# The computations underlying human confidence reports are probabilistic, but not Bayesian

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

[1]Center for Neural Science, [2]Department of Psychology,
New York University, New York, NY

*Corresponding author: will.adler@nyu.edu

**Humans can meaningfully rate their confidence in a perceptual or cognitive decision. It is widely believed that these reports reflect the estimated probability that the decision is correct, but, upon closer look, this belief is a hypothesis rather than an established fact. In a pair of perceptual categorization tasks, we tested whether explicit confidence reports reflect the Bayesian posterior probability of being correct. This Bayesian hypothesis predicts that subjects take sensory uncertainty into account in a specific way in the computation of confidence ratings. We find that confidence reports are probabilistic: subjects take sensory uncertainty into account on a trial-to-trial basis. However, they do not do so in the way predicted by the Bayesian hypothesis. Instead, heuristic probabilistic models provide the best fit to human confidence ratings. This conclusion is robust to changes in the uncertainty manipulation, task, response modality, additional flexibility in the Bayesian model, and model comparison metric. To better understand the origins of the heuristic computation, we trained feedforward neural networks consisting of generic units with error feedback, mapped the output of the trained networks to confidence ratings, and fitted our behavioral models to the resulting synthetic datasets. We find that the synthetic confidence ratings are also best fit by heuristic probabilistic models. This suggests that implementational constraints cause explicit confidence reports to deviate from being Bayesian.**

## Introduction

People often have a sense of a level of confidence about their decisions. Such a "feeling of knowing"[1] may serve to improve performance in subsequent decisions[2], learning[1], and group decision-making[3]. Much recent work has focused on identifying brain regions and neural mechanisms responsible for the computation of confidence in humans[4,5,6], non-human primates[7,8,9], and rodents[10]. 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[1,11,12,13]. In human studies, however, naïve subjects can give a meaningful answer when you ask them to rate their confidence about a decision[14]; 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[15,16,17]. 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[18,19] but, because of their experimental designs, are unable to meaningfully distinguish the Bayesian model from any other model of confidence.

We test the quantitative 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 rarely done but very important for the epistemological standing of Bayesian claims[20,21]. We find that the BCH qualitatively describes human behavior, but that quantitatively, even the most flexible Bayesian model is outperformed by non-Bayesian probabilistic models. To better understand why, we trained neural networks to perform one of our tasks. We found that the same non-Bayesian models provided the best description of the output of the trained networks, suggesting that neural network architecture constrains the computation of confidence.

# Results

## Experiment 1

During each session, each subject completed two orientation categorization tasks, Tasks A and B. On each trial, the observer categorized a single oriented stimulus as category 1 or 2 and reported their confidence on a 4-point scale. Category and confidence were reported simultaneously, with a single button press (**Fig. 1a**). 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 ($\sigma_C$); leftward-tilting stimuli were more likely to be from category 1. Variants of Task A are common in decision-making studies[22]. In Task B, the distributions had the same mean (0°) and different standard deviations ($\sigma_1, \sigma_2$); stimuli around the horizontal were more likely to be from category 1. Variants of Task B are less common[23,24,25] 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).

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; during testing, correctness feedback was withheld to avoid the possibility that confidence simply reflects a learned mapping between stimuli and the probability of being correct, something that no other confidence studies have done[25,26,27].

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). If we had, we would have essentially been training subjects to use a specific model of confidence.

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 (Supplementary Information).

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. 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. $\sigma$ (**Fig. 1d,e**). In the model, $\sigma$ is a function of stimulus reliability (Supplementary Information).

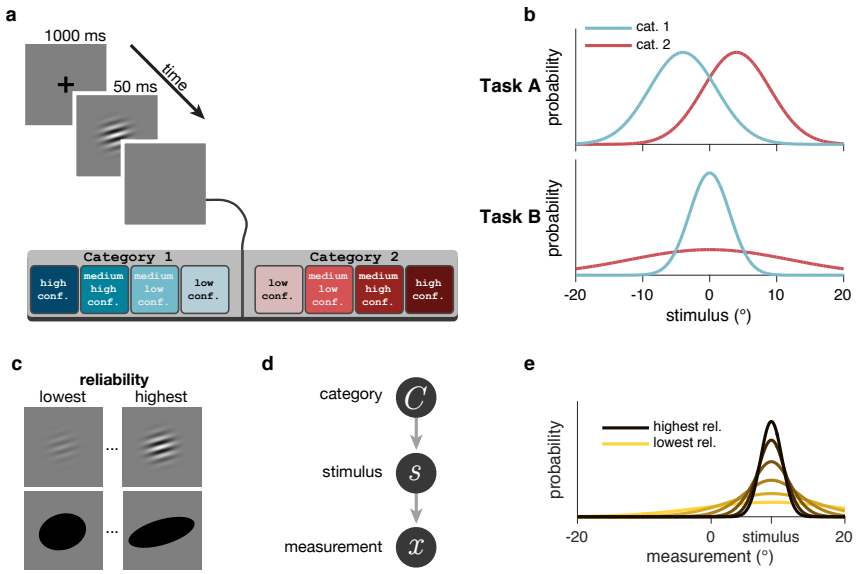


Figure 1: Task design and generative model. (**a**) Schematic of a test block trial. After stimulus offset, subjects reported category and confidence level with a single button press. (**b**) Stimulus distributions for Tasks A and B. (**c**) Examples of low and high reliability stimuli. Six (out of eleven) 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.

## 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 $d_{\mathrm{A}} = \frac{2x\mu_C}{\sigma^2+\sigma_C^2}$. Therefore, the ideal observer reports category 1 when $x$ is positive; this is the structure of many psychophysical tasks[28]. In Task B, however, $d$ is $d_{\mathrm{B}} = \frac{1}{2}\log\frac{\sigma^2+\sigma_2^2}{\sigma^2+\sigma_1^2} - \frac{\sigma_2^2-\sigma_1^2}{2(\sigma^2+\sigma_1^2)(\sigma^2+\sigma_2^2)}x^2$; the observer needs both $x$ and $\sigma$ in order to make an optimal decision. See Supplementary Informationfor derivations of $d_{\mathrm{A}}$ and $d_{\mathrm{B}}$.

From the point of view of the observer, $\sigma$ is the trial-to-trial level of sensory uncertainty associated with the measurement[29]. 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 $d_{\mathrm{A}}$ and $d_{\mathrm{B}}$.

We introduce the Bayesian Confidence Hypothesis (BCH), stating that confidence reports depend on the internal representation of the stimuli (here $x$) only via the posterior probability. 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. Another way of thinking about this is: 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 formulate several levels of strength of the BCH, with weaker versions having fewer assumptions and more sets of mappings between the posterior probability and the confidence report (Supplementary Information,

**Fig. S1**). In the *ultrastrong BCH*, confidence is a function solely of the posterior probability of the chosen category. In the *strong BCH*, it is additionally a function of the current task. In the *weak BCH*, it is additionally a function of the identity of the chosen category. 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, Task B has this feature (**Fig. S1**, bottom row); most experimental tasks do not. We compared Bayesian models ($\text{Bayes}_{\text{Ultrastrong}}$, $\text{Bayes}_{\text{Strong}}$, $\text{Bayes}_{\text{Weak}}$) corresponding to each of these versions of the BCH.

The observer's decision can be summarized as a mapping from a combination of a measurement and an uncertainty level $(x, \sigma)$ to a response that indicates both category and confidence. We can visualize this mapping as in **Figure 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 and can even be non-monotonic functions of uncertainty.

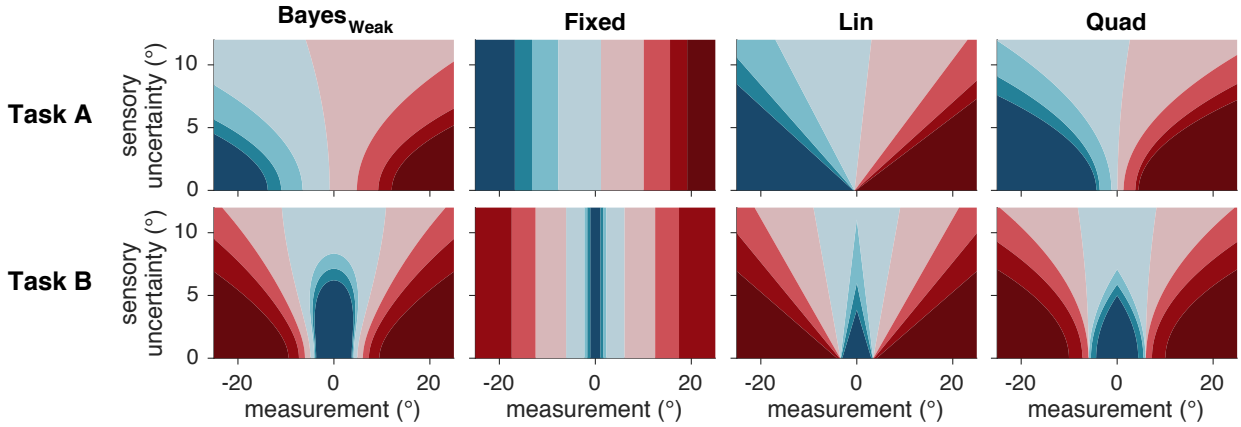


Figure 2: Decision rules/mappings in four models. Each model corresponds to a different mapping from a measurement and uncertainty level to a category and confidence response. Colors correspond to category and confidence response, as in **Figure 1a**. Plots were generated from the mean of subject 4's posterior distribution over parameters.

## 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[9,10,30]. 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).

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[31]. 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. In the Lin and Quad models, subjects base their response on a linear or a quadratic function of $x$ and $\sigma$, respectively (Supplementary Information). A comparison of the Lin and Quad columns to the $\text{Bayes}_{\text{Weak}}$ column in **Figure 2** demonstrates that Lin

and Quad can approximate the Bayesian mapping from $(x, \sigma)$ to response without actually computing $d$. All of the models we tested were variants of the eight models described so far ($\text{Bayes}_{\text{Ultrastrong}}$, $\text{Bayes}_{\text{Strong}}$, $\text{Bayes}_{\text{Weak}}$, Fixed, Orientation Estimation, Linear Neural, Lin, Quad). We will refer to these models, when fitted jointly to category and confidence data from Tasks A and B, as our core models.

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 **Figure 3**. We find the following effects: performance and confidence increase as a function of reliability (**Fig. 3a,b,h,i**), and high-confidence reports are less frequent than low-confidence reports (**Fig. 3e,f**). Note **Figure 3c,d** 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.

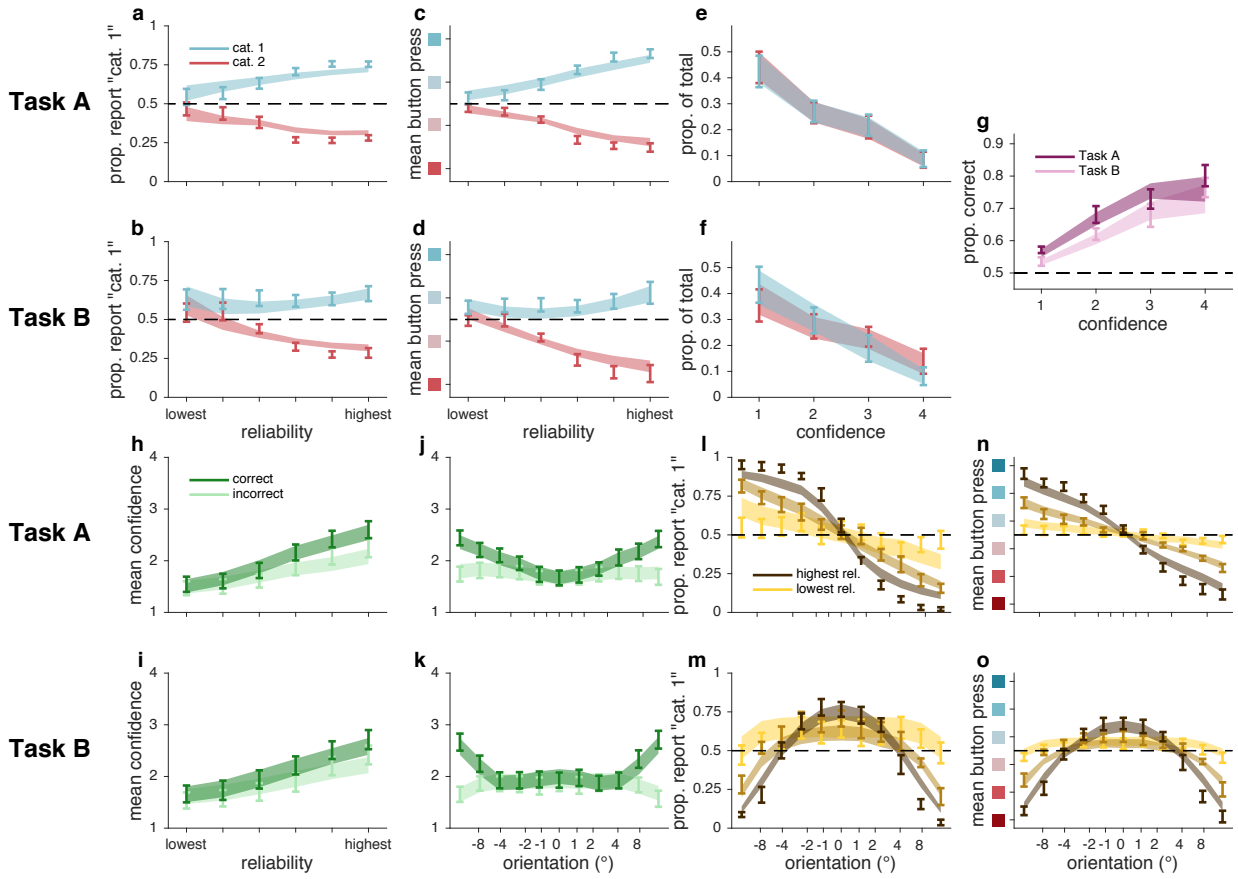


Figure 3: Behavioral data and fits from best model (Quad), experiment 1. Error bars represent $\pm 1$ s.e.m. across 11 subjects. Shaded regions represent $\pm 1$ s.e.m. on model fits. (**a,b**) Proportion "category 1" reports as a function of stimulus reliability and true category. (**c,d**) Mean button 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 **Figure 1a**. (**j-o**) See **Visualization of model fits** for details on orientation binning. (**l-o**) For clarity, only 3 of 6 reliability levels are shown, although models were fit to all reliability levels.

## 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[32]. A model recovery analysis ensured that our models are meaningfully distinguishable (Supplementary Information, **Fig. S2**). Unless otherwise noted, models were fit jointly to Task A and B category and confidence responses.

**Use of sensory uncertainty.** We first compared the Bayesian models to the Fixed model, in which the observer does not take trial-to-trial sensory uncertainty into account (**Fig. 4**). Fixed provides a poor fit to the data, which indicates that observers use not only a point estimate of their measurement, but also their uncertainty about that measurement. All three Bayesian models outperform Fixed, by summed LOO differences (median and 95% CI of bootstrapped sums across subjects) of 2265 [498, 4253] ($\text{Bayes}_{\text{Ultrastrong}}$), 3595 [1995, 5323] ($\text{Bayes}_{\text{Strong}}$), and 4312 [2610, 6151] ($\text{Bayes}_{\text{Weak}}$). For the rest of this paper, we will report model comparison results using this format.

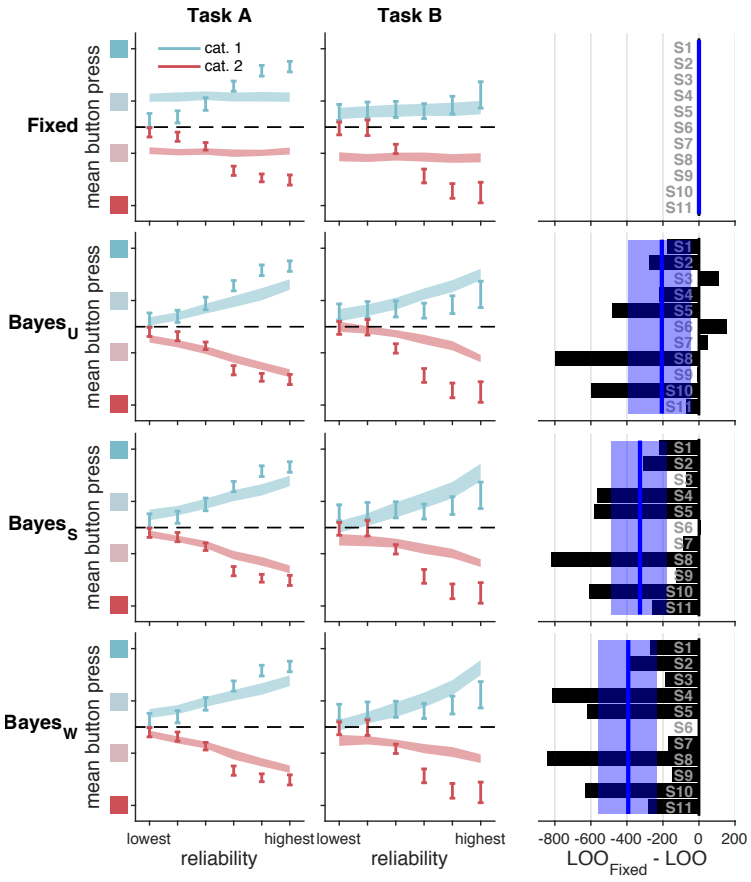


Figure 4: Model fits and model comparison for Fixed and Bayesian models. Left and middle columns: model fits to mean button press as a function of reliability, true category, and task. Error bars represent $\pm 1$ s.e.m. across 11 subjects. Shaded regions represent $\pm 1$ s.e.m. on model fits, with each model on a separate row. Right column: LOO model comparison. Bars represent individual subject LOO scores for every model, relative to Fixed. Negative values indicate that the model in the corresponding row had a better (higher) 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.

Out of the three Bayesian models, $\text{Bayes}_{\text{Weak}}$ provides the best fit, indicating that, if confidence is a function of the posterior probability, people rescale their mappings from posterior to response depending on task and category. However, all Bayesian models still show systematic deviations from the data, especially at high reliabilities.

**Noisy log posterior ratio.** To see if we could improve the fits of the Bayesian models, we tried a set of Bayesian models 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. $\sigma_d$. This is almost equivalent to the probability of a response being a logistic (softmax) function of $d$[33]. Adding $d$ noise improves the fits of the Bayesian models by 657 [214, 1217] ($\text{Bayes}_{\text{Ultrastrong}}$), 662 [307, 1150] ($\text{Bayes}_{\text{Strong}}$), and 804 [510, 1134] ($\text{Bayes}_{\text{Weak}}$). However, there are still clear deviations from the data at high reliabilities, and the fits are qualitatively poor. The ranking of the three models remains the same; $\text{Bayes}_{\text{Weak}}$ with $d$ noise is now the best-performing Bayesian model (**Fig. S5**). For the rest of this paper, we will only consider Bayesian models with $d$ noise.

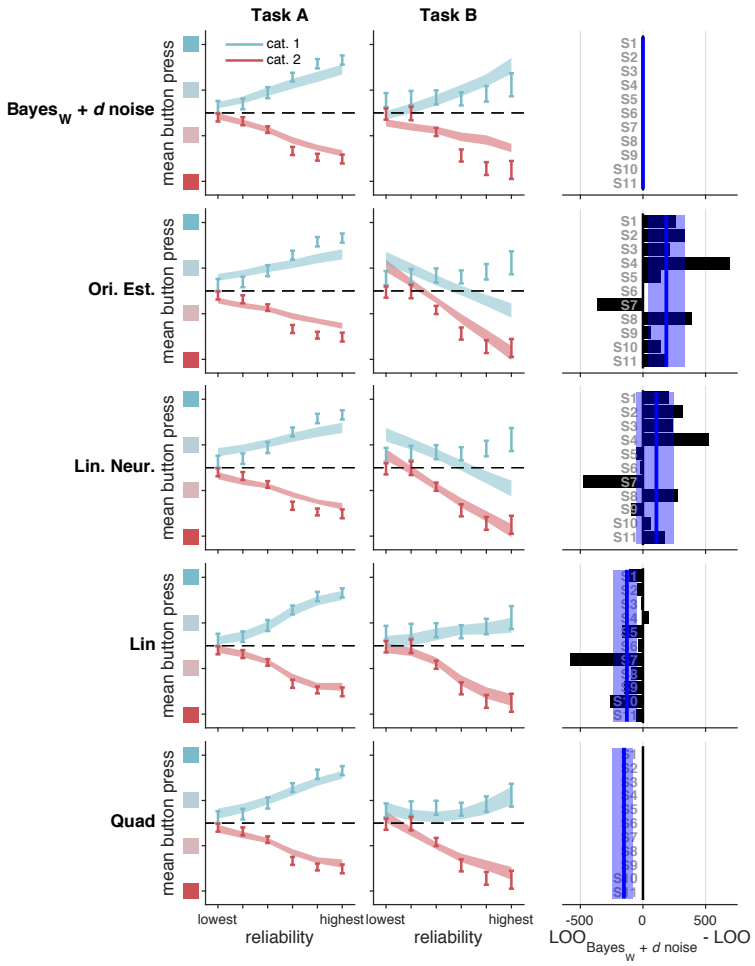


Figure 5: Model fits and model comparison for best-fitting Bayesian model and heuristic models, as in **Figure 4**.

**Heuristic models.** Orientation Estimation performs worse than our best performing Bayesian model, 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 model performs about as well as our best performing Bayesian model,

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 the best-performing Bayesian model by 3545 [2401, 5177] (Lin) and 3799 [2521, 5455] (Quad). 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 **Fig. S8n,o** to **Fig. S12n,o** and **Fig. 3n,o**).

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[34], the heuristic models Lin and Quad perform much better than the Bayesian models. Furthermore, we can decisively rule out Fixed. We will now test variants of our core models.

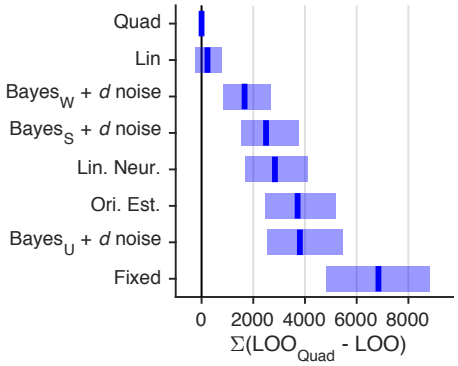


Figure 6: Comparison of core models, experiment 1. Models were fit jointly to Task A and B category and confidence responses. Blue lines and shaded regions represent, respectively, medians and 95% CI of bootstrapped summed LOO differences across subjects.

**Non-parametric relationship between reliability and $\sigma$.** One potential criticism of our fitting procedure is that we assumed a parameterized relationship between reliability and $\sigma$ (Supplementary Information). To see if our results were dependent on that assumption, we modified the models such that $\sigma$ was non-parametric (i.e., there was a free parameter for $\sigma$ at each level of reliability). With this feature added to our core models, Quad still fits better than the best-fitting Bayesian model by 1676 [839, 2730], and better than Fixed by 6097 [4323, 7901] (**Fig. S13** and **Table S1**). This feature improved Quad's performance by 325 [141, 535]. For the rest of this paper, we will only report the best-fitting Bayesian model, 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."[35,36,37] Up to now, our Bayesian models have assumed that observers have accurate knowledge of the parameters of the generative model. To test whether this assumption prevents the Bayesian models from fitting the data well, we tested a series of Bayesian models in which the observer has inaccurate knowledge of the generative model.

The previously described Bayesian models assumed that, because subjects were well-trained, they knew the true values of $\sigma_C$, $\sigma_1$, and $\sigma_2$, the standard deviations of the stimulus distributions. We tested models 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 the Bayesian model in the case where subjects were not trained enough to sufficiently learn the stimulus distributions. This feature improves the fit of the best Bayesian model by 908 [318, 1661], but it still underperforms Quad by 768 [399, 1144] (**Fig. S13** and **Table S1**).

Previous models also assumed that subjects had full knowledge of their own measurement noise; the $\sigma$ used in the computation of $d$ was identical to the $\sigma$ that determined their measurement noise. We tested models in which we fit $\sigma_{\text{measurement}}$ and $\sigma_{\text{inference}}$ as two independent functions of reliability[35]. This feature improves the fit of the best Bayesian model by 1310 [580, 2175], but it still underperforms Quad by 362 [162, 602] (**Fig. S13** and **Table S1**).

**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 the Bayesian model by 581 [278, 938], and better than Fixed by 3534 [2529, 4552] (**Fig. S14** and **Table S2**). In Task B, Quad fits better than the best-fitting Bayesian model by 978 [406, 1756], and better than Fixed by 3234 [2099, 4390] (**Fig. S15** and **Table S3**).

**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 the only Bayesian model by 595 [311, 927], and better than Fixed by 1690 [976, 2534] (**Fig. S16** and **Table S4**).

**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 fitted the Fixed model 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 the best-fitting Bayesian model by 1773 [451, 2845], and better than Fixed by 5016 [3090, 6727] (**Fig. S17** and **Table S5**).

**Experiment 2: Separate category and confidence responses, and testing feedback.** 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 separate responses[19,38,39,40]. Aitchison *et al.*[41] found that confidence appears more Bayesian when subjects use separate responses. To confirm this, we ran a second experiment that was like theirs in two ways. First, subjects chose a category by pressing one of two buttons, then reported confidence by pressing one of four buttons. Second, correctness feedback was given on every trial, rather than only on training blocks. After fitting our core models, our results did not differ substantially from experiment 1: Lin fits better than the best-fitting Bayesian model by 396 [186, 622], and better than Fixed by 2095 [1344, 2889] (**Fig. S18** and **Table S6**).

**Experiment 3: Task B only.** 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. Quad fits better than the best-fitting Bayesian model by 1361 [777, 2022], and better than Fixed by 7326 [4905, 9955] (**Fig. S19** and **Table S7**). 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 the Bayesian model, with summed LOO differences of 117 [-76, 436], and better than Fixed by 1084 [619, 1675] (**Fig. S20** and **Table S8**). This approximately replicates our previously published results[25].

**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 (Supplementary Information).

## Neural network

Taken together, the model comparisons in experiments 1 to 3 convince us that there is no obvious way to explain human confidence ratings as Bayesian. Does this mean that the normative framework must be entirely rejected? We should instead consider the possibility that implementational constraints restrict the brain's ability to perform fully Bayesian computation[20,21]. To explore this possibility, we trained biologically plausible feedforward neural networks (**Fig. 7a**) to perform Task B[42], and fitted their output with the same models that we used to fit subject data. We found that Lin is the best-fitting model, outperforming $\text{Bayes}_{\text{Strong}}$ by summed AIC differences of 22504 [19898, 25500] (**Fig. 7b**, blue). Although the training procedure was different for the networks than for the human subjects (they were trained, using back-propagation, at all 6 noise levels), the networks perform the task in a way that is qualitatively similar to the subjects (**Fig. 7c,d**). This suggests that the architecture or the training procedure of the neural networks constrains the type of behavior that can be produced (Supplementary Information).

One possibility is that the Bayesian model is too inflexible to fit any behavioral dataset based on neural activity. To rule out this possibility, we decoded optimal posterior probabilities from input unit activity on a

per-trial basis, mapped these onto button presses using quantiles, and fit the behavioral models. $\text{Bayes}_{\text{Strong}}$ provides the best fit, fitting these datasets better than Lin by 5845 [4032, 8103] (**Fig. 7b**, black). Thus, the fact that Lin wins is not due to general inflexibility of the Bayesian models.

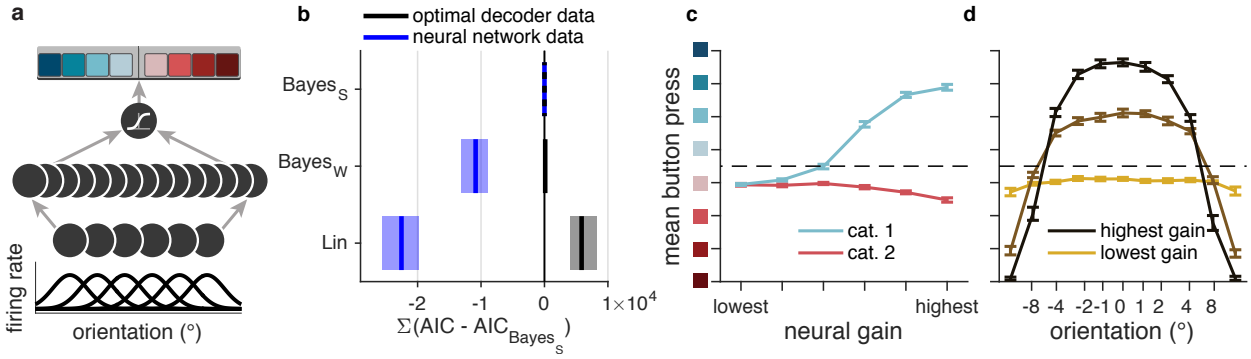


Figure 7: Neural network architecture, model comparison, and task behavior. (**a**) Feedforward neural network architecture. Input units were independent Poisson neurons with Gaussian tuning curves that were evenly spaced and of identical width. Input units were connected, all-to-all, to hidden rectified linear units. Hidden units were connected to a sigmoidal output unit. The output was mapped onto a category and confidence response (as in **Fig. 1a**) using eight quantiles (Supplementary Information). (**b**) Model comparison (as in **Fig. 6**) for data generated by the optimal decoder, and by trained neural networks. Fixed, Orientation Estimation, Linear Neural, and Quad were also fit to those data (**Fig. S4**), but are not shown because they all fit more poorly than Lin. (**c**) Task behavior of trained neural networks. Mean button press as a function of neural gain and true category. Compare to **Figure 3d**. (**d**) Mean button press as a function of stimulus orientation and neural gain. Compare to **Figure 3o**.

# Discussion

Although people can report subjective feelings of confidence, the computations that produce this feeling are not well-understood. Confidence has been defined as the observer's computed posterior probability that a decision is correct. However, this hypothesis has not been fully tested. We used model comparison to investigate the computational underpinnings of human confidence reports. We also trained neural networks to perform a perceptual task, treating the network output as if it were subject-generated data for the purpose of model comparison[42]. We carried out a strong and comprehensive test of a set of cognitive models, varying task components such as stimulus reliability and stimulus distributions[26].

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 subject ignores their sensory uncertainty provided a poor fit to the data (**Fig. 4**). Our second finding is that, unlike the optimal observer, subjects do not appear to use knowledge of their sensory uncertainty in a Bayesian way. Instead, heuristic models that approximate Bayesian computation—but do not compute a posterior probability over category—outperform the Bayesian models in a variety of experimental contexts (**Fig. 5**). 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 were meaningfully distinguishable (Supplementary Information).

We trained neural networks to perform one of our tasks. Although the training procedure necessarily differed from that of the humans, we found that the trained networks produced confidence responses that, like the human data, were best fit by heuristic models. This suggests that the structure of the neural network—and by extension, the structure of the brain—limits its ability to produce accurate posterior estimates in categorization tasks.

We do not advocate for the heuristic Lin and Quad models as general descriptions of human confidence; our main message is that human behavior is best described by models that are non-Bayesian. Some may argue that non-Bayesian models should be rejected because they are not generalizable. Bayesian models derive their generalizability from their normative nature: in any task, one can determine the performance-maximizing strategy. However, it is not clear that this property should override a bad fit.

Moreover, performance maximization is only one of several ecologically relevant organizing principles. The brain is also limited by the kinds of operations that neurons can perform and the ways by which organisms learn[20,21]; our neural network analysis suggests that architectural or learning constraints may cause the brain to deviate from Bayesian computations. Future work could investigate the possibility that the brain is near-optimal under implementational constraints; this would connect Marrian levels within a single rational framework[43].

We will now describe how our results relate to recently published experimental findings. Rahnev *et al.*[30] reported that subjective decision criteria are fixed across conditions of uncertainty. However, their study did not test models in which the criteria was a function of uncertainty, so they cannot make this conclusion very strongly. Additionally, their study used visibility ratings, which differ from confidence ratings[44]. Finally, their results may be specific to the case where sensory uncertainty is a function of attention rather than stimulus reliability. We leave this question for future work.

Sanders *et al.*[19] reported that confidence has a "statistical" nature. However, their experiment was unable to determine whether confidence is probabilistic or Bayesian[16], because the stimuli vary along only one dimension. As noted by Aitchison *et al.*, to distinguish models of confidence, the experimenter must use stimuli that are characterized by two dimensions (e.g., contrast and orientation)[41]. 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.*, the measurement and the posterior probability of being correct). Therefore, the only alternative model proposed by Sanders *et al.* is based on reaction time, rather than on the presented stimuli.

Like the present study, Aitchison *et al.*[41] 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 (i.e., whether observers take sensory uncertainty into account on a trial-to-trial basis). Such a test requires that the experimenter vary the reliability of the stimulus feature of interest.

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[6] and prefrontal cortex[45], monkey lateral intraparietal cortex[7] and pulvinar[9], rodent orbitofrontal cortex[10]) 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,[41] neural activity associated with confidence may represent either the measurement or the probability of being correct. In addition to the recommendation of Aitchison *et al.* to distinguish between these possibilities by varying stimuli along two dimensions, we recommend fitting both Bayesian and non-Bayesian probabilistic models to behavior. Many physiological studies of decision-making focus on correlating neural activity to parameters of behavioral models. This approach only makes sense when the behavioral model is a good description of the behavior. Our results suggest that the Bayesian model is a relatively poor description of confidence behavior. Therefore, the proposal to do this kind of correlational analysis with parameters of the Bayesian confidence models[11] should be viewed with skepticism.

**ACKNOWLEDGEMENTS**
The authors would like to thank Emin Orhan for assistance on the neural network analysis, and Luigi Acerbi for helpful ideas and tools related to model fitting and model comparison. We would also like to thank Luigi Acerbi, Aspen Yoo, Andra Mihali, and Rachel Denison for helpful conversations and comments on the manuscript. W.T.A. was supported by a Graduate Research Fellowship from the National Science Foundation.

**AUTHOR CONTRIBUTIONS**
W.T.A. and W.J.M. designed the experiments, analyzed the data, and wrote the manuscript. W.T.A. performed the experiments.

**COMPETING FINANCIAL INTERESTS**
The authors declare no competing financial interests.

# References

[1] Meyniel, F., Sigman, M. & Mainen, Z. F. Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron* **88**, 78–92 (2015).

[2] Persaud, N., McLeod, P. & Cowey, A. Post-decision wagering objectively measures awareness. *Nat. Neurosci.* **10**, 257–261 (2007).

[3] Bahrami, B., Olsen, K., Latham, P. E. & Roepstorff, A. Optimally interacting minds. *Science* (2010).

[4] Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J. & Rees, G. Relating Introspective Accuracy to Individual Differences in Brain Structure. *Science* **329**, 1541–1543 (2010).

[5] Fleming, S. M. & Dolan, R. J. The neural basis of metacognitive ability. *Phil. Trans. R. Soc. B* **367**, 1338–1349 (2012).

[6] Rutishauser, U. *et al.* Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nat. Neurosci.* **18**, 1041–1050 (2015).

[7] Kiani, R. & Shadlen, M. N. Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. *Science* **324**, 759–764 (2009).

[8] Fetsch, C. R., Kiani, R., Newsome, W. T. & Shadlen, M. N. Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron* **83**, 797–804 (2014).

[9] Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T. & Miyamoto, A. Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nat. Neurosci.* **16**, 749–755 (2013).

[10] Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227–231 (2008).

[11] Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19**, 366–374 (2016).

[12] Kepecs, A. & Mainen, Z. F. A computational framework for the study of confidence in humans and animals. *Phil. Trans. R. Soc. B* **367**, 1322–1337 (2012).

[13] Drugowitsch, J., Moreno-Bote, R. & Pouget, A. Relation between Belief and Performance in Perceptual Decision Making. *PLoS ONE* **9**, e96511 (2014).

[14] Peirce, C. S. & Jastrow, J. On Small Differences in Sensation. *Memoirs of the National Academy of Sciences* **3**, 73–83 (1884).

[15] Knill, D. C. & Richards, W. *Perception as Bayesian Inference* (Cambridge University Press, 1996).

[16] Ma, W. J. & Jazayeri, M. Neural coding of uncertainty and probability. *Annu. Rev. Neurosci.* (2014).

[17] Kording, K. Decision Theory: What "Should" the Nervous System Do? *Science* **318**, 606–610 (2007).

[18] Aitchison, L. & Latham, P. E. Bayesian synaptic plasticity makes predictions about plasticity experiments in vivo. *arXiv* (2014). 1410.1029v2.

[19] Sanders, J. I., Hangya, B. & Kepecs, A. Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron* **90**, 499–506 (2016).

[20] Bowers, J. S. & Davis, C. J. Bayesian just-so stories in psychology and neuroscience. *Psychol. Bull.* **138**, 389–414 (2012).

[21] Jones, M. & Love, B. C. Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behav. Brain. Sci.* **34**, 169–188 (2011).

[22] Britten, K. H., Shadlen, M. N., Newsome, W. T. & Movshon, J. A. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J. Neurosci.* **12**, 4745–4765 (1992).

[23] Liu, Z., Knill, D. C. & Kersten, D. Object classification for human and ideal observers. *Vis. Res.* **35**, 549–568 (1995).

[24] Sanborn, A. N., Griffiths, T. L. & Shiffrin, R. M. Uncovering mental representations with Markov chain Monte Carlo. *Cogn. Psychol.* **60**, 63–106 (2010).

[25] Qamar, A. T. *et al.* Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *PNAS* **110**, 20332–20337 (2013).

[26] Maloney, L. T. & Mamassian, P. Bayesian decision theory as a model of human visual perception: testing Bayesian transfer. *Vis. Neurosci.* **26**, 147–155 (2009).

[27] Körding, K. P. & Wolpert, D. M. Bayesian integration in sensorimotor learning. *Nature* **427**, 244–247 (2004).

[28] Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics* (1966).

[29] Ma, W. J. Signal detection theory, uncertainty, and Poisson-like population codes. *Vis. Res.* **50**, 2308–2319 (2010).

[30] Rahnev, D. *et al.* Attention induces conservative subjective biases in visual perception. *Nat. Neurosci.* **14**, 1513–1515 (2011).

[31] Ma, W. J. Organizing probabilistic models of perception. *Trends. Cogn. Sci.* **16**, 511–518 (2012).

[32] Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *arXiv* (2015). 1507.04544v4.

[33] Keshvari, S., van den Berg, R. & Ma, W. J. Probabilistic Computation in Human Perception under Variability in Encoding Precision. *PLoS ONE* **7**, e40216–9 (2012).

[34] Kass, R. E. & Raftery, A. E. Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773–795 (2012).

[35] Acerbi, L., Vijayakumar, S. & Wolpert, D. M. On the Origins of Suboptimality in Human Probabilistic Inference. *PLoS Comput. Biol.* **10**, e1003661 (2014).

[36] Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E. & Pouget, A. Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. *Neuron* **74**, 30–39 (2012).

[37] Orhan, A. E. & Jacobs, R. A. Are Performance Limitations in Visual Short-Term Memory Tasks Due to Capacity Limitations or Model Mismatch? *arXiv* (2014). 1407.0644v1.

[38] Kiani, R., Corthell, L. & Shadlen, M. N. Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron* **84**, 1329–1342 (2014).

[39] Navajas, J., Bahrami, B. & Latham, P. E. Post-decisional accounts of biases in confidence. *Curr. Opin. Behav. Sci.* **11**, 55–60 (2016).

[40] 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* **8**, 7.1–10 (2008).

[41] Aitchison, L., Bang, D., Bahrami, B. & Latham, P. E. Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making. *PLoS Comput. Biol.* **11**, e1004519–23 (2015).

[42] Orhan, A. E. & Ma, W. J. The Inevitability of Probability: Probabilistic Inference in Generic Neural Networks Trained with Non-Probabilistic Feedback. *arXiv* (2016). 1601.03060v1.

[43] Marr, D. *Vision.* A Computational Investigation Into the Human Representation and Processing of Visual Information (MIT Press, 1982).

[44] Rausch, M. & Zehetleitner, M. Visibility Is Not Equivalent to Confidence in a Low Contrast Orientation Discrimination Task. *Front. Psychol.* **7**, 47–15 (2016).

[45] Fleming, S. M., Huijgen, J. & Dolan, R. J. Prefrontal Contributions to Metacognition in Perceptual Decision Making. *J. Neurosci.* **32**, 6117–6125 (2012).

[46] Pelli, D. G. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* **10**, 437–442 (1997).

[47] Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10**, 433–436 (1997).

[48] Girshick, A. R., Landy, M. S. & Simoncelli, E. P. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nat. Neurosci.* **14**, 926–932 (2011).

[49] Acerbi, L., Wolpert, D. M. & Vijayakumar, S. Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS Comput. Biol.* **8**, e1002771 (2012).

[50] Neal, R. M. Slice sampling (2003).

[51] Gelman, A., Hwang, J. & Vehtari, A. Understanding predictive information criteria for Bayesian models. *arXiv* (2013). 1307.5928v1.

[52] van den Berg, R., Awh, E. & Ma, W. J. Factorial comparison of working memory models. *Psychol. Rev.* (2014).

[53] Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *NeuroImage* **46**, 1004–1017 (2009).

[54] Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).

[55] Beck, J. M. *et al.* Marginalization in neural circuits with divisive normalization. *J. Neurosci.* **31**, 15310–15319 (2011).