nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-15581-6 OPEN

Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis

Hsin-Hung Li [1]✉ & Wei Ji Ma [1,2]

Decision confidence reflects our ability to evaluate the quality of decisions and guides subsequent behavior. Experiments on confidence reports have almost exclusively focused on two-alternative decision-making. In this realm, the leading theory is that confidence reflects the probability that a decision is correct (the posterior probability of the chosen option). There is, however, another possibility, namely that people are less confident if the best two options are closer to each other in posterior probability, regardless of how probable they are in absolute terms. This possibility has not previously been considered because in two-alternative decisions, it reduces to the leading theory. Here, we test this alternative theory in a three-alternative visual categorization task. We found that confidence reports are best explained by the difference between the posterior probabilities of the best and the next-best options, rather than by the posterior probability of the chosen (best) option alone, or by the overall uncertainty (entropy) of the posterior distribution. Our results upend the leading notion of decision confidence and instead suggest that confidence reflects the observer's subjective probability that they made the best possible decision.

[1] Department of Psychology, New York University, New York, NY, USA. [2] Center for Neural Science, New York University, New York, NY, USA. ✉email: hsin.hung.li@nyu.edu

Confidence refers to the "sense of knowing" that comes with a decision. Confidence affects the planning of subsequent actions after a decision[1,2], learning[3], and cooperation in group decision making[4]. Failures of utilizing confidence information have been linked to psychiatric disorders[5].

While human observers can report their self-assessment of the quality of their decisions[6–12], the computations underlying confidence reports are still insufficiently understood. The leading theory of confidence suggested that confidence reflects the probability that a decision is correct[7,8,13–17]. We refer to this idea as the "Bayesian confidence hypothesis", meaning that the decision-makers use the posterior probability of the chosen category (i.e. the subjective probability that decision is correct) for their confidence reports. Accordingly, in neurophysiological studies, a brain region or a neural process is considered to represent confidence if its responses correlate with the probability that a decision is correct[18–20]. Behavioral studies testing whether human confidence reports follow Bayesian confidence hypothesis have shown mixed results: While some studies found resemblances between Bayesian confidence and empirical data[18,19,21,22], others have suggested that confidence reports deviate from the Bayesian confidence hypothesis[23–25].

Even though the Bayesian confidence hypothesis is the leading theory of confidence, there is currently no evidence to rule out the possibility that confidence is affected by the probability of correct of the unchosen options. Specifically, people could be less confident if the next-best option is very close to the best option. In other words, confidence could depend on the *difference* between the posterior probabilities of the best and the next-best options, rather than on the absolute value of the posterior of the best option. The reason that this idea has not been tested before might be that previous studies of decision confidence predominantly used two-alternative decision tasks; in such tasks, the alternative hypothesis is equivalent to the Bayesian confidence hypothesis, because the difference between the two posterior probabilities in a two-alternative task is a monotonic function of the highest posterior probability. Thus, to dissociate these two models of confidence, we need more than two alternatives. Here, we use a three-alternative decision task. To preview our main result, we find that the difference-based model accounts well for the data, whereas the model corresponding to the Bayesian confidence hypothesis and a third, entropy-based model do not.

To investigate the computations underlying confidence reports in the presence of multiple alternatives, we designed a three-alternative categorization task. On each trial, participants viewed a large number of exemplar dots from each of the three categories (color-coded), along with one target dot in a different color (Fig. 1a). Each category corresponded to an uncorrelated, isotropic Gaussian distribution in the plane. We asked participants to regard the stimulus as a bird's eye view of three groups of people. People within a group wear shirts of the same color, and the target dot represents a person from one of the three groups. Participants made two responses: the category of the target, and their confidence in their decision on a four-point Likert scale.

To manipulate participants' beliefs (posterior probability distribution), we used different configurations of the category distributions and varied the position of the target dot within each configuration (Fig. 1b, c). This design allowed us to test quantitative models of how the posterior distribution gives rise to confidence reports (see an illustration of this idea in Supplementary Fig. 1).

Results

Model. Generative model. Each category is equally probable. We assume that the observer makes a noisy measurement $\boldsymbol{x}$ of the position $\boldsymbol{s}$ of the target dot. We model the noise as obeying an isotropic Gaussian distribution centered at the target dot.

Decision model: We now consider a Bayesian observer. We assume that the observer knows that each category is equally probable ($p(C) = 1/3$), and knows the distribution associated with each category (group) based on the exemplar dots. Given a measurement $\boldsymbol{x}$, the posterior probability of category C is then

$$p(C|\mathbf{x}) = \frac{p(\mathbf{x}|C)}{\sum_{C'=1}^{3} p(\mathbf{x}|C')} \quad (1)$$

We further assume that due to decision noise or inference noise, the observer does not maintain the exact posterior

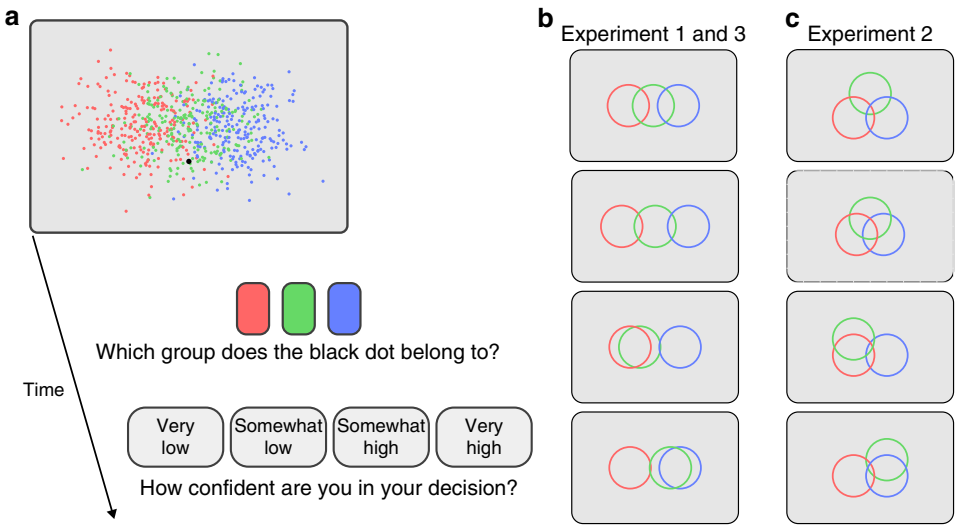


Fig. 1 Experimental procedure and stimuli. a Each trial started with the presentation of the stimulus including exemplar dots in three different colors representing the distribution of each of the three categories and one target dot, the black dot. Observers first reported their decisions in the categorization task and then reported their confidence by using the rectangular buttons presented at the bottom of the screen. **b**, **c** Schematic representation of the distribution of the categories. The circles are centered at the mean location of each category. The width of the circles corresponds to 2.5 times the standard deviation of the category distribution. (**b**) The four conditions tested in Experiment 1 and 3. **c** The four conditions tested in Experiment 2. The exemplar dots in (**a**) are based on the distribution depicted in the top panel in (**b**).

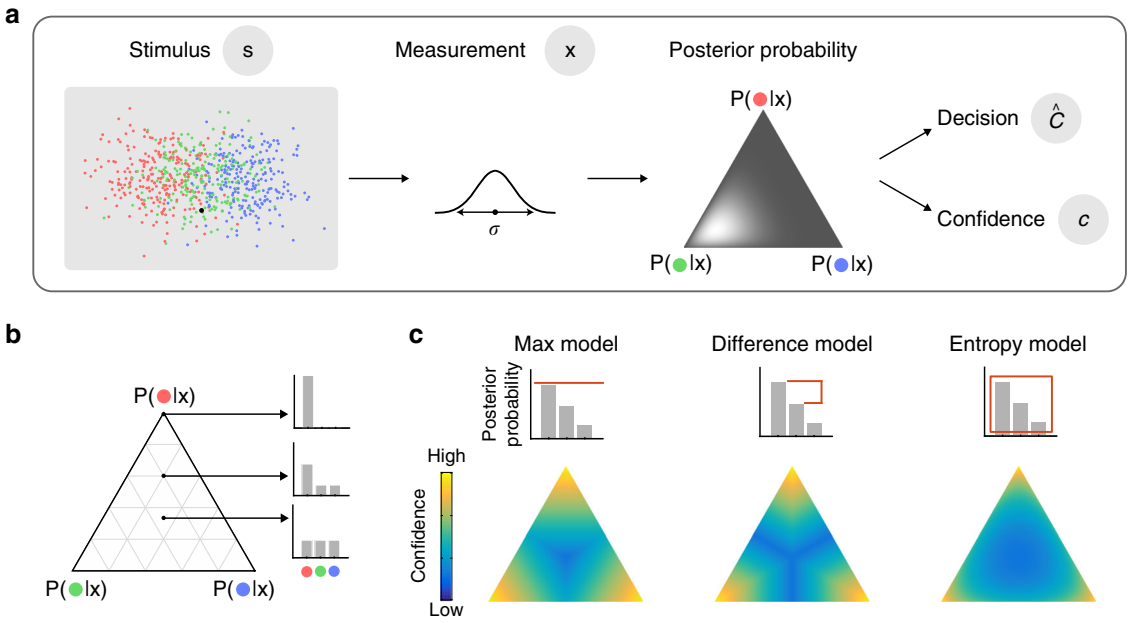


Fig. 2 Models. a Generative model. Target position is represented by **s**. Two sources of variability are considered in the model: First, observers have access to noisy measurement **x**, a Gaussian distribution centered at **s** with a standard deviation σ. Second, given the same measurement **x**, the posterior distribution varies across trials due to decision noise, modeled by Dirichlet distribution, of which spread (represented by the shade of the ternary plot) is controlled by a parameter α (see Methods). On each trial, a decision $\hat{C}$ and a confidence c are read out from the posterior distribution of that trial. **b** We use ternary plots to represent all possible posterior distributions. For example, a point at the center represents a uniform posterior distribution; at the corners of the ternary plot, the posterior probability of one category is one while the posterior for the other two categories are zeros. **c** The bar graphs illustrate how confidence is read out from posterior probabilities in each model. For the purposes of these plots, we did not include decision noise here. The color of each ternary plot represents the confidence as a function of posterior distribution for each model. The color is scaled for each ternary plot (independently) to take the whole range of the color bar.

distribution, $p(C \mid \mathbf{x})$, but instead a noisy version of it. This type of decision noise is consistent with the notion that a portion of variability in behavior is due to "late noise" at the level of decision variable[26–28]. We modeled decision noise by drawing a noisy posterior distribution from a Dirichlet distribution around the true posterior (Fig. 2a, b; See details in Methods). In our case, the true posterior, which we denote by $\mathbf{p}$, consists of the three posterior probabilities from Eq.(1): $\mathbf{p} = (p(C=1 \mid \mathbf{x}), p(C=2 \mid \mathbf{x}), p(C=3 \mid \mathbf{x}))$. The magnitude of decision noise, the amount of variation around $\mathbf{p}$, is (inversely) controlled by a concentration parameter $\alpha>0$. When $\alpha \to \infty$, the variation vanishes and the posterior is noiseless. In general, the "noisy posterior", which we denote as a vector $\mathbf{q}$, satisfies $\mathbf{q} \sim \text{Dirichlet}(\alpha \mathbf{p})$. We assume that when reporting the category of the target, the observer chooses the category C with the highest $q(C \mid \mathbf{x})$. Unless otherwise specified, we will from now on refer to the noisy posterior distribution as simply the posterior distribution.

We introduce three models of confidence reports: the Max model, the Entropy model and the Difference model. Each of these models contains two steps: (a) mapping the posterior distribution ($\mathbf{q}$) to a real-valued internal confidence variable; (b) applying three criteria to this confidence variable to divide its space into four regions, which, when ordered, map to the four confidence ratings. The second step accounts for every possible monotonic mapping from the internal confidence variable to the four-point confidence rating. The three models differ only in the first step.

The Max model corresponds to the Bayesian confidence hypothesis. In this model, the confidence variable is the probability that the chosen category is correct, or in other words, it is the highest of the three posterior probabilities (Fig. 2c). In this model, the observer is least confident when the posterior distribution is uniform. Importantly, after the posterior distribution is computed, the posterior probability of the unchosen options does not further contribute to the computation of confidence.

In the Difference model, the confidence variable is the difference between the highest and second-highest posterior probabilities. In this model, confidence is low if the evidence for the next-best option is strong, and the observer is least confident whenever the two most probable categories are equally probable. One interpretation of this model is that confidence reflects the observer's subjective probability that they made the best possible choice, regardless of the actual posterior probability of that choice. An alternative interpretation is that decision-making consists of an iterative process in which the observer reduces a multiple-alternative task to simpler (two-alternative) tasks (see the Discussion section). (Note that a model that uses the difference of the probability of the best option and the average of the non-chosen options is equivalent to the Max model.)

In the Entropy model, the confidence variable is the negative of the uncertainty conveyed by the entire posterior distribution, as quantified by its negative entropy. High confidence is associated with low entropy, and vice versa. Like in the Max model, the observer is least confident when the posterior distribution is uniform. Unlike in the Max model, however, the posterior probabilities of the non-chosen categories directly affect confidence. For the details of the models, see Methods.

All three models are Bayesian in the sense that they compute the posterior probability distribution and categorize the target dot into the category with the highest posterior. Thus, in all three models, the unchosen options "implicitly" affect confidence by contributing to the denominator in the computation of the posterior probabilities. In the Discussion, we discuss a model in which the unchosen option (e.g., the least probable category) is disregarded before even contributing to the normalization of the posterior. The three models differ in how the confidence variable

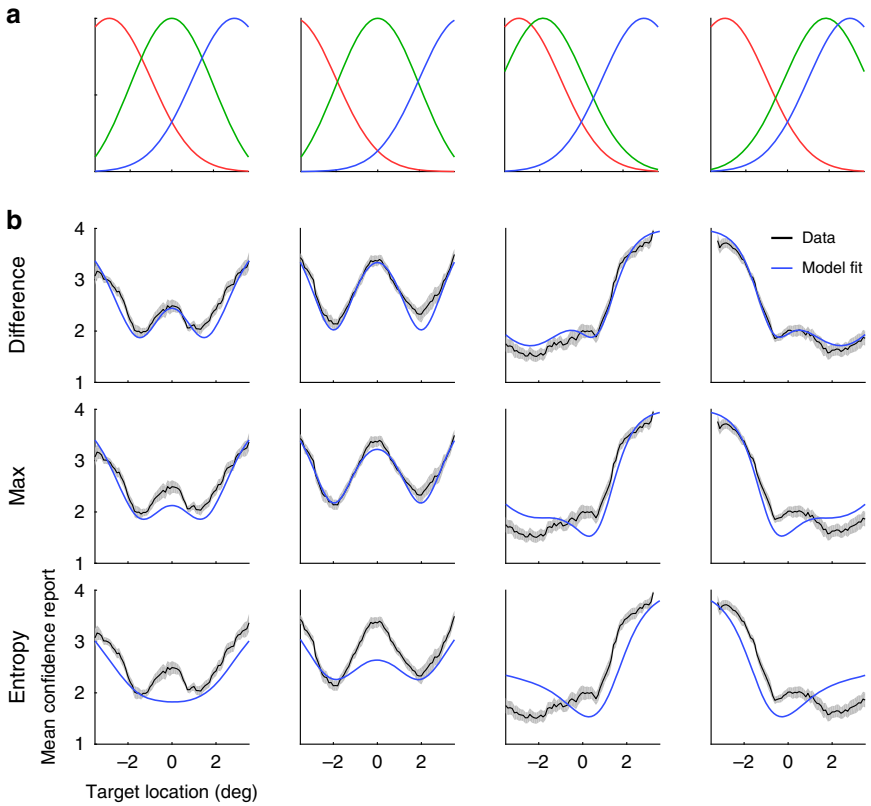


Fig. 3 Experiment 1. a The distribution of the reference dots in each condition. **b** Mean confidence report as a function of target position for each of the four conditions. The black curves represent group mean ± 1s.e.m. Blue curves represent the model fit averaged across individuals.

is read out from the posterior distribution. The Max model is unique in assuming that after the computation of the posterior probabilities, the unchosen categories do not further affect the computation of confidence.

In our three-alternative task, these models generate qualitatively different mappings from the posterior distribution to the confidence variable (Fig. 2c). In a standard two-alternative task, however, the models would have been indistinguishable, because the probability of the chosen category would have determined the probability of the non-chosen category.

The Max, Difference and Entropy models are our three main models. So far, the sources of variability in these models are sensory noise (Sen) and Dirichlet decision noise (Dir). We name the corresponding models Max-Sen-Dir, Diff-Sen-Dir, Ent-Sen-Dir models in the supplementary figures and supplementary tables in order to distinguish them from model variants that consider different sources of variability (introduced later).

We fitted the free parameters to the data of each individual subject using maximum-likelihood estimation, where the data on a given trial consist of a decision-confidence pair. Thus, we accounted for the joint distribution of decisions and confidence ratings[24,25,29] (see Methods). We compared models using the Akaike Information Criterion (AIC[30]). A model recovery analysis suggests that if the true model is among our tested models, our model comparison procedure is able to identify the correct model (see Methods and Supplementary Fig. 3).

Experiment 1. In Experiment 1, the centers of the three category distributions were aligned vertically (Fig. 1b). There were four conditions: In the first two conditions, the centers were evenly spaced horizontally. In the last two conditions, the center of the central distribution was closer to the center of either the left or the right distribution. The vertical position of the target dot was sampled from a normal distribution, and the horizontal position of the target dot was sampled uniformly between the center of the leftmost and right-most classes plus an extension to the left and the right (see Methods).

We plotted the psychometric curves (mean confidence report as a function of the horizontal position of the target dot) by averaging confidence reports across trials using a sliding window (Fig. 3). Mean confidence report varied as a function of the horizontal position of the target. In the first two conditions (Fig. 3), where the three distributions were evenly spaced, the psychometric curves showed two dips, with the lowest confidence attained at two positions symmetric around 0°.

We simulated the predicted psychometric curves using the best-fitting parameters of each model (Fig. 3b). The fits of the Max and the Difference models resembled the data, but the best fit of the Entropy model showed a dip at the center in the first condition.

In the third and fourth conditions, in which the three distributions were unevenly spaced, mean confidence was lowest around the centers of the two distributions that were closest to each other. Only the Difference model exhibited this pattern, while the Max and the Entropy models deviated more clearly from the data.

The models not only make predictions for confidence reports, but also for the category decisions (Supplementary Fig. 2). Participants categorized the target dot based on its location, and when the target dot was close to the boundary between two neighboring categories (the location where two categories have equal likelihood), they assigned the target to those two neighboring categories with nearly equal probabilities. In general, this pattern is consistent with an observer who chooses the category associated with the highest posterior probability. The Entropy model fits worst, even though all three models used the

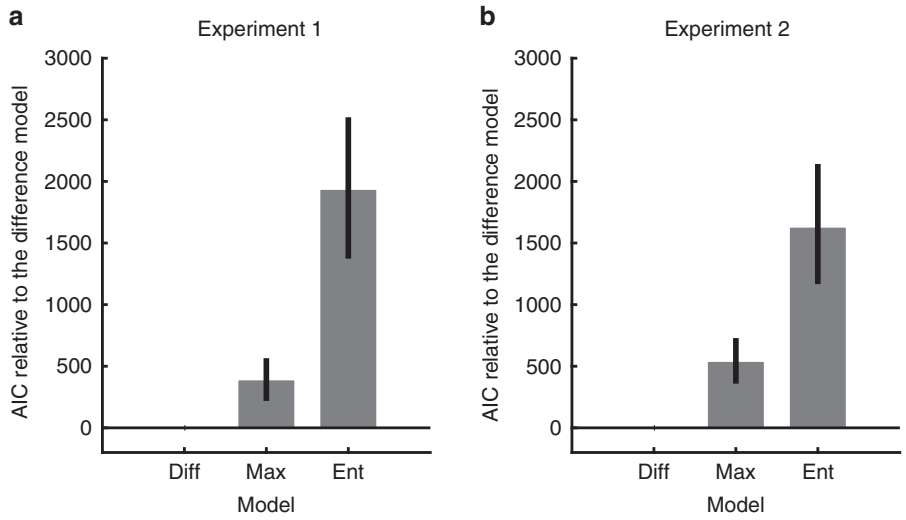


Fig. 4 Model comparisons using ΔAIC: AIC of each model compared with the Difference model. The bars represent ΔAIC summed across participants. The error bars represent 95% bootstrapped confidence interval. **a** Experiment 1. **b** Experiment 2.

same rule for the category decision; this is because the confidence data also need to be accounted for. The Difference model outperformed the Max model by a group-summed AIC score of 391 (95% CI [222, 569]) and the Entropy model by 1937 (95% CI [1363, 2562]) (Fig. 4a and Supplementary Table 1).

We further tested reduced versions of each of the three confidence models by removing either the sensory noise or the decision noise from the model. The Difference model outperformed the Max model and the Entropy model regardless of these manipulations (Supplementary Fig. 4A and Supplementary Table 1). The sensory noise played a minor role in this task compared to the decision noise. For example, removing the sensory noise from the Difference model increased the AIC by 121 (95% CI [48, 199]), while removing the inference noise increased the AIC by 737 (95% CI [590, 914]). Using the Bayesian information criterion (BIC)[31] for model comparison led to the same conclusions (Supplementary Fig. 5A and Supplementary Table 2).

So far, we jointly fitted the category decision and confidence reports. One could wonder whether independently fitting the confidence reports would lead to different results. We found the same results when only fitting the confidence reports: The Difference model outperformed the other two models, and the decision noise had a stronger influence on the model fit (Supplementary Figs. 6 and 7). Because the Max, the Difference and the Entropy used the same rule for category decisions, we compared category decision models that used the same decision rule (reporting the category with the highest posterior probability), but included sensory noise only, decision noise only, or both. We fitted the category decisions alone and found that the models including the decision noise fit the data better than the model with the sensory noise alone (Supplementary Figs. 8 and 9). This is similar to the results obtained by fitting the confidence reports alone or by jointly fitting both category decisions and confidence reports.

We tested various alternative models (see details in Supplementary Information). We found that the Difference model outperformed the Max and the Entropy models when we replaced Dirichlet decision noise by drawing samples from the true posterior, or when we added noise in the measurement of the category means (Supplementary Fig. 10A). In addition, we tested heuristic models that made category decisions and confidence reports based on the category means and the noisy measurement of the target location ($\mathbf{x}$) but did not compute posterior probabilities. Still, the heuristic models did not fit the data better than the Difference model (Supplementary Fig. 10A).

Experiment 2. In Experiment 2, we aimed to test whether the findings in Experiment 1 could be generalized to other stimulus configurations, where the centers of the categories varied in a two-dimensional space. We tested four conditions in which the centers of the three groups varied along both horizontal and vertical axis (Fig. 1c). We sampled the target dot positions uniformly within a circular area centered on the screen. In addition, the distribution of the categories used in Experiment 2 allowed us to probe confidence reports in a wider range of posterior distributions (Supplementary Fig. 1B). For example, we can probe the confidence report when the target dot had the same distance to all three categories in Experiment 2, but not in Experiment 1.

The "psychometric curve" is now a heat map in two dimensions (Fig. 5). The fits to these psychometric curves showed different patterns among the three models: When the three groups formed an equilateral triangle (Fig. 5, the first and second columns), the confidence (as a function of target location) estimated by the Entropy model exhibited contours that were more convex than that in the data. In the last two conditions (Fig. 5, the third and fourth columns), compared to the other two models, the Difference model showed stronger resemblance to the data, as the model exhibited an extended low confidence region at the side where two categories were positioned closely. The results of model comparisons were consistent with Experiment 1. The Difference model outperformed the Max model by a group-summed AIC score of 541 (95% CI [371, 735]) and the Entropy model by 1631 (95% CI [1179, 2159]) (Fig. 4b). The model with both sensory and inference noise explained the data the best, and the inference noise had a stronger influence on the model fit than the sensory noise (Supplementary Fig. 4B, Supplementary Fig. 5B, Supplementary Tables 1 and 2).

Consistent with Experiment 1, we found that the Difference model outperformed the Max and the Entropy model when we only fitted the confidence reports (Supplementary Fig. 6B). Models that considered other sources of variability or used heuristic decision rules did not perform better than the Difference model (Supplementary Fig. 10B).

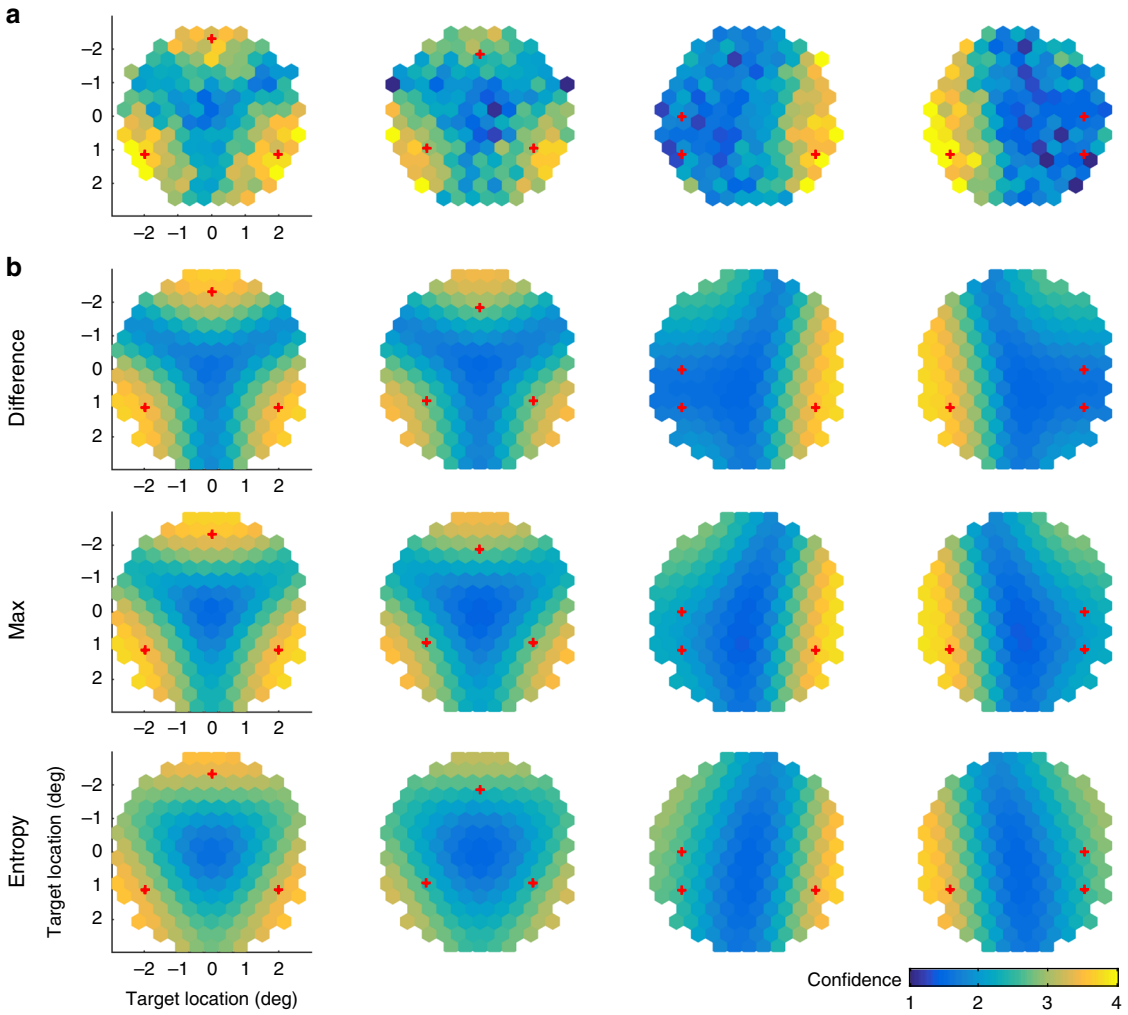


Fig. 5 Experiment 2. a The mean confidence report as a function of target positions. **b** Model fit averaged across individuals. The red crosses in each panel represent the center of each of the three categories.

Experiment 3. So far, we found that the Difference model fits the data better than the Max model and the Entropy model. However, whether participants report the probability that a decision is correct (the Max model) might depend on the experimental design. In Experiment 1 and 2, participants received no feedback on their category decision. Thus, the probability of being correct in the task could be difficult to learn. To investigate this issue, in Experiment 3, using the same four stimulus configurations as those in Experiment 1 (Fig. 1b), we randomly chose one of the three groups as the true target category in each trial, and sampled the target position from the distribution of the true category. Feedback was presented at the end of each trial, informing participants of the true category.

The results of model comparison were consistent with Experiment 1 and 2. The Difference model outperformed the Max model by a group-summed AIC score of 100 (95% CI [46, 156]) and the Entropy model by 1113 (95% CI [817, 1447]) (Supplementary Fig. 11, Supplementary Tables 1 and 2). The model with both sensory and inference noise explained the data the best, and the inference noise had a stronger influence on the model fit than the sensory noise (Supplementary Figs. 4C and 5C). These results held when we fitted the confidence reports alone, or when other sources of variability were considered (Supplementary Figs. 6C and 10C). Heuristic models did not fit the data better than the Difference model (Supplementary Fig. 10C).

Discussion

To distinguish the leading model of perceptual confidence (the Bayesian confidence hypothesis) from a new alternative model in which confidence is affected by the posterior probabilities of unchosen options, we studied human confidence reports in a three-alternative perceptual decision task. We found that confidence is best described by the Difference model, in which confidence reflects the difference between the strength of observers' belief (posterior probability) of the top two options in a decision. The Max model (which corresponds to the Bayesian confidence hypothesis) and the Entropy model (in which confidence is derived from the entropy of the posterior distribution) fell short in accounting for the data. Our results were robust under changes of stimulus configurations (Experiment 1 and 2), and when trial-by-trial feedback was provided (Experiment 3). Our results demonstrate that the posterior probabilities of the unchosen categories impact confidence in decision-making.

Decision tasks with multiple alternatives not only allow us to dissociate different computational models of confidence, they are also ecologically important. In the real world, human and other animals often face decisions with multiple alternatives, such as identifying the color of a traffic light, recognizing a person, categorizing a species of an animal, or making a medical diagnosis.

Our models can be generalized to categorical choice with more than three alternatives. Specifically, the Difference model predicts that besides the posterior probabilities of the top two options, the posterior of the other options does not matter as long as they add up to the same total. A special type of categorical choice is when the world state variable is continuous (e.g., in an orientation estimation task) but gets discretized for the purpose of the experiment. Consider the specific case that the posterior distribution is Gaussian. An observer following the Difference model would compute the difference between the posteriors of the two discrete options closest to the peak. This serves as a coarse approximation to the curvature of the posterior distribution at its peak, which, for Gaussians, is monotonically related to its inverse variance, consistent with an earlier model by van den Berg et al.[29], in which confidence is based on the precision parameter of the posterior in continuous estimation tasks[29]. Outside the realm of Gaussian and similar distributions, the Difference model and van den Berg et al.[29]'s model might be distinguishable. For example, when the posterior distribution is bimodal, with the modes slightly different in height, the variance of the posterior is dominated by the separation between the modes, whereas the Difference model will use the difference in height for confidence reports.

Although many behavioral studies have emphasized similarities between human confidence reports and predictions of Bayesian models e.g[18,21,22]., the Bayesian confidence hypothesis has been questioned before[8,13–16]. In addition to the probability of being correct, confidence is influenced by various factors such as reaction time[32], post-decision processing[33–36], and the magnitude of positive evidence[37–40]. Two model comparison studies have shown deviations from Bayesian confidence hypothesis in two-alternative decision tasks[24,25]. However, in one study[24], the experimental design did not allow the authors to strongly distinguish the model that was based on Bayesian confidence hypothesis from those that were not. Moreover, in both studies[24,25], the alternative models were based on heuristic decision rules without a broader theoretical interpretation. Here, we have identified a type of deviation from the Bayesian confidence hypothesis that is not only of a qualitatively different nature, but that also raises new theoretical questions.

Specifically, the Difference model is currently a descriptive model. We have two suggestions to interpret it as an outcome of approximate inference. First, the Difference model might be an approximation to a model in which confidence depends on the probability that an observer made the best possible decision. In this view, the observer possesses metacognitive knowledge that their decision is based on the noisy posterior $\mathbf{q}$ rather than the true posterior $\mathbf{p}$, and consequently, realizes that it is possible that the chosen category is not the category with the highest true posterior probability. Confidence would then be derived from the probability that the chosen category has the highest true posterior probability. The stronger the evidence for the next-best option, the less likely is the case, which would lead to lower confidence. This interpretation is consistent with recent work that showed that subjective confidence guides information seeking during decision-making[41]. Under the Difference model, during information seeking, the observer's goal is to make sure that the best option is better than the alternative options. Low confidence would encourage the observer to collect more information in order to strengthen the belief that the best option is better than the next-best option.

Second, the finding that confidence is best described by the relative strength of the evidence of the top two options might be related to other findings in multiple-alternative decision-making. In one experiment, the observer watched columns of bricks build up on the screen, and reported which column had the highest accumulation rate[42]. A heuristic model in which the observer makes a decision when the height of the tallest column exceeds the height of the next-tallest column by a fixed threshold captured the overall pattern of people's behavior. In a study on self-directed learning in a three-alternative categorization task, observers had to learn the category distributions by sampling from the feature space and receiving feedback. Instead of choosing the most informative samples, human observers chose ones for which the likelihood of two categories were similar, namely those located at boundaries between pairs of two categories[43]. This literature allows us to speculate that observers might decompose a multiple-alternative decision into several simpler (perhaps two-alternative) decisions. This notion is reminiscent of the concept in prospect theory that before a phase of evaluation, extremely unlikely outcomes might be first discarded in an "editing" phase[44]. Hence, an alternative interpretation of our results is that confidence reports deviate from the Bayesian confidence hypothesis (the Max model) because the observer estimates the probability of correct in a way that ignores the options that are discarded before final evaluation. In the Difference model, the least favorite option is not completely discarded because it decreases the posterior probabilities of the other two options (and thus their difference) by contributing to the normalization pool[45,46]. Therefore, we consider an extreme version of editing, the Ratio model, in which the least-favorite option does not even participate in normalization, and thus confidence solely depends on the likelihood ratio between the top two options. The Difference model and the Ratio model are not distinguishable in Experiment 1 and 2 (Supplementary Fig. 12). In Experiment 3, the Difference model has a slight advantage over the Ratio model by a group-summed AIC of 51 (95% CI [18, 90]). Testing variable numbers of categories within an experiment might help to differentiate between these two models.

We found that compared to the sensory noise, the noise associated with the computation of posterior probability plays a more important role in our task. This is consistent with the findings of a recent study[26]. The relative unimportance of sensory noise could be partly due to our experimental design, which used stimuli with strong signal strength (saturated color and unlimited duration). Differently from our study, Drugowitsch et al.[26] used an evidence accumulation task and further distinguished two types of decision noise: inference noise that was added with each new stimulus sample, and selection noise that was injected only once, right before the final response. Because our experiment only had one stimulus in each trial, it was not set up to distinguish these two sources of variability. While modeling decision noise using Dirichlet distribution was successful, we found that models in which the category mean is not known exactly but measured in a noisy fashion also fit the data in two experiments quite well (Diff-Sen-Mean model in Experiment 1 and Experiment 2; Supplementary Fig. 10). This is consistent with a recent finding that imperfect knowledge about the experimental parameters explained a significant portion of the behavioral variability in two-alternative decision tasks[47].

Across the three experiments, we did not find evidence that any of the heuristic models we tested outperformed the Difference model. In only one experiment, a heuristic model was indistinguishable from the Difference model (DistW-Sen-Mean model in Experiment 2; Supplementary Fig. 10B). These results are different from a recent study reporting that some probabilistic but heuristic models outperformed the Bayesian models in fitting confidence data in two-alternative tasks[25]. The causes of this discrepancy may lie in the experimental design. Adler and Ma manipulated different levels of uncertainty by presenting brief stimuli (50 ms) at various contrast levels. To perform the task optimally, observers had to track the sensory uncertainty varied

in a range (e.g., from 0.4 to 13% contrast in their main experiment) in a trial-by-trial fashion. Instead, we purposely reduced uncertainty by presenting all the stimuli at the highest achievable contrast with unlimited duration. In addition, the distributions of the categories were not presented explicitly in Adler and Ma's study whereas the distributions of the categories were presented throughout each trial in the present study. These factors may contribute to the fact that heuristic models performed better in Adler and Ma's study but not in our experiments.

The ΔAIC (relative to the Difference model) was smaller in Experiment 3 than in Experiment 1 and 2. Intuitively, this is not surprising, since when directly sampling from the stimulus distributions in Experiment 3, there were more target dots positioned at the far left and far right, compared to the target dots in Experiment 1 and 2. These trials would not have been informative to distinguish the models. All three models have the same prediction (high confidence) for far-left and the far-right locations, and this may lead to a smaller ΔAIC. To examine whether stimulus selection alone could account for the smaller AIC differences, we performed a model recovery analysis. We synthesized data based on the observers' best-fitted parameters of the Max, Difference and Entropy models, and we fitted the synthesized data with these three models. We found that the performance of the Max and the Difference models are closer in Experiment 3 than in Experiment 1 and 2, similar to the real data (Supplementary Fig. 3). Thus, stimulus selection alone can account for the smaller AIC differences.

Whereas we propose a theoretical framework for how decisions and confidence reports are computed in multi-alternative tasks, we are agnostic about how the decision-making process unfolds over time. Other models exist that consider the temporal dynamics of decision-making. In particular, drift-diffusion models and race models jointly account for accuracy and reaction times in many tasks[48]. Some studies have employed such accumulation models to account for confidence judgments[34,49–52]. However, these studies only considered confidence judgments in two-alternative decision tasks. Conceptually, our findings might be related to the "balance of evidence" (BoE) in Vickers and colleagues' work[51,53]. In a race model with two accumulators, they suggested that confidence is computed as the difference between the accumulated evidence of the two accumulators[51]. Vickers and Lee suggested that in theory, this idea could be extended to three-alternative tasks, but they speculated that confidence in the chosen category (option A) might be computed as the average of the confidence in comparing option A to option B and the confidence in comparing option A to option C[53]. This algorithm is more similar to the Max model than to the Difference model here: Assuming that A, B and C represent the evidence accumulated for each of the three categories, and A is the chosen category, confidence is computed as $c^*=((A-B)+(A--C))/2=(3A-1)/2$. Then, confidence only depends on the chosen category A. It remains to be seen whether evidence accumulation models designed to explain decisions in multiple-alternative tasks (review in refs. [57,58]) could be extended to generate confidence reports that are consistent with our data and with the Difference model.

Do our results generalize beyond perceptual decision-making? In a two-alternative value-based decision task, observers reported confidence in a way that was similar to that in perceptual decision tasks:[10] When observers were asked to choose the good with the higher value, confidence increased with the posterior probability that a decision is correct, which in turn increased with the difference in value between the two goods. In addition, choice accuracy was higher in high-confidence trials then in low-confidence trials, reflecting observers' ability to evaluate their own performance. A recent study also reported that observers are able to reflect on their decisions and report confidence in three-alternative value-based decision tasks[54]. Given that the computation of subjective value may involve a Bayesian inference process similar to that in perception[12], it might be worth investigating whether confidence reports in multiple-alternative value-based decisions also deviate from the Bayesian confidence hypothesis. The Difference model would predict that, confidence is determined by the difference between the probability that the chosen item is the most valuable and the probability that the next-best item is the most valuable.

How does the present study advance our understanding of the neural basis of confidence? Most neurophysiological studies of confidence have considered the neural activity that correlates with the probability of being correct as the neural representation of confidence (but see ref. [55]). Neural responses in parietal cortex[19], orbitofrontal cortex[18] and pulvinar[20] have been associated with that representation of confidence. These studies all used two-alternative decision tasks. Multiple-alternative decision tasks have been used in neurophysiological studies on non-human primates but not with the objective of studying confidence[46,56–58]. By utilizing multiple-alternative tasks, neural studies could dissociate the neural correlates of probability correct from that of the "difference" confidence variable in the Difference model, which according to our results, might be the basis of human subjective confidence. A potentially important difference between human and non-human animal studies is that in the latter, confidence is not explicitly reported but operationalized through some aspect of behavior, such as the probability of choosing a "safe" (opt-out) option[19,20,55,59,60], or the time spent on waiting for reward[18]. Thus, one should be careful when directly comparing these implicit reports with explicit confidence reports in human studies.

Methods

Setup. Participants sat in a dimly lit room with the chin rest positioned 45 cm from the monitor. The stimuli and the experiment were controlled by customized programs written in Javascript. The monitor had a resolution of 3840 by 2160 pixels and a refresh rate of 30 Hz. The spectrum and the luminance of the monitor were measured with a spectroradiometer.

Participants. Thirteen participants took part in Experiment 1. Eleven participants took part in Experiment 2. Eleven participants took part in Experiment 3. All participants had normal or corrected-to-normal vision. The experiments were conducted with the written consent of each participant. The University Committee on Activities involving Human Subjects at New York University approved the experimental protocols.

Stimulus. On each trial, three categories of exemplar dots (375 dots per category) were presented along with one target dot, a black dot (Fig. 1a). The exemplar dots within a category were distributed as an uncorrelated, circularly symmetric Gaussian distribution with a standard deviation of 2° (degree visual angle) along both horizontal and vertical directions. Exemplar dots from the different categories were coded with different colors. The three colors were randomly chosen on each trial, and were equally spaced in Commission Internationale de l'Eclairage (CIE) $L^*a^*b^*$ color space. The three colors were at a fixed lightness of $L^*=70$ and were equidistant from the gray point ($a^*=0$, and $b^*=0$).

In Experiment 1 and 3, the centers of the three categories were aligned vertically to the center of the screen, and were located at different horizontal positions (Fig. 1b). In four configurations, the horizontal positions of the centers of the three categories were (−3°, 0°, 3°), (−4°, 0°, 4°), (−3°, −2°, 3°), and (−3°, 2°, 3°), from the center of the screen respectively. In Experiment 2, the centers of the three categories varied on a 2-dimensional space (Fig. 1c). In four configurations, the horizontal positions of the centers of the three categories were (−2°, 0°, 2°), (−1.59°, 0°, 1.59°), (−2°, −2°, 2°), and (−2°, 2°, 2°), from the center of the screen, respectively. The vertical positions of the centers were (1.16°, −2.31°, 1.16°), (0.94°, −1.84°, 0.94°), (1.16°, 0°, 1.16°), and (1.16°, 0°, 1.16°) from the center of the screen respectively.

Procedures. We told participants that the three groups of exemplar dots represented a bird's eye view of three groups of people. The three groups contained equal numbers of people. The black dot (the target) is a person from one of the three groups, but we do not know the color of her/his T-shirt. We asked participants to categorize the target to one of the three groups based on the (position)

information conveyed by the dots, and report their confidence on a four-point Likert scale.

Each trial started with the onset of the stimulus and three rectangular buttons positioned at the bottom of the screen (Fig. 1a). On each trial, participants first categorized the target to one of the three groups (based on the position information conveyed by the dots) by using the mouse to click on one of the three buttons. After participants reported their decision, the three buttons were replaced by four buttons (labeled as "very unconfident", "somewhat unconfident", "somewhat confident", and "very confident") for participants to report their confidence on the decision they made. The stimuli were presented throughout each trial. Reaction time (for both category decision and confidence reports) was unlimited. After participants reported their confidence, all the exemplar dots and the rectangular buttons disappeared from the screen, and the next trial started after a 600 ms inter-trial-interval.

In Experiment 1, the vertical position of the target dot was sampled from a normal distribution (2° std), and the horizontal position of the target dot was sampled uniformly between the center of the leftmost and rightmost categories plus a 0.2° extension to the left and the right. In Experiment 2, the target dot was uniformly sampled from a circular area (2.6° radius) positioned at the center of the screen. No feedback was provided in Experiment 1 and Experiment 2.

In Experiment 3, in each trial, we randomly chose one of the three categories with equal probability as the true category. We then positioned the target dot by sampling from the distribution of the true category. A feedback regarding the true category was provided at the end of each trial: After participants reported their confidence, all exemplar dots disappeared except that the exemplar dots from the true category remained on the screen for an extra 500 ms. In each experiment, participants completed one 1-hr session (84 trials per configuration in Experiment 1 and 120 trials per configuration in Experiment 2 and 3). All the trials in one session were separated into eight blocks with equal number of trials. Different configurations were randomized and interleaved within each block.

Participants were well informed about the structure of the stimuli. We told observers that the distributions of the three groups are circular and symmetric, and the three groups have the same spread (standard deviation) throughout the experiments, and only differed in their centers. In Experiment 1 and 3, participants were informed that the centers of the three groups only varied horizontally.

Models. Generative model. The target belongs to category $C \in \{1, 2, 3\}$. The two-dimensional position $\mathbf{s}$ of a target in category C is drawn from a two-dimensional Gaussian $p(\mathbf{s} \mid C) = N(\mathbf{s}; \mathbf{m}_C, \sigma_s^2\mathbf{I})$, where $\mathbf{m}_C$ is the center of category C, σ_s^2 is the variance of the stimulus distribution, and $\mathbf{I}$ is the two-dimensional identity matrix. We assume that the observer makes a noisy sensory measurement $\boldsymbol{x}$ of the target position. We model the sensory noisy using a Gaussian distribution centered at $\mathbf{s}$ with covariance matrix $\sigma^2\mathbf{I}$. Thus, the distribution of $\boldsymbol{x}$ given category C is $p(\mathbf{x} \mid C) = N(\mathbf{x}; \mathbf{m}_C, (\sigma_s^2 + \sigma^2)\mathbf{I})$.

Inference on a given trial. We assume that the observer knows the mean and standard deviation of each category based on the exemplar dots, and that the observer assumes that the three categories have equal probabilities. The posterior probability of category C given the measurement $\boldsymbol{x}$ is then $p(C \mid \mathbf{x}) \propto p(\mathbf{x} \mid C) = N(\mathbf{x}; \mathbf{m}_C, (\sigma_s^2 + \sigma^2)\mathbf{I})$. Instead of the true posterior $p(C \mid \mathbf{x})$, the observer makes the decisions based on $q(C \mid \mathbf{x})$, a noisy version of the posterior probability. We obtain a noisy posterior $q(C \mid \mathbf{x})$ by drawing from a Dirichlet distribution. The Dirichlet distribution is a generalization of the beta distribution. Just like the beta distribution is a continuous distribution over the probability parameter of a Bernoulli random variable, the Dirichlet distribution is a distribution over a vector that represents the probabilities of any number of categories. The Dirichlet distribution is parameterized as

$$p(\mathbf{q}|\mathbf{p};\alpha) = \frac{1}{B(\alpha\mathbf{p})}\prod_{i=1}^{3} q_i^{\alpha p_i - 1}$$

$$B(\alpha\mathbf{p}) = \frac{\prod_{i=1}^{3}\Gamma(\alpha p_i)}{\Gamma\left(\alpha\sum_{i=1}^{3} p_i\right)}$$

$\boldsymbol{\Gamma}$ represents the gamma function. $\mathbf{p}$ is a vector consisting of the three posterior probabilities, $\boldsymbol{p} = (p_1, p_2, p_3) = (p(C = 1 \mid \mathbf{x}), p(C = 2 \mid \mathbf{x}), p(C = 3 \mid \mathbf{x}))$. $\mathbf{q}$ is a vector consisting of the three posterior probabilities perturbed by decision noise, $\mathbf{q} = (q_1, q_2, q_3) = (q(C = 1 \mid \mathbf{x}), q(C = 2 \mid \mathbf{x}), q(C = 3 \mid \mathbf{x}))$. The expected value of $\mathbf{q}$ is $\mathbf{p}$. The concentration parameter α is a scalar whose inverse determines the magnitude of the decision noise; as α increases, the variance of $\mathbf{q}$ decreases. To make a category decision, the observer chooses the category that maximizes the posterior probability: $\hat{C} = \underset{C}{\operatorname{argmax}}\, q(C|\mathbf{x})$.

We considered three models of confidence reports. We first specify in each model an internal continuous confidence variable c^*. In the Max (maximum a posteriori) model, c^* is the posterior probability of the chosen category: $c^* = q(C = \hat{C}|\mathbf{x})$. In the Difference model, c^* is a difference: $c^* = q(C = \hat{C}|\mathbf{x}) - q(C = \hat{C}_2|\mathbf{x})$, where $\hat{C}_2$ is the category with the second-highest posterior probability. In the Entropy model, c^* is the negative entropy of the posterior distribution: $c^* = \sum_{C=1}^{3} q(C|\mathbf{x}) \log q(C|\mathbf{x})$.

In each model, the internal confidence variable c^* is converted to a four-point confidence report c by imposing three confidence criteria b_1, b_2 and b_3. For example, $c = 3$ when $b_2 < c^* < b_3$. This implementation accommodated any type of mapping between the internal confidence variables c^* and the four-level button press, as long as the reported levels monotonically increased with the internal confidence variables c^*. We also included a lapse rate λ in each model; on a lapse trial, the observer presses a random button for both the decision and the confidence report. In addition to the models that included both sensory and Dirichlet decision noise, we took a factorial approach and tested various combinations of confidence model and sources of variability[61–63]. For each of the three main confidence models (Max, Difference and Entropy), we tested two reduced models by removing either the sensory noise (by setting $\sigma = 0$) or the decision noise (by setting $q(C \mid \mathbf{x}) = p(C \mid \mathbf{x})$) from the model, leading to nine models reported in Supplementary Figs. 4 and 5. In addition, we fitted these nine models with confidence reports only, without jointly fitting the category decisions (Supplementary Figs. 6 and 7). We also fitted the category decisions alone by three different models. These three models all chose the category with the highest posterior probability, but considered different sources of variability (sensory noise only, decision noise only, or both; Supplementary Fig. 8).

In addition to the nine models reported in Supplementary Fig. 4 and Supplementary Fig. 5, we furthermore tested 21 alternative models (Supplementary Fig. 10), including Bayesian models with various sources of variability and heuristic models that made decisions without computing posterior probability. The details of these models are described in Supplementary Information.

Response probabilities. So far, we have described the mapping from a measurement $\mathbf{x}$ to a decision $\hat{C}$ and a confidence report c. The measurement, however, is internal to the observer and unknown to the experimenter. Therefore, to obtain model predictions for a given parameter combination (σ, α, b_1, b_2, b_3, λ), we perform a Monte Carlo simulation. For every true target position $\mathbf{s}$ that occurs in the experiment, we simulated a large number (10,000) of measurements $\mathbf{x}$. For each of these measurements, we compute the posterior $p(C \mid \mathbf{x})$, add decision noise to obtain $q(C \mid \mathbf{x})$, and finally obtain a category decision $\hat{C}$ and a confidence report c. Across all simulated measurements, we obtain a joint distribution $p(\hat{C}, c|\mathbf{s};\sigma, \alpha, b_1, b_2, b_3, \lambda)$ that represents the response probabilities of the observer.

Model fitting and model comparison. We denote the parameters (σ, α, b_1, b_2, b_3, λ) collectively by θ. We fit each model to individual-subject data by maximizing the log likelihood of θ, $\log L(\theta) = \log p(\text{data}|\theta)$. We assume that the trials are conditionally independent. We denote the target position, category response, and four-point confidence report on the ith trial by $\mathbf{s}_i$, $\hat{C}_i$, and c_i, respectively. Then, the log likelihood becomes

$$\log L(\theta) = \log \prod_i p(\hat{C}_i, c_i|\mathbf{s}_i,\theta) = \sum_i \log p(\hat{C}_i, c_i|\mathbf{s}_i,\theta),$$

where $p(\hat{C}_i, c_i|\mathbf{s}_i, \theta)$ is obtained from the Monte Carlo simulation described above. We optimized the parameters for each individual using a new method called Bayesian Adaptive Direct Search[64]. We used AIC for model comparison. To report the AIC, we computed the AIC for each individual and then summed the AIC across participants. The confidence interval of the group-summed AIC was estimated by bootstrapping. We also reported BIC in Supplementary Information.

Parameterization. The three main models (Max, Difference and Entropy models reported in Fig. 4) have the same set of free parameters including the magnitude of sensory noise (σ), the magnitude (concentration parameter) of decision noise (α), three boundaries for converting internal continuous confidence variable to button press (b_1, b_2, b_3) and a lapse rate λ. For each of the three models, we tested two versions of the reduced models (Supplementary Figs. 4–6). In one version, we kept the sensory noise (σ) in the model while removing the decision noise (α). In the other version we kept the decision noise (α) in the model while removing the sensory noise (σ). The details of other alternative models are described in Supplementary Information.

Model recovery. To evaluate our ability to distinguish the three models, we performed a model recovery analysis. Based on the design of each experiment (including the stimulus distributions, target locations and the number trials), we synthesized a dataset based on the best-fit parameters of each participant. We then fit each of the datasets with the three models. Supplementary Fig. 3 illustrates the results summed over all participants in each experiment.

Data visualization. For Experiments 1 and 3, we used a sliding window to visualize the psychometric curves, defined as the confidence ratings as a function of horizontal location of the target dot. The sliding window had a width of 0.6°. We moved the window horizontally (in a step of 0.1°) from the left to the right of the screen center. At each step, we computed mean confidence rating by averaging the confidence reports c of all the trials that fell within the window (based on the horizontal target location of each trial). We first applied this procedure to individual data, and then averaged the individual psychometric curves across subjects

(Fig. 3b, Supplementary Figs. 2, 7, 9 and 11). For Experiment 1, we visualized the data ranging from −3.5° to +3.5° from the screen center. For Experiment 3, we visualized the data ranging from −5° to +5° from the center. These ranges were chosen so that each steps along the curves in Fig. 3b, Supplementary Figs. 2, 7, 9 and 11 contained at least five trials per subject on average. To visualize the model fit, we sampled a series of evenly spaced target dot locations along the horizontal axis (in a step of 0.1°), and we used the best-fitting parameters to compute the confidence reports predicted by the models for each target location. We then used the same procedure (a sliding window) to compute the mean confidence rating predicted by the models (the model-fit curves in Fig. 3b, Supplementary Figs. 2, 7, 9 and 11).

For Experiment 2, the "psychometric curve" became a heat map in a two-dimensional space (Fig. 5). We tiled the two-dimensional space with non-overlapped hexagonal spatial windows (with a radius of 0.25°) positioned from −3° to +3° (Fig. 5a) along both horizontal and vertical axis. To compute the mean confidence rating for each hexagonal window, we averaged the confidence ratings across all the trials fell within that window for each participant. If the number of trials was zero among all the participants for a window, that window was left as white in Fig. 5a. To visualize the model fit, we used the best-fitting parameters and computed the confidence reports predicted by the models for an array of target locations (a grid tiling the two-dimensional space with a step of 0.1° along both horizontal and vertical axis). The predicted confidence reports were then averaged within each hexagonal window.

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

Data availability

The data that support the findings of this paper are available on https://github.com/hsinhungli/confidence-multiple-alternatives

Code availability

The analysis code used in this paper is available on https://github.com/hsinhungli/confidence-multiple-alternatives

Received: 20 March 2019; Accepted: 12 March 2020;
Published online: 24 April 2020

References

1. Persaud, N., McLeod, P. & Cowey, A. Post-decision wagering objectively measures awareness. *Nat. Neurosci.* **10**, 257 (2007).
2. Van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N. & Wolpert, D. M. Confidence is the bridge between multi-stage decisions. *Curr. Biol.* **26**, 3157–3168 (2016).
3. Meyniel, F., Schlunegger, D. & Dehaene, S. The sense of confidence during probabilistic learning: a normative account. *PLoS Comput. Biol.* **11**, e1004305 (2015).
4. Bahrami, B. et al. Optimally interacting minds. *Science* **329**, 1081–1085 (2010).
5. Vaghi, M. M. et al. Compulsivity reveals a novel dissociation between action and confidence. *Neuron* **96**, 348–354. e344 (2017).
6. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Front. Hum. Neurosci.* **8**, 443 (2014).
7. Mamassian, P. Visual confidence. *Annu. Rev. Vis. Sci.* **2**, 459–481 (2016).
8. Kepecs, A. & Mainen, Z. F. A computational framework for the study of confidence in humans and animals. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* **367**, 1322–1337 (2012).
9. Yeung, N. & Summerfield, C. Metacognition in human decision-making: confidence and error monitoring. *Philos. Trans. R. Soc. B* **367**, 1310–1321 (2012).
10. De Martino, B., Fleming, S. M., Garrett, N. & Dolan, R. J. Confidence in value-based choice. *Nat. Neurosci.* **16**, 105 (2013).
11. Lebreton, M., Abitbol, R., Daunizeau, J. & Pessiglione, M. Automatic integration of confidence in the brain valuation signal. *Nat. Neurosci.* **18**, 1159 (2015).
12. Polania, R., Woodford, M. & Ruff, C. C. Efficient coding of subjective value. *Nat. Neurosci.* **22**, 134 (2019).
13. Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19**, 366 (2016).
14. Drugowitsch, J., Moreno-Bote, R. & Pouget, A. Relation between belief and performance in perceptual decision making. *PloS ONE* **9**, e96511 (2014).
15. Clarke, F. R., Birdsall, T. G. & Tanner, W. P. Jr Two types of ROC curves and definitions of parameters. *J. Acoustical Soc. Am.* **31**, 629–630 (1959).
16. Galvin, S. J., Podd, J. V., Drga, V. & Whitmore, J. Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. *Psychonomic Bull. Rev.* **10**, 843–876 (2003).
17. Peirce, C. S. & Jastrow, J. On small differences in sensation. (1884).
18. Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227 (2008).
19. Kiani, R. & Shadlen, M. N. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759–764 (2009).
20. 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 (2013).
21. Sanders, J. I., Hangya, B. & Kepecs, A. Signatures of a statistical computation in the human sense of confidence. *Neuron* **90**, 499–506 (2016).
22. Barthelmé, S. & Mamassian, P. Flexible mechanisms underlie the evaluation of visual confidence. *Proc. Natl Acad. Sci. USA* **107**, 20834–20839 (2010).
23. Navajas, J. et al. The idiosyncratic nature of confidence. *Nat. Hum. Behav.* **1**, 810 (2017).
24. Aitchison, L., Bang, D., Bahrami, B. & Latham, P. E. Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Comput. Biol.* **11**, e1004519 (2015).
25. Adler, W. T. & Ma, W. J. Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLoS Comput. Biol.* **14**, e1006572 (2018).
26. Drugowitsch, J., Wyart, V., Devauchelle, A.-D. & Koechlin, E. Computational precision of mental inference as critical source of human choice suboptimality. *Neuron* **92**, 1398–1411 (2016).
27. Keshvari, S., Van den Berg, R. & Ma, W. J. Probabilistic computation in human perception under variability in encoding precision. *PLoS ONE* **7**, e40216 (2012).
28. Shen, S. & Ma, W. J. Variable precision in visual perception. *Psychol. Rev.* **126**, 89–132 (2019).
29. van den Berg, R., Yoo, A. H. & Ma, W. J. Fechner's law in metacognition: A quantitative model of visual working memory confidence. *Psychol. Rev.* **124**, 197 (2017).
30. Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control.* **19**, 716–723 (1974).
31. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
32. Kiani, R., Corthell, L. & Shadlen, M. N. Choice certainty is informed by both evidence and decision time. *Neuron* **84**, 1329–1342 (2014).
33. Moran, R., Teodorescu, A. R. & Usher, M. Post choice information integration as a causal determinant of confidence: novel data and a computational account. *Cogn. Psychol.* **78**, 99–147 (2015).
34. Pleskac, T. J. & Busemeyer, J. R. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychol. Rev.* **117**, 864 (2010).
35. Yu, S., Pleskac, T. J. & Zeigenfuse, M. D. Dynamics of postdecisional processing of confidence. *J. Exp. Psychol.: Gen.* **144**, 489 (2015).
36. Navajas, J., Bahrami, B. & Latham, P. E. Post-decisional accounts of biases in confidence. *Curr. Opin. Behav. Sci.* **11**, 55–60 (2016).
37. Koizumi, A., Maniscalco, B. & Lau, H. Does perceptual confidence facilitate cognitive control? *Atten., Percept., Psychophys.* **77**, 1295–1306 (2015).
38. Zylberberg, A., Barttfeld, P. & Sigman, M. The construction of confidence in a perceptual decision. *Front Integr. Neurosci.* **6**, 2359–2374 (2012).
39. Peters, M. A. et al. Perceptual confidence neglects decision-incongruent evidence in the brain. *Nat. Hum. Behav.* **1**, 0139 (2017).
40. Maniscalco, B. & Lau, H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cognition* **21**, 422–430 (2012).
41. Desender, K., Boldt, A. & Yeung, N. Subjective confidence predicts information seeking in decision making. *Psychol. Sci.* **29**, 761–778 (2018).
42. Brown, S., Steyvers, M. & Wagenmakers, E.-J. Observing evidence accumulation during multi-alternative decisions. *J. Math. Psychol.* **53**, 453–462 (2009).
43. Markant, D. B., Settles, B. & Gureckis, T. M. Self directed learning favors local, rather than global, uncertainty. *Cogn. Sci.* **40**, 100–120 (2016).
44. Kahneman, D. & Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica.* **47**, 263–292 (1979).
45. Carandini, M. & Heeger, D. J. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2012).
46. Louie, K., Grattan, L. E. & Glimcher, P. W. Reward value-based gain control: divisive normalization in parietal cortex. *J. Neurosci.* **31**, 10627–10639 (2011).
47. Stengård, E., Van & den Berg, R. Imperfect Bayesian inference in visual perception. *PLoS Comput. Biol.* **15**, e1006465 (2019).
48. Ratcliff, R., Smith, P. L., Brown, S. D. & McKoon, G. Diffusion decision model: Current issues and history. *Trends Cogn. Sci.* **20**, 260–281 (2016).
49. Ratcliff, R. & Starns, J. J. Modeling confidence judgments, response times, and multiple choices in decision making: recognition memory and motion discrimination. *Psychol. Rev.* **120**, 697 (2013).
50. Ratcliff, R. & Starns, J. J. Modeling confidence and response time in recognition memory. *Psychol. Rev.* **116**, 59 (2009).
51. Vickers, D. *Decision processes in visual perception*. (Academic Press, 1979).

52. Vickers, D. Where does the balance of evidence lie with respect to confidence? In *Proceedings of the seventeenth annual meeting of the international society for psychophysics*. pp. 148–153 (Lengerich, Germany: Pabst Science Publishers, 2001).
53. Vickers, D. & Lee, M. D. Dynamic models of simple judgments: I. Properties of a self-regulating accumulator module. *Nonlinear Dynamics. Psychol., Life Sci.* **2**, 169–194 (1998).
54. Folke, T., Jacobsen, C., Fleming, S. M. & De Martino, B. Explicit representation of confidence informs future value-based decisions. *Nature Human. Behaviour* **1**, 0002 (2017).
55. Odegaard, B. et al. Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proceedings of the National Academy of Sciences*, **115**, E1588–E1597 (2018).
56. Churchland, A. K., Kiani, R. & Shadlen, M. N. Decision-making with multiple alternatives. *Nat. Neurosci.* **11**, 693 (2008).
57. Churchland, A. K. & Ditterich, J. New advances in understanding decisions among multiple alternatives. *Curr. Opin. Neurobiol.* **22**, 920–926 (2012).
58. Ditterich, J. A comparison between mechanisms of multi-alternative perceptual decision making: ability to explain human behavior, predictions for neurophysiology, and relationship with decision theory. *Front. Neurosci.* **4**, 184 (2010).
59. Hampton, R. R. Rhesus monkeys know when they remember. *Proc. Natl Acad. Sci.* **98**, 5359–5362 (2001).
60. Foote, A. L. & Crystal, J. D. Metacognition in the rat. *Curr. Biol.* **17**, 551–555 (2007).
61. Acerbi, L., Wolpert, D. M. & Vijayakumar, S. Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS computational Biol.* **8**, e1002771 (2012).
62. van den Berg, R., Awh, E. & Ma, W. J. Factorial comparison of working memory models. *Psychological Rev.* **121**, 124 (2014).
63. Daunizeau, J., Preuschoff, K., Friston, K. & Stephan, K. Optimizing experimental design for comparing models of brain function. *PLoS computational Biol.* **7**, e1002280 (2011).
64. Acerbi, L. & Ma, W. J. Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search. In *Advances in Neural Information Processing Systems*. pp. 1836–1846 (2017).

Acknowledgements

We thank members of the Ma Lab, Hui-Kuan Chung, Rachel Denison, and Michael Landy for helpful comments on the manuscript.

Author contributions

H.-H. L. and W.J.M. designed the experiment. H.-H. L. performed the experiment and analyzed the data. H.-H. L. and W.J.M. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-15581-6.

Correspondence and requests for materials should be addressed to H.-H.L.

Peer review information *Nature Communications* thanks Stephen Fleming, Samuel Gershman, and Rafael Polania for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Information

Supplementary Text

In addition to the three main models and their reduced versions discussed in the main text, we considered two types of alternative models. Alternative models of the first type are still Bayesian in the sense that the posterior probabilities of the categories are computed, but they contain different sources of variability than those considered in the main text. In alternative models of the second type (heuristic models), posterior probabilities are not computed; observers' category decisions and confidence reports are instead based on ad-hoc functions of the noisy measurement **x** and the stimuli. All the alternative models we tested contain a lapse rate as a free parameter. The model-fitting procedures are described in the main text (see **Methods**).

Bayesian models with different sources of variability

Sampling noise. Some studies have suggested that the brain approximates posterior probabilities by Monte Carlo sampling or simulation[1, 2, 3]. On each trial, the brain has access to n samples drawn from a multinomial distribution with parameters given by the true posterior probabilities, $\mathbf{p}=(p_1,\ p_2,\ p_3)=(p(C=1|\mathbf{x}),\ p(C=2|\mathbf{x}),\ p(C=3|\mathbf{x}))$. The probability mass function of this multinomial distribution is $f(n_1,n_2,n_3;n,\mathbf{p})=\frac{n!}{n_1!n_2!n_3!}p_1^{n_1}p_2^{n_2}p_3^{n_3}$, in which n_i represents the number of samples drawn from the i^{th} category and $\sum_{i=1}^{3}n_i=n$. The number of samples n is a free parameter.

We assume that the observer's category decisions and confidence reports are based on the noisy posterior $\mathbf{q}=(q_1,\ q_2,\ q_3)$, where $q_i=\frac{n_i}{n}$. Thus, sampling noise is a form of decision noise, and the higher n, the less decision noise there is.

We took the Max, Difference and Entropy models (with both sensory noise and decision noise) and replaced the Dirichlet decision noise by sampling noise while keeping all the other aspects of the models unchanged. We named the resulting models Max-Sen-Samp, Diff-Sen-Samp and Ent-Sen-Samp. We also tested versions of the models that do not include sensory noise (Max-Samp, Diff-Samp and Ent-Samp models).

Noisy measurement of the category mean. In the three main models (Max, Difference and Entropy), given a uniform prior and the assumption that the observer has perfect

knowledge about the stimulus distribution, the posterior probability of category C given the measurement of the target location $\mathbf{x}$ is $p(C|\mathbf{x}) \propto p(\mathbf{x}|C) = N(\mathbf{x};\mathbf{m}_c,(\sigma_s^2+\sigma^2)\mathbf{I})$ (see **Methods**). Here, we allow for the possibility that the category mean $\mathbf{m}_c$ is not known exactly but measured in a noisy fashion. Then, the posterior probability of category C is computed as $p(C|\mathbf{x},\hat{\mathbf{m}}_c) \propto N(\mathbf{x};\hat{\mathbf{m}}_c,(\sigma_s^2+\sigma^2)\mathbf{I})$, in which $\hat{\mathbf{m}}_c \sim N(\mathbf{m}_c,\sigma_m^2\mathbf{I})$ is the measurement of the mean of category C. σ_m is a free parameter that controls the amount of noise in the measurement of the category mean. We assume that across three categories, this measurement noise is identical and independent.

We took the Max, Difference and Entropy models (with both sensory noise and decision noise) and replaced the Dirichlet decision noise by noise in measuring the category mean while keeping all the other aspects of the models unchanged. The resulting models are named Max-Sen-Mean, Diff-Sen-Mean and Ent-Sen-Mean. We also tested versions of the models that only include noise in measuring the category mean (Max-Mean, Diff-Mean, Ent-Mean models), without sensory noise or Dirichlet decision noise.

Heuristic models

Distance model. This model makes decisions based on the distance from the measurement to the center of each category. The observer chooses the category with the shortest distance $\hat{C} = \underset{i}{\operatorname{argmin}}\, d_i$, in which $d_i = ||\mathbf{x}_i - \mathbf{m}_i||$ is the distance between the measurement and the center of category i. We further assume that confidence depends on the difference between the two shortest distances. For example, if $d_1 < d_2 < d_3$, an internal confidence variable is computed as $c^* = d_2 - d_1$. The internal confidence variable c^* is then converted to a four-point confidence report c by applying three criteria b_1, b_2 and b_3.

Weighted-distance model. This model is similar to the distance model, except that the continuous confidence variable is a linear function of the distance to each group. For example, if $d_1 < d_2 < d_3$, the continuous confidence variable is computed as $c^* = -0.5d_1 + ad_2 + bd_3$, in which a and b are free parameters representing the weights for the medium and the longest distance respectively. The weight for the shortest distance is fixed at -0.5. Allowing the weights of all three distances to be free parameters would have been redundant: infinitely many combinations of the weights and the three criteria (b_1, b_2 and b_3) would have produced the same model predictions. Choosing a=0.5 and b=0 reduces the Weighted-distance model to the Distance model.

Distance-to-Boundary model. This model is inspired by the finding by Kepecs et al. (2008) that behavioral and neural correlates of confidence showed responses that varied as a function of the distance between the target and the category boundary in the stimulus space. As in the Distance model, to make category decisions, the observer chooses the category that is closest to the measurement: $\hat{C} = \underset{i}{\operatorname{argmin}}\, d_i$. The internal confidence variable c^* is computed as the distance between the measurement **x** and a decision boundary. This decision boundary is defined as a line perpendicular to and goes through the midpoint of the line connecting the centers of the two nearest categories.

We tested three versions of each of the heuristic models above. A version that only considers the sensory noise σ^2 (Dist-Sen, DistW-Sen and Bound-Sen models), a version that only considers the noisy estimate of the category center modeled by σ_m (Dist-Mean, DistW-Mean and Bound-Mean models), and a version that considers both (Dist-Sen-Mean, DistW-Sen-Mean and Bound-Sen-Mean models).

Supplementary Tables

	#	Experiment 1	Experiment 2	Experiment 3
Diff-Sen-Dir	6	0 [-]	0 [-]	0 [-]
Max-Sen-Dir	6	391 [222, 569]	541 [371, 735]	100 [46, 156]
Ent-Sen-Dir	6	1937 [1363, 2562]	1631 [1179, 2159]	1113 [817, 1447]
Diff-Dir	5	121 [48, 199]	132 [30, 237]	36 [3, 77]
Max-Dir	5	440 [276, 621]	616 [421, 816]	113 [48, 176]
Ent-Dir	5	1913 [1314, 2544]	1683 [1208, 2198]	1092 [797, 1395]
Diff-Sen	5	737 [590, 914]	921 [664, 1196]	1171 [982, 1363]
Max-Sen	5	1504 [1217, 1792]	1223 [933, 1520]	2011 [1719, 2299]
Ent-Sen	5	4114 [3394, 4920]	2190 [1658, 2733]	3835 [3356, 4287]
Diff-Samp	5	154 [73, 238]	292 [162, 426]	122 [54, 194]
Max-Samp	5	581 [426, 753]	812 [573, 1047]	325 [248, 403]
Ent-Samp	5	1744 [1219, 2282]	1540 [1068, 2053]	1134 [911, 1413]
Diff-Sen-Samp	6	-25 [-65, 17]	91 [2, 198]	-5 [-51, 43]
Max-Sen-Samp	6	411 [242, 574]	573 [397, 754]	207 [128, 308]
Ent-Sen-Samp	6	1751 [1225, 2325]	1503 [1034, 1983]	1121 [877, 1376]
Diff-Mean	5	183 [13, 340]	42 [-124, 182]	263 [115, 392]
Max-Mean	5	507 [242, 768]	535 [337, 703]	352 [165, 533]
Ent-Mean	5	2607 [1955, 3335]	1730 [1272, 2239]	2236 [1778, 2716]
Diff-Sen-Mean	6	29 [-102, 161]	-65 [-190, 76]	132 [1, 245]
Max-Sen-Mean	6	429 [183, 686]	420 [259, 561]	292 [106, 480]
Ent-Sen-Mean	6	2632 [1975, 3356]	1677 [1200, 2178]	2254 [1806, 2741]
Dist-Sen	5	1749 [1350, 2169]	1381 [1174, 1587]	2069 [1808, 2321]
Dist-Mean	5	745 [496, 1029]	265 [109, 421]	759 [492, 1030]
Dist-Sen-Mean	6	713 [434, 1026]	179 [68, 312]	730 [470, 975]
DistW-Sen	7	829 [667, 976]	1014 [757, 1266]	1204 [991, 1428]
DistW-Mean	7	361 [179, 566]	103 [-63, 281]	455 [256, 666]
DistW-Sen-Mean	8	269 [130, 403]	16 [-103, 132]	344 [208, 495]
Bound-Sen	5	1464 [1127, 1847]	2305 [1784, 2762]	1726 [1196, 2231]
Bound-Mean	5	1829 [1435, 2300]	1566 [1178, 1954]	1834 [1388, 2206]
Bound-Sen-Mean	6	1306 [995, 1701]	1492 [1085, 1863]	1354 [944, 1735]
Ratio-Sen-Dir	6	25 [-19, 62]	-19 [-54, 15]	51 [18, 90]

Supplementary Table 1. ΔAIC of each model and experiment, computed as the AIC of each model minus the AIC of the Diff-Sen-Dir model (the Difference model with both the sensory noise and Dirichlet decision noise). ΔAIC is computed for individual participants and then summed across participants. The first two columns are the model name, and the number of the free parameters. For each model and experiment, the group-summed ΔAIC and bootstrapped 95% confidence interval are reported. Names of the model are denoted as decision rules paired with the sources of variability separated by hyphens (-). Diff: Difference model; Max: Max model; Ent: Entropy model; Dist: Distance model; DistW: Weighted distance model; Bound: Distance-to-bound model; Ratio: Ratio model; Sen: sensory noise; Dir: Dirichlet decision noise; Samp: Sampling noise; Mean: noisy estimation of category mean.

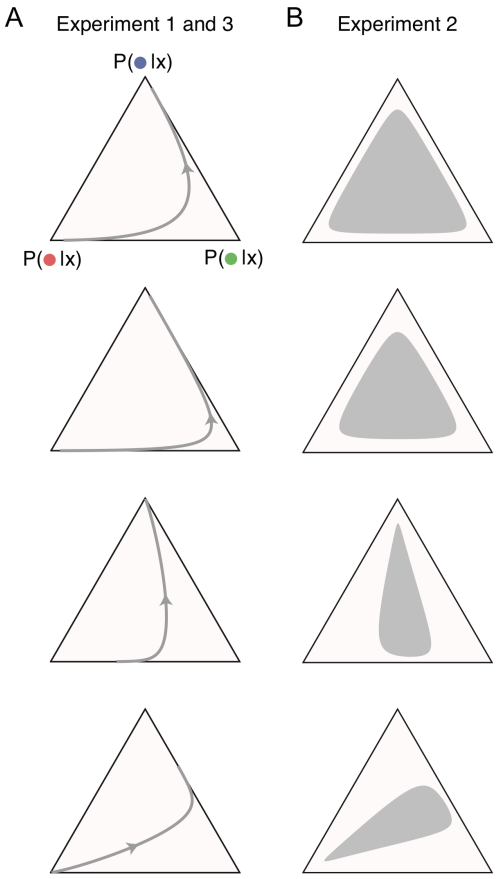
	#	Experiment 1	Experiment 2	Experiment 3
Diff-Sen-Dir	6	0 [-]	0 [-]	0 [-]
Max-Sen-Dir	6	391 [224, 574]	541 [372, 728]	100 [45, 155]
Ent-Sen-Dir	6	1937 [1356, 2569]	1631 [1182, 2142]	1113 [796, 1433]
Diff-Dir	5	72 [-3, 150]	85 [-21, 195]	-10 [-44, 33]
Max-Dir	5	390 [225, 553]	570 [361, 774]	67 [3, 133]
Ent-Dir	5	1863 [1275, 2509]	1637 [1191, 2142]	1046 [725, 1371]
Diff-Sen	5	688 [534, 859]	875 [615, 1147]	1125 [941, 1317]
Max-Sen	5	1455 [1156, 1749]	1176 [887, 1480]	1965 [1684, 2246]
Ent-Sen	5	4065 [3349, 4899]	2143 [1627, 2710]	3789 [3340, 4238]
Diff-Samp	5	104 [24, 191]	245 [110, 373]	76 [6, 146]
Max-Samp	5	531 [376, 684]	765 [526, 1001]	280 [206, 354]
Ent-Samp	5	1695 [1209, 2248]	1493 [990, 2023]	1088 [851, 1347]
Diff-Sen-Samp	6	-25 [-66, 18]	91 [2, 194]	-5 [-52, 46]
Max-Sen-Samp	6	411 [243, 589]	573 [411, 761]	207 [126, 306]
Ent-Sen-Samp	6	1751 [1226, 2277]	1503 [1047, 2013]	1121 [894, 1384]
Diff-Mean	5	133 [-38, 293]	-4 [-161, 130]	217 [76, 353]
Max-Mean	5	457 [194, 712]	489 [300, 652]	306 [112, 502]
Ent-Mean	5	2557 [1904, 3278]	1683 [1204, 2169]	2190 [1724, 2681]
Diff-Sen-Mean	6	29 [-103, 163]	-65 [-199, 74]	132 [5, 251]
Max-Sen-Mean	6	429 [180, 680]	420 [245, 564]	292 [105, 482]
Ent-Sen-Mean	6	2632 [1961, 3368]	1677 [1196, 2187]	2254 [1777, 2756]
Dist-Sen	5	1700 [1276, 2105]	1334 [1109, 1538]	2023 [1753, 2266]
Dist-Mean	5	695 [425, 978]	219 [59, 384]	713 [426, 985]
Dist-Sen-Mean	6	713 [439, 1014]	179 [68, 320]	730 [473, 975]
DistW-Sen	7	879 [723, 1023]	1060 [821, 1319]	1250 [1013, 1475]
DistW-Mean	7	411 [225, 600]	149 [-23, 333]	501 [304, 699]
DistW-Sen-Mean	8	369 [233, 505]	109 [-5, 226]	436 [295, 587]
Bound-Sen	5	1415 [1085, 1775]	2259 [1793, 2738]	1680 [1146, 2180]
Bound-Mean	5	1779 [1404, 2238]	1520 [1142, 1900]	1789 [1365, 2175]
Bound-Sen-Mean	6	1306 [980, 1678]	1492 [1102, 1871]	1354 [949, 1734]
Ratio-Sen-Dir	6	25 [-15, 62]	-19 [-55, 18]	51 [16, 93]

Supplementary Table 2. ΔBIC of each model and experiment, computed as the BIC of each model minus the BIC of the Diff-Sen-Dir model (the Difference model with both the sensory noise and Dirichlet decision noise). ΔBIC is computed for individual participants and then summed across participants. The first two columns are the model name, and the number of the free parameters. For each model and experiment, group-summed ΔAIC and bootstrapped 95% confidence interval are reported. Names of the model are denoted as decision rules paired with the sources of variability separated by hyphens (-). Diff: Difference model; Max: Max model; Ent: Entropy model; Dist: Distance model; DistW: Weighted distance model; Bound: Distance-to-bound model; Ratio: Ratio model; Sen: sensory noise; Dir: Dirichlet decision noise; Samp: Sampling noise; Mean: noisy estimation of category mean.

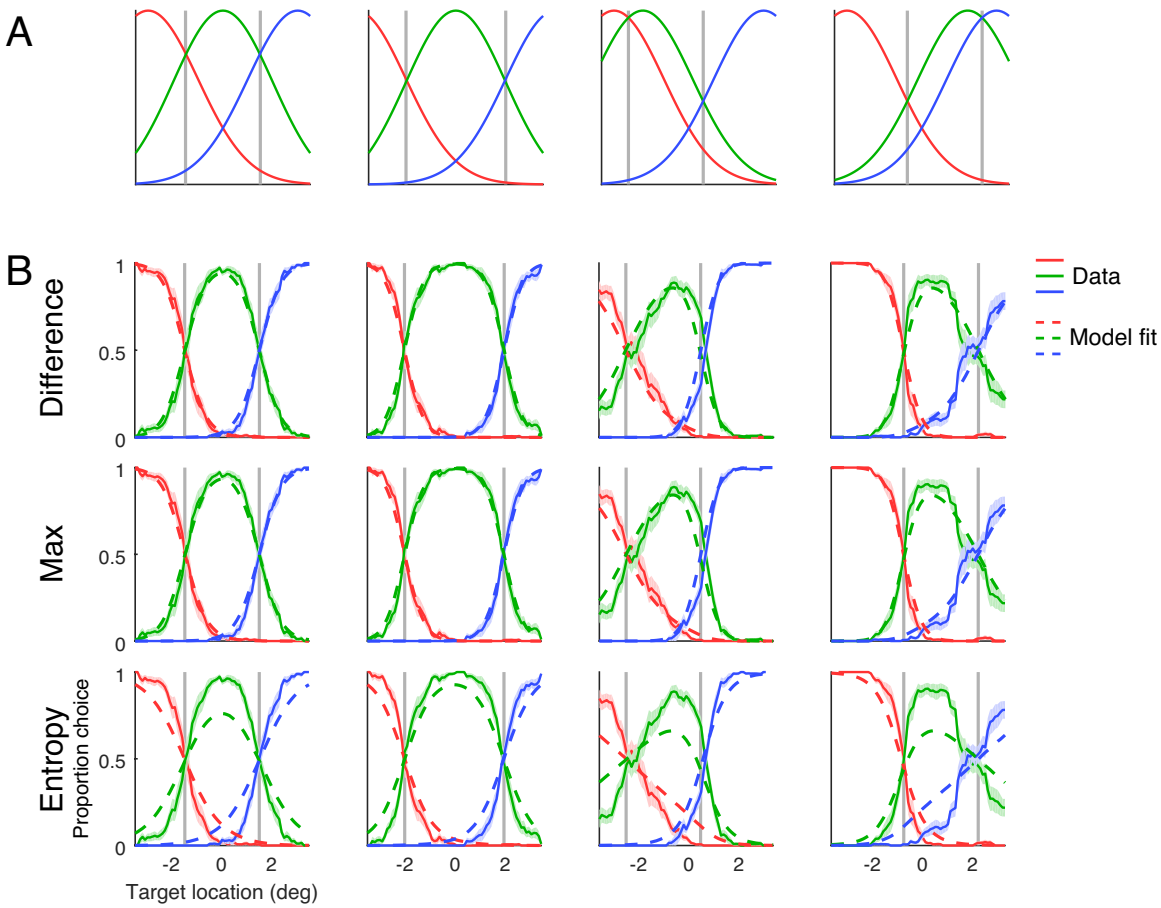
	#	Experiment 1	Experiment 2	Experiment 3
ΔAIC				
Diff-Sen-Dir	6	0 [-]	0 [-]	0 [-]
Max-Sen-Dir	6	349 [203, 508]	547 [371, 736]	114 [58, 169]
Ent-Sen-Dir	6	1333 [935, 1776]	1436 [991, 1885]	875 [531, 1274]
Diff-Dir	5	116 [42, 197]	115 [41, 197]	20 [-2, 51]
Max-Dir	5	422 [270, 580]	589 [383, 794]	105 [51, 157]
Ent-Dir	5	1377 [965, 1898]	1424 [1002, 1876]	877 [532, 1244]
Diff-Sen	5	234 [152, 333]	252 [146, 375]	504 [393, 615]
Max-Sen	5	628 [450, 809]	632 [443, 859]	882 [716, 1046]
Ent-Sen	5	1818 [1369, 2330]	1555 [1062, 2103]	2182 [1929, 2423]
ΔBIC				
Diff-Sen-Dir	6	0 [-]	0 [-]	0 [-]
Max-Sen-Dir	6	349 [204, 519]	547 [362, 741]	114 [56, 174]
Ent-Sen-Dir	6	1333 [933, 1767]	1436 [977, 1898]	875 [522, 1237]
Diff-Dir	5	66 [-8, 151]	69 [-6, 149]	-26 [-48, 6]
Max-Dir	5	372 [218, 530]	543 [340, 743]	59 [5, 114]
Ent-Dir	5	1328 [916, 1828]	1378 [945, 1814]	831 [490, 1227]
Diff-Sen	5	184 [105, 279]	206 [100, 323]	458 [349, 580]
Max-Sen	5	578 [409, 760]	585 [410, 798]	836 [677, 1006]
Ent-Sen	5	1769 [1320, 2283]	1509 [1044, 2064]	2136 [1893, 2366]

Supplementary Table 3. ΔAIC and ΔBIC of the models when fitting the confidence reports alone.

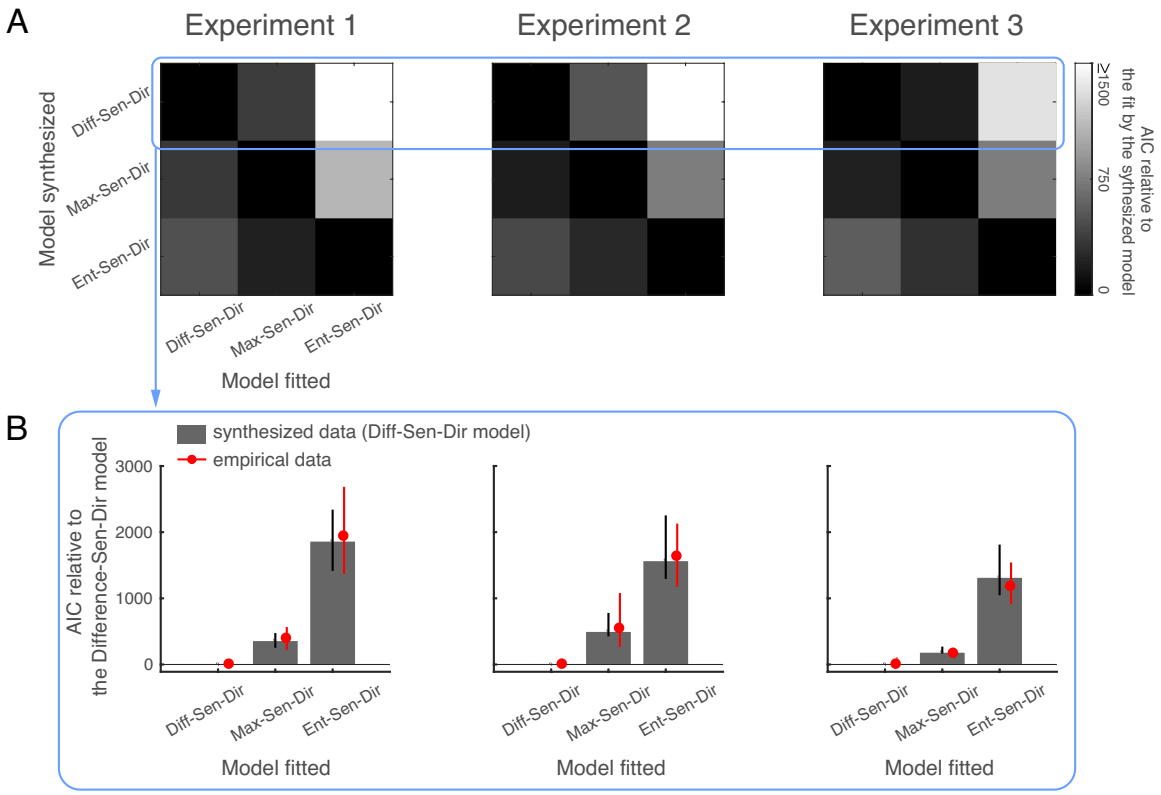
Supplementary Figures



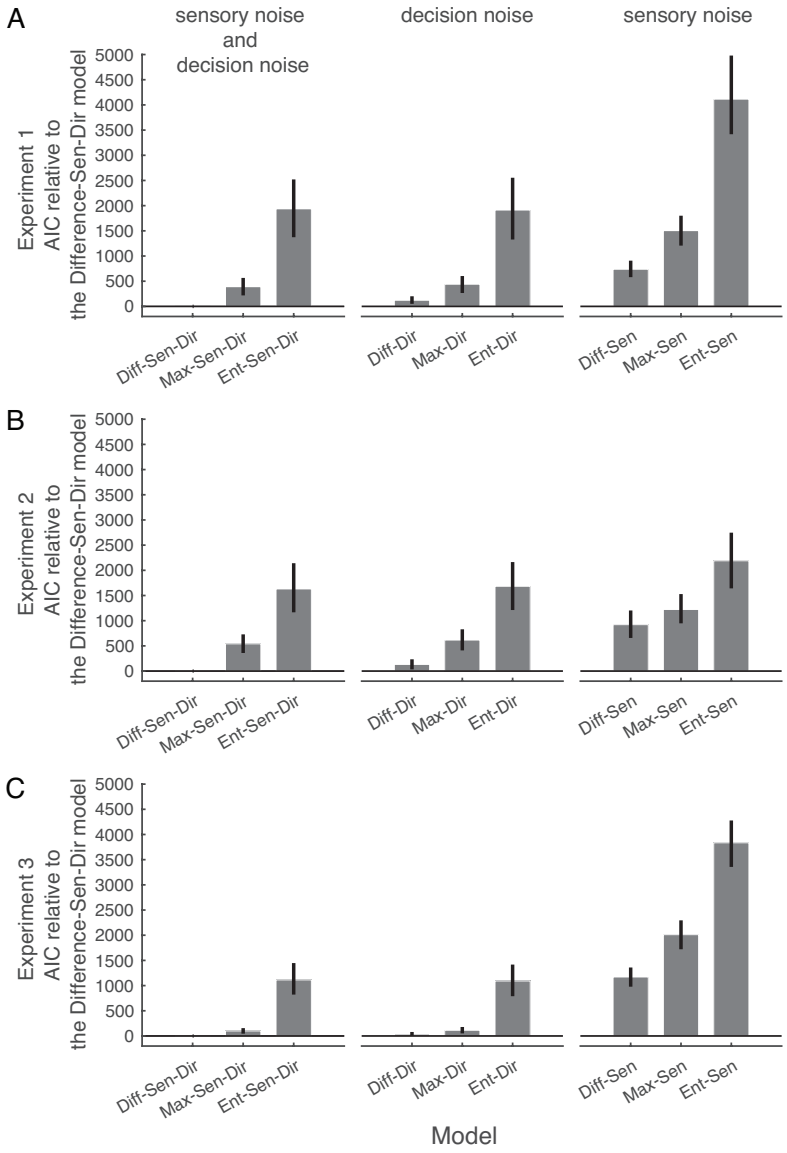
Supplementary Figure 1. Illustration of how observers' belief, posterior distribution, about the target category could change as a function of the target dot position. For illustration purpose, we consider a simplified case in which there is no sensory noise and no decision noise, so the posterior distribution only depends the target dot position and the distribution of each category. We use ternary plots to represent all possible posterior distributions. (A) Experiment 1 and 3: The four panels correspond to the four conditions depicted in Figure 1B. The gray lines and the arrows indicate the trajectory of the posterior distribution on the ternary plot as a target dot move from the left-end to the right-end of the screen. (B) Experiment 2: The four panels correspond to the four conditions depicted in Figure 1C. In the experiment, the target dot was uniformly sampled within a circle at the center of the screen with a radius of 2.6° (see Methods). All possible target dot locations within the circle correspond to a range of posterior probabilities indicated by the gray region in each panel.



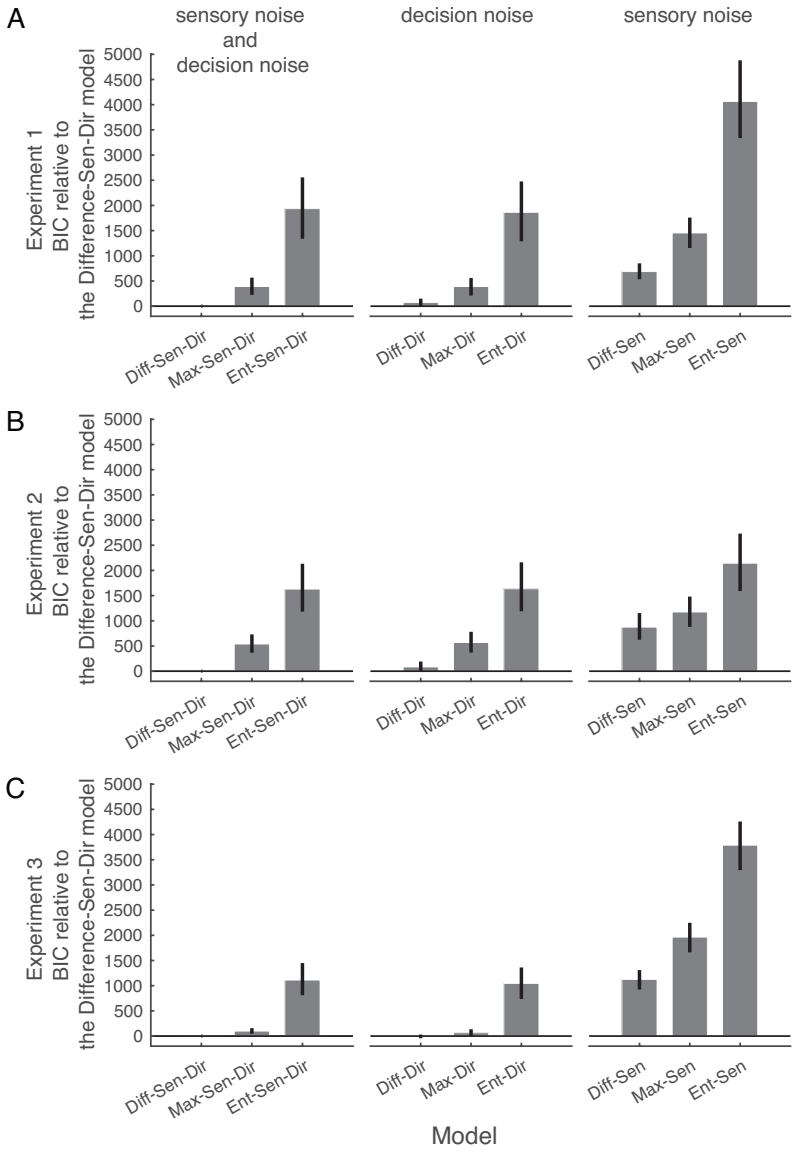
Supplementary Figure 2. Experiment 1. (A) Distribution of the reference dots in each condition. (B) The red (green, blue) lines represent the probability that the observers categorize the target dot to the red (green, blue) category as a function of the target dot location. Solid lines represent the group mean ± 1 s.e.m. The dashed lines represent the model fit averaged across individuals. In both (A) and (B), the gray vertical lines represent the boundary between two neighboring categories, the location where two neighboring categories have the same likelihood.



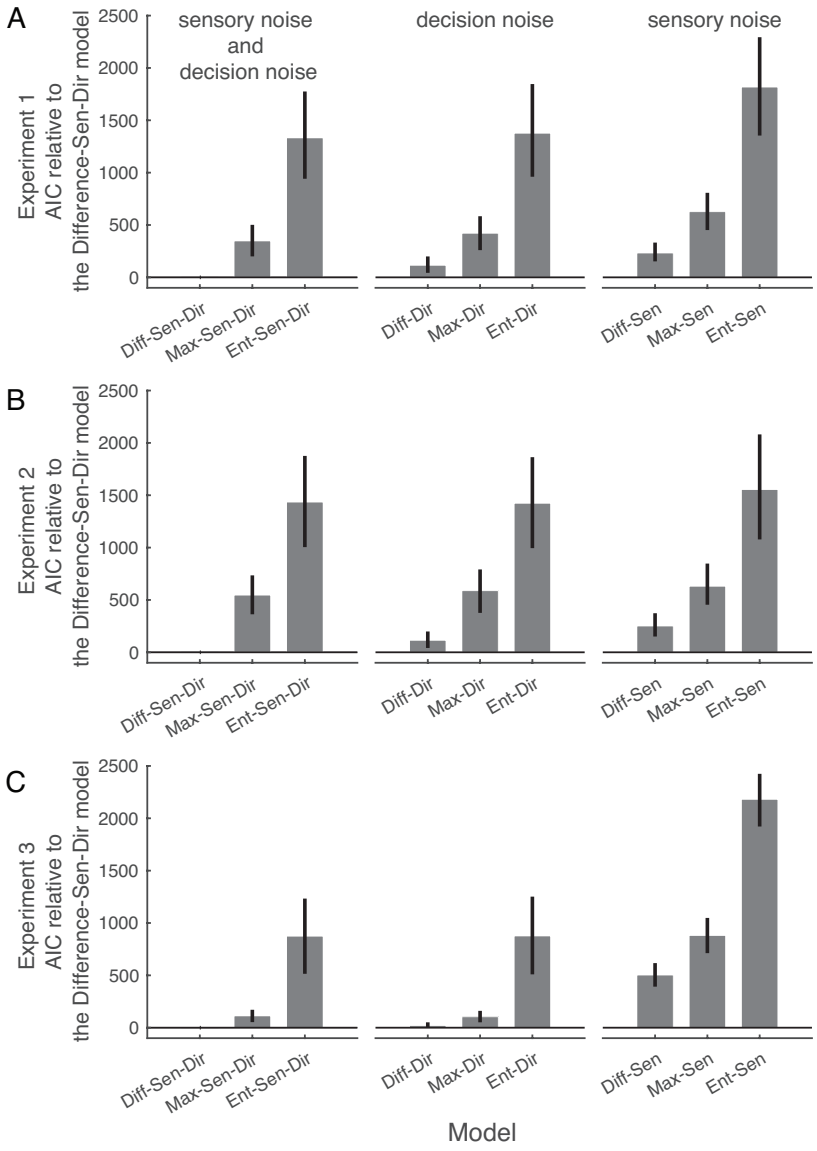
Supplementary Figure 3. Model recovery. (A) The darkness of the images represents ΔAIC (computed as the AIC of each fitted model minus the AIC of the fitted model that is the model used to synthesize the data) summed across participants. (B) The bars represent ΔAIC of the datasets synthesized based on the Difference model, corresponding to the top row of the images in (A). The red data points are the ΔAIC obtained in the experiments. Names of the model are denoted as decision rules paired with the sources of variability separated by hyphens (-). Diff: Difference model; Max: Max model; Ent: Entropy model; Sen: sensory noise; Dir: Dirichlet decision noise.



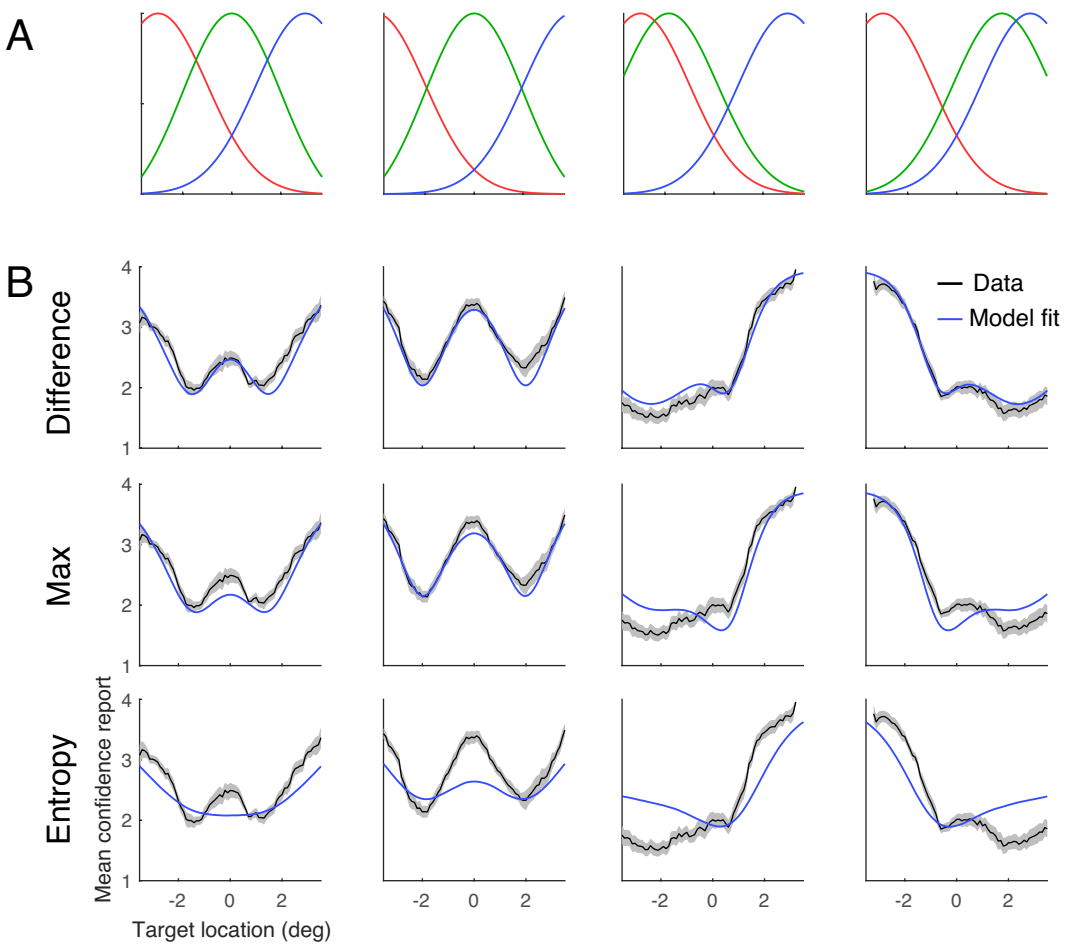
Supplementary Figure 4. Model comparison using AIC for both the full models (with both sensory and decision noise in the model) and the reduced models (with only the decision noise or only the sensory noise in the model). (A) Experiment 1. (B) Experiment 2. (C) Experiment 3. The bars represent ΔAIC (AIC of each model compared with the main Difference model, the Diff-Sen-Dir model) summed across participants. The error bars represent 95% bootstrapped confidence interval. Names of the model are denoted as decision rules paired with the sources of variability separated by hyphens (-). Diff: Difference model; Max: Max model; Ent: Entropy model; Sen: sensory noise; Dir: Dirichlet decision noise.



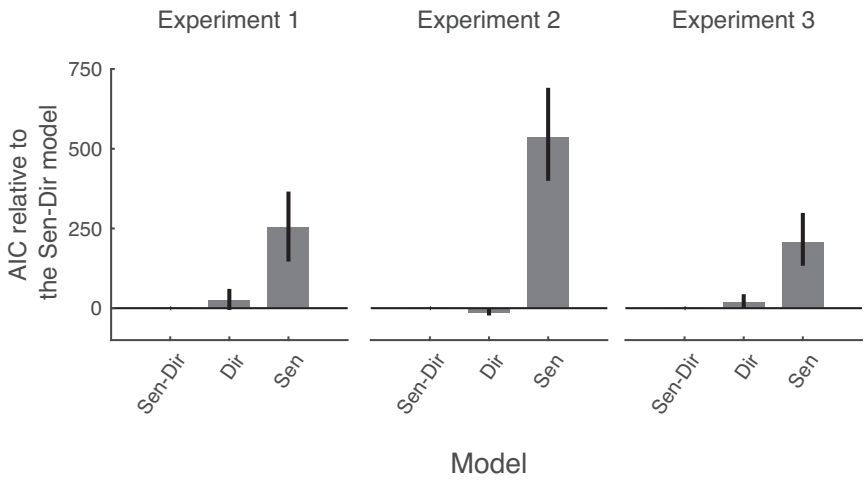
Supplementary Figure 5. Model comparison using BIC for both the full models (with both sensory and decision noise in the model) and the reduced models (with only the decision noise or only the sensory noise in the model). (A) Experiment 1. (B) Experiment 2. (C) Experiment 3. The bars represent ΔBIC (BIC of each model compared with the main Difference model, the Diff-Sen-Dir model) summed across participants. The error bars represent 95% bootstrapped confidence interval. Names of the model are denoted as decision rules paired with the sources of variability separated by hyphens (-). Diff: Difference model; Max: Max model; Ent: Entropy model; Sen: sensory noise; Dir: Dirichlet decision noise.



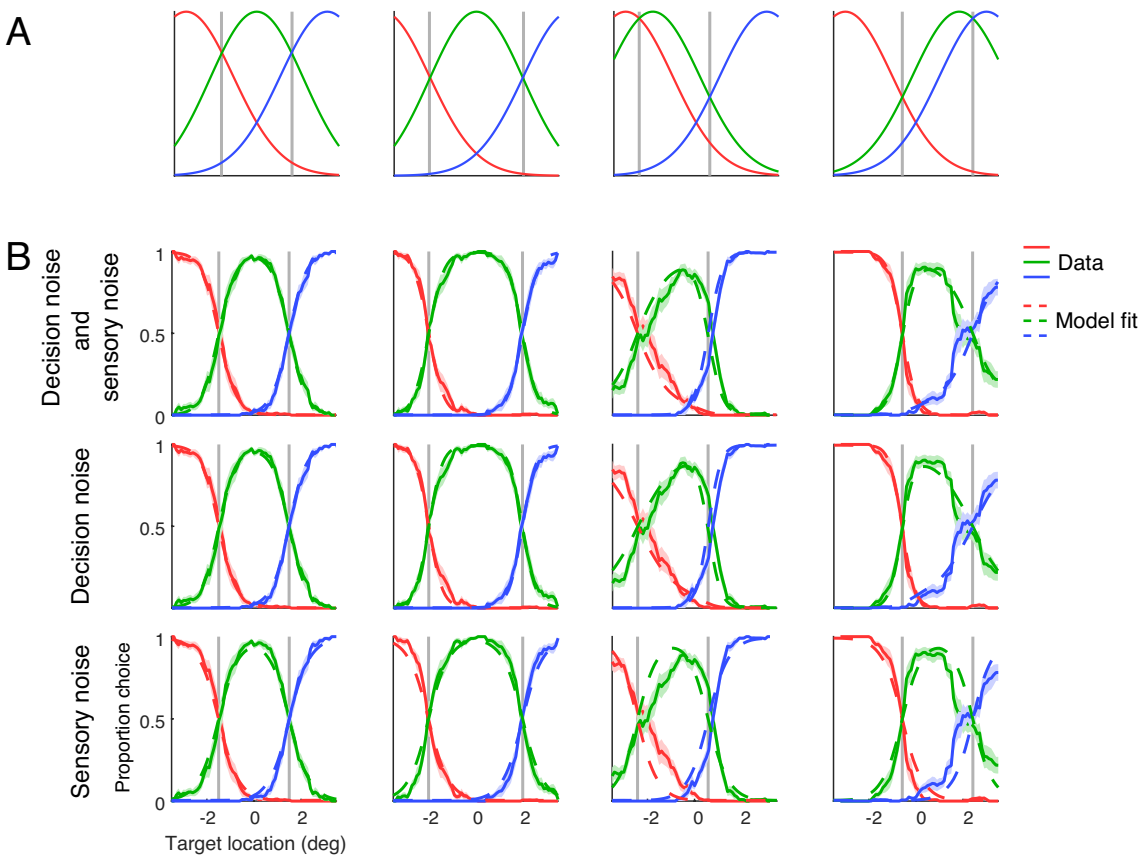
Supplementary Figure 6. Model comparison using confidence reports only. AIC for both the full models (with both sensory and decision noise in the model) and the reduced models (with only the decision noise or only the sensory noise in the model). (A) Experiment 1. (B) Experiment 2. (C) Experiment 3. The bars represent ΔAIC (AIC of each model compared with the main Difference model, the Diff-Sen-Dir model) summed across participants. The error bars represent 95% bootstrapped confidence interval. Names of the model are denoted as decision rules paired with the sources of variability separated by hyphens (-). Diff: Difference model; Max: Max model; Ent: Entropy model; Sen: sensory noise; Dir: Dirichlet decision noise.



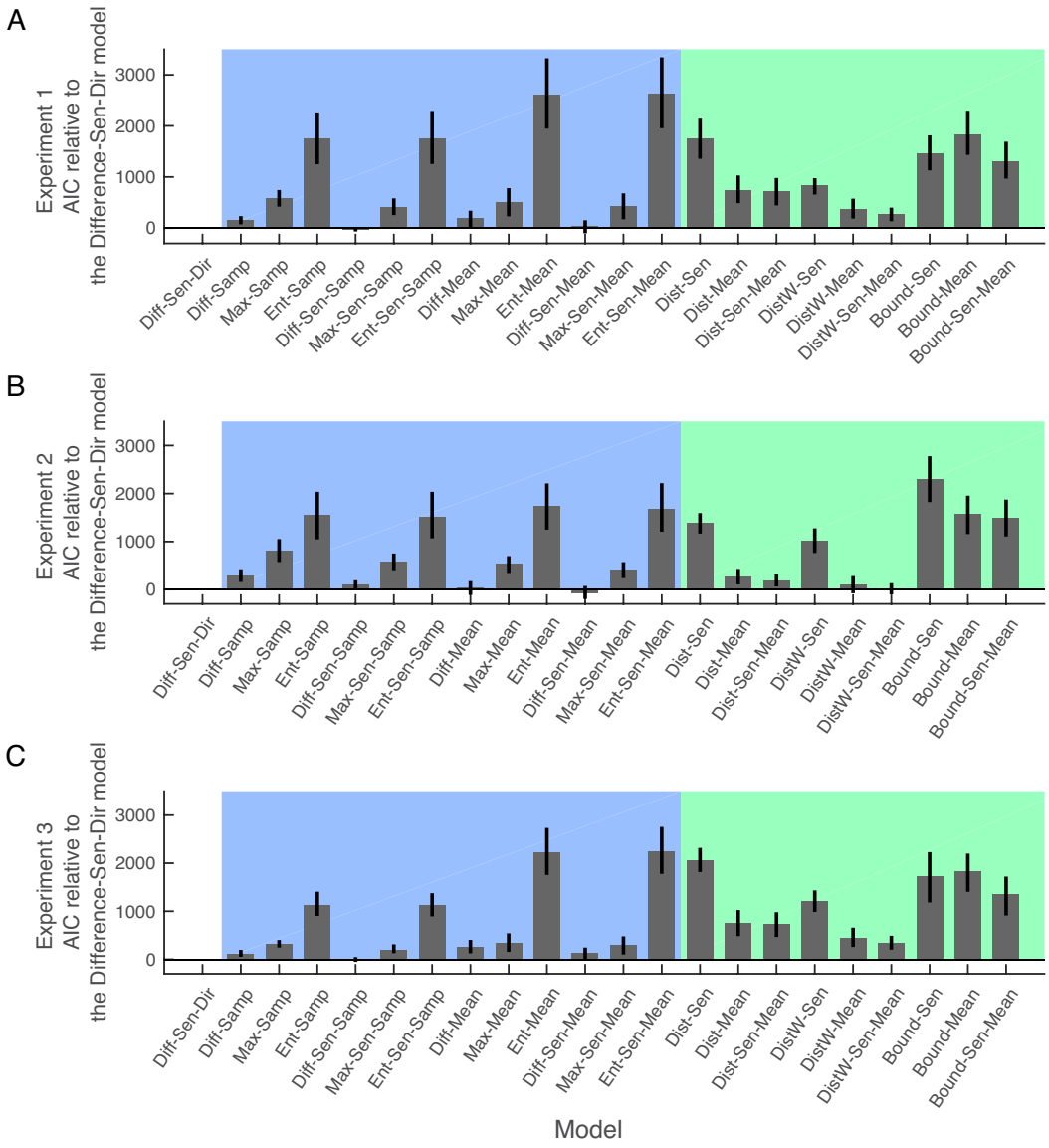
Supplementary Figure 7. Experiment 1. Model fit with confidence reports only, without jointly fitting the category decisions. (A) The distribution of the reference dots in each condition. (B) Mean confidence report as a function of target position for each of the four conditions. The black curves represent group mean ± 1 s.e.m. Blue curves represent the model fit averaged across individuals.



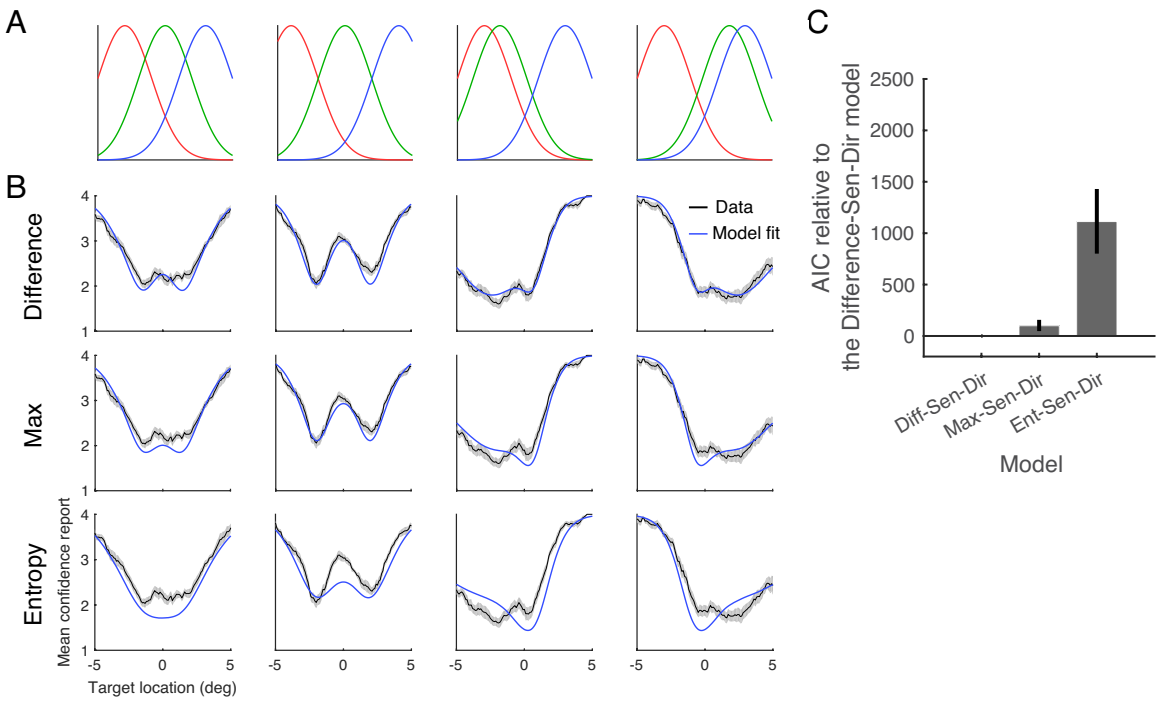
Supplementary Figure 8. Model comparison using category decisions only. Three models all choose the category with the highest posterior probability but consider different sources of variability: sensory and Dirichlet decision noise (Sen-Dir), Dirichlet decision noise only (Dir), and sensory noise only (Sen). The bars represent ΔAIC (AIC of each model compared with the Sen-Dir model) summed across participants. The error bars represent 95% bootstrapped confidence interval.



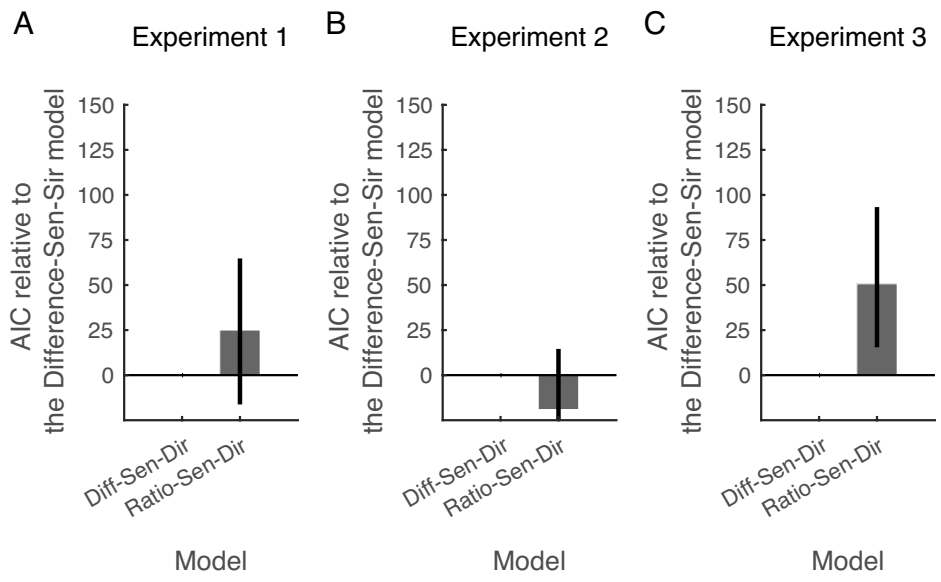
Supplementary Figure 9. Experiment 1. Model fit with category decisions only, without jointly fitting confidence reports. (A) Distribution of the reference dots in each condition. (B) The red (green, blue) lines represent the probability that the observers categorize the target dot to the red (green, blue) category as a function of the target dot location. Solid lines represent the group mean ± 1 s.e.m. The dashed lines represent the model fit averaged across individuals. In both (A) and (B), the gray vertical lines represent the boundary between two neighboring categories, the location where two neighboring categories have the same likelihood.



Supplementary Figure 10. Alternative models. Blue background: Bayesian models that compute posterior probabilities. Three confidence models (Difference, Max and Entropy) are paired with different sources of variability. Green background: Heuristic models that use the measurement of the target and the stimuli to perform the tasks. (A) Experiment 1. (B) Experiment 2. (C) Experiment 3. The bars represent ΔAIC (AIC of each model compared with the main Difference model, the Difference-Sen-Dir model) summed across participants. The error bars represent 95% bootstrapped confidence interval. Names of the model are denoted as decision rules paired with the sources of variability separated by hyphens (-). Diff: Difference model; Max: Max model; Ent: Entropy model; Dist: Distance model; DistW: Weighted distance model; Bound: Distance-to-bound model; Sen: sensory noise; Dir: Dirichlet decision noise; Samp: Sampling noise; Mean: noisy measurement of category mean.



Supplementary Figure 11. Experiment 3. (A) The distribution of the reference dots in each condition. (B) Mean confidence report as a function of target position for each of the four conditions. The black curves represent group mean ± 1 s.e.m. Blue curves represent the model fit averaged across individuals. (C) Model comparisons using ΔAIC: AIC of each model compared with the Difference model. The bars represent ΔAIC summed across participants. The error bars represent 95% bootstrapped confidence interval.



Supplementary Figure 12. Model comparison between the Difference model and the Ratio model using AIC. Sensory noise and Dirichlet decision noises are implemented in both models. (A) Experiment 1 (B) Experiment 2 and (C) Experiment 3. The bars represent ΔAIC (AIC of each model compared with the Difference model) summed across participants. The error bars represent 95% bootstrapped confidence interval..

Supplementary References

1. Shi L, Griffiths TL, Feldman NH, Sanborn AN. Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic bulletin & review* **17**, 443-464 (2010).

2. Gershman SJ, Vul E, Tenenbaum JB. Multistability and perceptual inference. *Neural computation* **24**, 1-24 (2012).

3. Fiser J, Berkes P, Orbán G, Lengyel M. Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences* **14**, 119-130 (2010).