. (a) Medians and 95% CI of bootstrapped sums of LOO differences, relative to the best model. Higher values indicate worse fits. (**b**) The protected exceedance probability, i.e., the posterior probability that a model occurs more frequently than the others [91]. (**c**) The expected posterior probability that a model generated the data of a randomly chosen subject [92].
(TIF)

S2 Fig. Model comparison, experiment 1. Models were fit to Task A category and confidence responses. See S1 Fig caption.
(TIF)

S3 Fig. Model comparison, experiment 1. Models were fit to Task B category and confidence responses. See S1 Fig caption.
(TIF)

S4 Fig. Model comparison, experiment 1. Models were fit jointly to Task A and B category choices. See S1 Fig caption.
(TIF)

S5 Fig. Model comparison, experiment 1. Noise parameters were fit to Task A category choices and then fixed during the fitting of Task B category and confidence responses. See S1 Fig caption.
(TIF)

S6 Fig. Model comparison, experiment 2. Models were fit jointly to Task A and B category and confidence responses. See S1 Fig caption.
(TIF)

S7 Fig. Model comparison, experiment 3. Models were fit to Task B category and confidence responses. See S1 Fig caption.
(TIF)

S8 Fig. Model comparison, experiment 3. Models were fit to Task B category choices. See S1 Fig caption.
(TIF)

S9 Fig. Model fits and model comparison for the three strengths of the Bayesian model, as in Fig 5. In the main text, $\text{Bayes}_{\text{Weak}}$-$d$N is referred to simply as Bayes.
(TIF)

S10 Fig. $\text{Bayes}_{\text{Weak}}$-$d$N fits, as in Fig 3. In the main text, $\text{Bayes}_{\text{Weak}}$-$d$N is referred to simply as Bayes.
(TIF)

S11 Fig. Fixed fits, as in Fig 3.
(TIF)

S12 Fig. Orientation Estimation fits, as in Fig 3.
(TIF)

S13 Fig. Linear Neural fits, as in Fig 3.
(TIF)

S14 Fig. Lin fits, as in Fig 3.
(TIF)

S15 Fig. Quad fits, as in Fig 3, but for data in experiment 2.
(TIF)

S1 Table. Cross comparison of all models in S1 Fig. Cells indicate medians and 95% CI of bootstrapped summed LOO score differences. A negative median indicates that the model in the corresponding row had a higher score (better fit) than the model in the corresponding column.
(PDF)

S2 Table. Cross comparison of all models in S2 Fig. See S1 Table caption.
(PDF)

S3 Table. Cross comparison of all models in S3 Fig. See S1 Table caption.
(PDF)

S4 Table. Cross comparison of all models in S4 Fig. See S1 Table caption.
(PDF)

S5 Table. Cross comparison of all models in S5 Fig. See S1 Table caption.
(PDF)

S6 Table. Cross comparison of all models in S6 Fig. See S1 Table caption.
(PDF)

S7 Table. Cross comparison of all models in S7 Fig. See S1 Table caption.
(PDF)

S8 Table. Cross comparison of all models in S8 Fig. See S1 Table caption.
(PDF)

S9 Table. List of parameters for each model. Each sheet corresponds with the sets of models pictured in S1 Fig and S1; S2 Fig and S2 Table; and so on.
(XLS)

Acknowledgments

The authors would like to thank Luigi Acerbi for helpful ideas and tools related to model fitting and model comparison. We would also like to thank Luigi Acerbi, Rachel N. Denison, Andra Mihali, A. Emin Orhan, Bas van Opheusden, and Aspen H. Yoo for helpful conversations and comments about the manuscript.

Author Contributions

Conceptualization: William T. Adler, Wei Ji Ma.

Data curation: William T. Adler.

Formal analysis: William T. Adler.

Funding acquisition: William T. Adler, Wei Ji Ma.

Investigation: William T. Adler.

Methodology: William T. Adler.

Project administration: William T. Adler, Wei Ji Ma.

Resources: William T. Adler.

Software: William T. Adler.

Supervision: Wei Ji Ma.

Validation: William T. Adler.

Visualization: William T. Adler.

Writing – original draft: William T. Adler.

Writing – review & editing: William T. Adler, Wei Ji Ma.

References

1. Meyniel F, Sigman M, Mainen ZF. Confidence as Bayesian probability: From neural origins to behavior. Neuron. 2015 Oct; 88(1):78–92. https://doi.org/10.1016/j.neuron.2015.09.039 PMID: 26447574
2. Brown AS. A review of the tip-of-the-tongue experience. Psychol Bull. 1991 Mar; 109(2):204–223. https://doi.org/10.1037/0033-2909.109.2.204 PMID: 2034750
3. Persaud N, McLeod P, Cowey A. Post-decision wagering objectively measures awareness. Nat Neurosci. 2007 Jan; 10(2):257–261. https://doi.org/10.1038/nn1840 PMID: 17237774
4. Bahrami B, Olsen K, Latham PE, Roepstorff A. Optimally interacting minds. Science. 2010; 329 (5995):1081–1085. https://doi.org/10.1126/science.1185718 PMID: 20798320
5. Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G. Relating introspective accuracy to individual differences in brain structure. Science. 2010 Sep; 329(5998):1541–1543. https://doi.org/10.1126/science.1191883 PMID: 20847276

6. Fleming SM, Dolan RJ. The neural basis of metacognitive ability. Phil Trans R Soc B. 2012 May; 367(1594):1338–1349. https://doi.org/10.1098/rstb.2011.0417 PMID: 22492751
7. Rutishauser U, Ye S, Koroma M, Tudusciuc O, Ross IB, Chung JM, et al. Representation of retrieval confidence by single neurons in the human medial temporal lobe. Nat Neurosci. 2015 Jun; 18(7):1041–1050. https://doi.org/10.1038/nn.4041 PMID: 26053402
8. Kiani R, Shadlen MN. Representation of confidence associated with a decision by neurons in the parietal cortex. Science. 2009 May; 324(5928):759–764. https://doi.org/10.1126/science.1169405 PMID: 19423820
9. Fetsch CR, Kiani R, Newsome WT, Shadlen MN. Effects of cortical microstimulation on confidence in a perceptual decision. Neuron. 2014 Aug; 83(4):797–804. https://doi.org/10.1016/j.neuron.2014.07.011 PMID: 25123306
10. Komura Y, Nikkuni A, Hirashima N, Uetake T, Miyamoto A. Responses of pulvinar neurons reflect a subject's confidence in visual categorization. Nat Neurosci. 2013 Jun; 16(6):749–755. https://doi.org/10.1038/nn.3393 PMID: 23666179
11. Kepecs A, Uchida N, Zariwala HA, Mainen ZF. Neural correlates, computation and behavioural impact of decision confidence. Nature. 2008 Sep; 455(7210):227–231. https://doi.org/10.1038/nature07200 PMID: 18690210
12. Pouget A, Drugowitsch J, Kepecs A. Confidence and certainty: Distinct probabilistic quantities for different goals. Nat Neurosci. 2016 Feb; 19(3):366–374. https://doi.org/10.1038/nn.4240 PMID: 26906503
13. Kepecs A, Mainen ZF. A computational framework for the study of confidence in humans and animals. Phil Trans R Soc B. 2012 May; 367(1594):1322–1337. https://doi.org/10.1098/rstb.2012.0037 PMID: 22492750
14. Drugowitsch J, Moreno-Bote R, Pouget A. Relation between belief and performance in perceptual decision making. PLoS ONE. 2014 May; 9(5):e96511. https://doi.org/10.1371/journal.pone.0096511 PMID: 24816801
15. Peirce CS, Jastrow J. On small differences in sensation. Memoirs of the National Academy of Sciences. 1884; 3:73–83.
16. Knill DC, Richards W. Perception as Bayesian inference. Cambridge: Cambridge University Press; 1996.
17. Ma WJ, Jazayeri M. Neural coding of uncertainty and probability. Annu Rev Neurosci. 2014; 37(1):205–220. https://doi.org/10.1146/annurev-neuro-071013-014017 PMID: 25032495
18. Körding K. Decision theory: What "should" the nervous system do? Science. 2007 Oct; 318(5850):606–610. https://doi.org/10.1126/science.1142998 PMID: 17962554
19. Aitchison L, Bang D, Bahrami B, Latham PE. Doubly Bayesian analysis of confidence in perceptual decision-making. PLoS Comput Biol. 2015 Oct; 11(10):e1004519. https://doi.org/10.1371/journal.pcbi.1004519 PMID: 26517475
20. Sanders JI, Hangya B, Kepecs A. Signatures of a statistical computation in the human sense of confidence. Neuron. 2016 May; 90(3):499–506. https://doi.org/10.1016/j.neuron.2016.03.025 PMID: 27151640
21. Hangya B, Sanders JI, Kepecs A. A mathematical framework for statistical decision confidence. Neural Computation. 2016 Sep; 28(9):1840–1858. https://doi.org/10.1162/NECO_a_00864 PMID: 27391683
22. Adler WT, Ma WJ. Limitations of proposed signatures of Bayesian confidence. Neur Comp. 2018;. https://doi.org/10.1162/neco_a_01141
23. Bowers JS, Davis CJ. Bayesian just-so stories in psychology and neuroscience. Psychol Bull. 2012; 138(3):389–414. https://doi.org/10.1037/a0026450 PMID: 22545686
24. Jones M, Love BC. Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behav Brain Sci. 2011 Aug; 34(4):169–188. https://doi.org/10.1017/S0140525X10003134 PMID: 21864419
25. Navajas J, Bahrami B, Latham PE. Post-decisional accounts of biases in confidence. Curr Opin Behav Sci. 2016; 11:55–60. https://doi.org/10.1016/j.cobeha.2016.05.005
26. Pleskac TJ, Busemeyer JR. Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. Psychol Rev. 2010 Jul; 117(3):864–901. https://doi.org/10.1037/a0019737 PMID: 20658856
27. Britten KH, Shadlen MN, Newsome WT, Movshon JA. The analysis of visual motion: A comparison of neuronal and psychophysical performance. J Neurosci. 1992 Dec; 12(12):4745–4765. https://doi.org/10.1523/JNEUROSCI.12-12-04745.1992 PMID: 1464765
28. Liu Z, Knill DC, Kersten D. Object classification for human and ideal observers. Vis Res. 1995 Feb; 35(4):549–568. https://doi.org/10.1016/0042-6989(94)00150-K PMID: 7900295

29. Sanborn AN, Griffiths TL, Shiffrin RM. Uncovering mental representations with Markov chain Monte Carlo. Cogn Psychol. 2010 Mar; 60(2):63–106. https://doi.org/10.1016/j.cogpsych.2009.07.001 PMID: 19703686

30. Qamar AT, Cotton RJ, George RG, Beck JM, Prezhdo E, Laudano A, et al. Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. PNAS. 2013 Dec; 110(50):20332–20337. https://doi.org/10.1073/pnas.1219756110 PMID: 24272938

31. Maniscalco B, Peters MAK, Lau H. Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. Atten Percept Psychophys. 2016 Jan; 78(3):923–937. https://doi.org/10.3758/s13414-016-1059-x PMID: 26791233

32. Maloney LT, Mamassian P. Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. Vis Neurosci. 2009 Jan; 26(1):147–155. https://doi.org/10.1017/S0952523808080905 PMID: 19193251

33. Körding KP, Wolpert DM. Bayesian integration in sensorimotor learning. Nature. 2004 Jan; 427(6971):244–247. https://doi.org/10.1038/nature02169 PMID: 14724638

34. Massoni S, Gajdos T, Vergnaud JC. Confidence measurement in the light of signal detection theory. Front Psychol. 2014; 5(325):1455. https://doi.org/10.3389/fpsyg.2014.01455 PMID: 25566135

35. Brier GW. Verification of forecasts expressed in terms of probability. Monthly Weather Review. 1950; 78(1):1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

36. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc. 2007 Mar; 102(477):359–378. https://doi.org/10.1198/016214506000001437

37. Green DM, Swets JA. Signal detection theory and psychophysics. New York: Wiley; 1966.

38. Ma WJ. Signal detection theory, uncertainty, and Poisson-like population codes. Vis Res. 2010 Oct; 50(22):2308–2319. https://doi.org/10.1016/j.visres.2010.08.035 PMID: 20828581

39. Rahnev D, Maniscalco B, Graves T, Huang E, de Lange FP, Lau H. Attention induces conservative subjective biases in visual perception. Nat Neurosci. 2011 Oct; 14(12):1513–1515. https://doi.org/10.1038/nn.2948 PMID: 22019729

40. Ma WJ. Organizing probabilistic models of perception. Trends Cogn Sci. 2012; 16(10):511–518. https://doi.org/10.1016/j.tics.2012.08.010 PMID: 22981359

41. Maniscalco B, Lau H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Conscious Cogn. 2012 Mar; 21(1):422–430. https://doi.org/10.1016/j.concog.2011.09.021 PMID: 22071269

42. Fleming SM, Lau HC. How to measure metacognition. Front Hum Neurosci. 2014; 8:443. https://doi.org/10.3389/fnhum.2014.00443 PMID: 25076880

43. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. arXiv. 2015 Jul;.

44. Keshvari S, van den Berg R, Ma WJ. Probabilistic computation in human perception under variability in encoding precision. PLoS ONE. 2012 Jun; 7(6):e40216. https://doi.org/10.1371/journal.pone.0040216 PMID: 22768258

45. Kass RE, Raftery AE. Bayes factors. J Am Stat Assoc. 1995; 90(430):773–795. https://doi.org/10.1080/01621459.1995.10476572

46. Acerbi L, Vijayakumar S, Wolpert DM. On the origins of suboptimality in human probabilistic inference. PLoS Comput Biol. 2014 Jun; 10(6):e1003661. https://doi.org/10.1371/journal.pcbi.1003661 PMID: 24945142

47. Beck JM, Ma WJ, Pitkow X, Latham PE, Pouget A. Not noisy, just wrong: The role of suboptimal inference in behavioral variability. Neuron. 2012 Apr; 74(1):30–39. https://doi.org/10.1016/j.neuron.2012.03.016 PMID: 22500627

48. Orhan AE, Jacobs RA. Are performance limitations in visual short-term memory tasks due to capacity limitations of model mismatch? arXiv. 2014 Jul;.

49. Navajas J, Hindocha C, Foda H, Keramati M, Latham PE, Bahrami B. The idiosyncratic nature of confidence. Nat Hum Behav. 2017 Sep; 11(11):1–12.

50. Kiani R, Corthell L, Shadlen MN. Choice certainty is informed by both evidence and decision time. Neuron. 2014 Dec; 84(6):1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015 PMID: 25521381

51. Wilimzig C, Tsuchiya N, Fahle M, Einhäuser W, Koch C. Spatial attention increases performance but not subjective confidence in a discrimination task. Journal of Vision. 2008 May; 8(5):7.1–10. https://doi.org/10.1167/8.5.7

52. Palminteri S, Wyart V, Koechlin E. The importance of falsification in computational cognitive modeling. Trends Cogn Sci. 2017 Jun; 21(6):425–433. https://doi.org/10.1016/j.tics.2017.03.011 PMID: 28476348

53. Zylberberg A, Barttfeld P, Sigman M. The construction of confidence in a perceptual decision. Frontiers in Integrative Neuroscience. 2012; 6:79. https://doi.org/10.3389/fnint.2012.00079 PMID: 23049504

54. Vickers DD. Decision processes in visual perception. New York: Academic Press; 1979.

55. van den Berg R, Anandalingam K, Zylberberg A, Kiani R, Shadlen MN, Wolpert DM. A common mechanism underlies changes of mind about decisions and confidence. eLife. 2016 Feb;5:e12192. https://doi.org/10.7554/eLife.12192 PMID: 26829590

56. Peters MAK, Thesen T, Ko YD, Maniscalco B, Carlson C, Davidson M, et al. Perceptual confidence neglects decision-incongruent evidence in the brain. Nat Hum Behav. 2017; 1(0139). https://doi.org/10.1038/s41562-017-0139 PMID: 29130070

57. Barthelmé S, Mamassian P. Evaluation of objective uncertainty in the visual system. PLoS Comput Biol. 2009 Sep; 5(9):e1000504. https://doi.org/10.1371/journal.pcbi.1000504 PMID: 19750003

58. Barthelmé S, Mamassian P. Flexible mechanisms underlie the evaluation of visual confidence. In: PNAS; 2010.

59. Rahnev DA, Maniscalco B, Luber B, Lau H, Lisanby SH. Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. J Neurophysiol. 2012 Mar; 107(6):1556–1563. https://doi.org/10.1152/jn.00985.2011 PMID: 22170965

60. Bressler DW, Silver MA. Spatial attention improves reliability of fMRI retinotopic mapping signals in occipital and parietal cortex. 2010 Aug;p. 1–8.

61. Pestilli F, Carrasco M, Heeger DJ, Gardner JL. Attentional enhancement via selection and pooling of early sensory responses in human visual cortex. Neuron. 2011; 72(5):832–846. https://doi.org/10.1016/j.neuron.2011.09.025 PMID: 22153378

62. Wyart V, Nobre AC, Summerfield C. Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. PNAS. 2012; 109(9):3593–3598. https://doi.org/10.1073/pnas.1120118109 PMID: 22331901

63. Solovey G, Graney GG, Lau H. A decisional account of subjective inflation of visual perception at the periphery. Atten Percept Psychophys. 2014; 77(1):258–271. https://doi.org/10.3758/s13414-014-0769-1

64. Denison RN, Adler WT, Carrasco M, Ma WJ. Humans flexibly incorporate attention-dependent uncertainty into perceptual decisions and confidence. PNAS. 2018;. https://doi.org/10.1073/pnas.1717720115 PMID: 30297430

65. Cortese A, Amano K, Koizumi A, Kawato M, Lau H. Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. Nat Commun. 2016 Dec; 7:1–18. https://doi.org/10.1038/ncomms13669

66. Koizumi A, Maniscalco B, Lau H. Does metacognitive awareness facilitate cognitive control? 2014 Jun; p. 1–22.

67. Rounis E, Maniscalco B, Rothwell JC, E PR, Lau H. Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Cogn Neurosci. 2010; 1:165–175. https://doi.org/10.1080/17588921003632529 PMID: 24168333

68. Simons JS, Peers PV, Mazuz YS, Berryhill ME, Olson IR. Dissociation between memory accuracy and memory confidence following bilateral parietal lesions. Cereb Cortex. 2010; 20:479–485. https://doi.org/10.1093/cercor/bhp116 PMID: 19542474

69. Lau HC, Passingham RE. Relative blindsight in normal observers and the neural correlate of visual consciousness. PNAS. 2006; 103:18763–18768. https://doi.org/10.1073/pnas.0607716103 PMID: 17124173

70. Koizumi A, Maniscalco B, Lau H. Does perceptual confidence facilitate cognitive control? Atten Percept Psychophys. 2015; 77:1295–1306. https://doi.org/10.3758/s13414-015-0843-3 PMID: 25737256

71. Samaha J, Barrett JJ, Sheldon AD, Larocque JJ, Postle BR. Dissociating perceptual confidence from discrimination accuracy reveals no influence of metacognitive awareness on working memory. Front Psychol. 2016; 7(851). https://doi.org/10.3389/fpsyg.2016.00851 PMID: 27375529

72. Fleming SM, Huijgen J, Dolan RJ. Prefrontal contributions to metacognition in perceptual decision making. J Neurosci. 2012 May; 32(18):6117–6125. https://doi.org/10.1523/JNEUROSCI.6489-11.2012 PMID: 22553018

73. Gigerenzer G, Hertwig R, Pachur T. Heuristics: The foundations of adaptive behavior. Oxford: Oxford University Press; 2011.

74. Folke T, Jacobsen C, Fleming SM, De Martino B. Explicit representation of confidence informs future value-based decisions. Nat Hum Behav. 2016 Nov; 1(2).

75. Bahrami B, Olsen K, Bang D, Roepstorff A, Rees G, Frith C. What failure in collective decision-making tells us about metacognition. Phil Trans R Soc B. 2012 May; 367(1594):1350–1365. https://doi.org/10.1098/rstb.2011.0420 PMID: 22492752

76. Bang D, Fusaroli R, Tylén K, Olsen K, Latham PE, Lau JYF, et al. Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. Conscious Cogn. 2014 May; 26:13–23. https://doi.org/10.1016/j.concog.2014.02.002 PMID: 24650632

77. Lennie P. The cost of cortical computation. Current Biology. 2003 Mar; 13(6):493–497. https://doi.org/10.1016/S0960-9822(03)00135-0 PMID: 12646132

78. Attwell D, Laughlin SB. An energy budget for signaling in the grey matter of the brain. J Cereb Blood Flow Metab. 2001 Oct; 21(10):1133–1145. https://doi.org/10.1097/00004647-200110000-00001 PMID: 11598490

79. Chklovskii DB, Koulakov AA. Maps in the brain: What can we learn from them? Annu Rev Neurosci. 2004; 27:369–392. https://doi.org/10.1146/annurev.neuro.27.070203.144226 PMID: 15217337

80. Clune J, Mouret JB, Lipson H. The evolutionary origins of modularity. Proceedings of the Royal Society B: Biological Sciences. 2013 Jan; 280:20122863. https://doi.org/10.1098/rspb.2012.2863 PMID: 23363632

81. Orhan AE, Ma WJ. Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback. Nat Commun. 2017 Jul; 8(1):138. https://doi.org/10.1038/s41467-017-00181-8 PMID: 28743932

82. Pelli DG. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spat Vis. 1997; 10(4):437–442. https://doi.org/10.1163/156856897X00366 PMID: 9176953

83. Brainard DH. The Psychophysics Toolbox. Spat Vis. 1997; 10(4):433–436. https://doi.org/10.1163/156856897X00357 PMID: 9176952

84. Naka KI, Rushton WA. S-potentials from luminosity units in the retina of fish (Cyprinidae). Journal of Physiology. 1966 Aug; 185(3):587–599. https://doi.org/10.1113/jphysiol.1966.sp008001 PMID: 5918060

85. DiMattina C. Comparing models of contrast gain using psychophysical experiments. Journal of Vision. 2016 Jul; 16(9):1–18. https://doi.org/10.1167/16.9.1 PMID: 27380470

86. Ma WJ, Beck JM, Latham PE, Pouget A. Bayesian inference with probabilistic population codes. Nat Neurosci. 2006 Nov; 9(11):1432–1438. https://doi.org/10.1038/nn1790 PMID: 17057707

87. Girshick AR, Landy MS, Simoncelli EP. Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. Nat Neurosci. 2011 Jun; 14(7):926–932. https://doi.org/10.1038/nn.2831 PMID: 21642976

88. Neal RM. Slice sampling. 2003; 31(3):705–767.

89. Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. Statistics and Computing. 2013; 24(6):997–1016. https://doi.org/10.1007/s11222-013-9416-2

90. Acerbi L, Dokka K, Angelaki DE, Ma WJ. Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. bioRxiv. 2017 Jun;.

91. Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group studies—Revisited. NeuroImage. 2014 Jan; 84:971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065 PMID: 24018303

92. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. NeuroImage. 2009 Jul; 46(4):1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025 PMID: 19306932

93. van den Berg R, Awh E, Ma WJ. Factorial comparison of working memory models. Psychol Rev. 2014 Jan; 121(1):124–149. https://doi.org/10.1037/a0035234 PMID: 24490791