PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Suprathreshold perceptual decisions constrain models of confidence

Shannon M. Locke[1]*, **Michael S. Landy**[2,3], **Pascal Mamassian**[1]

1 Laboratoire des Systèmes Perceptifs, Département d'Études Cognitives, École Normale Supérieure, PSL University, CNRS, Paris, France, **2** Department of Psychology, New York University, New York, New York, United States of America, **3** Center for Neural Science, New York University, New York, New York, United States of America

* ShannonLocke@protonmail.com

Citation: Locke SM, Landy MS, Mamassian P (2022) Suprathreshold perceptual decisions constrain models of confidence. PLoS Comput Biol 18(7): e1010318. https://doi.org/10.1371/journal.pcbi.1010318

Editor: Christoph Mathys, Scuola Internazionale Superiore di Studi Avanzati, ITALY

Received: December 13, 2021

Accepted: June 19, 2022

Published: July 27, 2022

Data Availability Statement: All data and data analysis code are available on a OSF repository at https://osf.io/k2nhq/ (DOI: 10.17605/OSF.IO/K2NHQ).

Funding: This work was supported by NIH Grant EY08266 (MSL, SML; https://www.nih.gov/), National Science Foundation Collaborative Research in Computational Neuroscience Grant 1420262 (MSL, PM, SML; https://www.nsf.gov/), an Anneliese Maier Award from the Alexander von Humboldt Foundation (PM, SML; https://www.humboldt-foundation.de/), a post-doctoral

Abstract

Perceptual confidence is an important internal signal about the certainty of our decisions and there is a substantial debate on how it is computed. We highlight three confidence metric types from the literature: observers either use 1) the full probability distribution to compute probability correct (Probability metrics), 2) point estimates from the perceptual decision process to estimate uncertainty (Evidence-Strength metrics), or 3) heuristic confidence from stimulus-based cues to uncertainty (Heuristic metrics). These metrics are rarely tested against one another, so we examined models of all three types on a suprathreshold spatial discrimination task. Observers were shown a cloud of dots sampled from a dot generating distribution and judged if the mean of the distribution was left or right of centre. In addition to varying the horizontal position of the mean, there were two sensory uncertainty manipulations: the number of dots sampled and the spread of the generating distribution. After every two perceptual decisions, observers made a confidence forced-choice judgement whether they were more confident in the first or second decision. Model results showed that the majority of observers were best-fit by either: 1) the Heuristic model, which used dot cloud position, spread, and number of dots as cues; or 2) an Evidence-Strength model, which computed the distance between the sensory measurement and discrimination criterion, scaled according to sensory uncertainty. An accidental repetition of some sessions also allowed for the measurement of confidence agreement for identical pairs of stimuli. This N-pass analysis revealed that human observers were more consistent than their best-fitting model would predict, indicating there are still aspects of confidence that are not captured by our modelling. As such, we propose confidence agreement as a useful technique for computational studies of confidence. Taken together, these findings highlight the idiosyncratic nature of confidence computations for complex decision contexts and the need to consider different potential metrics and transformations in the confidence computation.

fellowship from the Fyssen Foundation (SML; http://www.fondationfyssen.fr/en/), as well as the French ANR grants ANR-18-CE28-0015-01 "VICONTE" and ANR-17-EURE-0017 "FrontCog" (PM, SML; https://anr.fr/en/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Author summary

The feeling of confidence in what we perceive can influence our future behaviour and learning. Understanding how the brain computes confidence is an important goal of researchers. As such, researchers have identified a host of potential models. Yet, rarely are a wide range of models tested against each other to find those that best predict choice behaviour. Our study had human participants compare their confidence for pairs of easy perceptual decisions, reporting if they had higher confidence in the first or second decision. We tested twelve models, covering all three types of models proposed in previous studies, finding strong support for two models. The winning Heuristic model combines all three factors affecting choice uncertainty with an idiosyncratic weighting to compute confidence. The other winning model uses a transformation where the strength of the sensory signal is scaled according to sensory uncertainty. We also assessed the agreement of confidence reports in identical decision scenarios. Humans had higher agreement than almost all model predictions. We propose using confidence agreement intentionally as a second performance benchmark of model fit.

Introduction

Perceptual confidence is a metacognitive judgement accompanying a perceptual judgement that is thought to reflect the observer's belief about the quality or correctness of their perceptual decision. For example, a person is likely to have more confidence in their judgement of whether the road ahead bends left or right on a bright sunny day than on a foggy evening because the chances of making a mistake are higher in the latter scenario. Confidence is a decision about a decision, and so it is often referred to as a *Type 2* judgement to contrast it with the *Type 1* perceptual judgement [1]. Confidence judgements are ubiquitous in everyday life and are one of the many metacognitive evaluations that guide behaviour and learning [2–4].

Numerous models of perceptual confidence have been proposed by researchers, often to capture distinct aspects of decision-making behaviour. For example, some models focus on the relationship between the speed of the perceptual decision and confidence [5] and others on the degree to which confidence reports can distinguish correct from incorrect decisions [6]. The majority of confidence models are process models that extract the confidence decision variable from either the original Type 1 decision process, or a comparable decision process, such as a partially or completely independent reconstruction [7] or further evolved Type 1 decision process [8]. The output of these confidence models is a confidence decision variable that can then be mapped to a behaviour (e.g., pressing a button on a keyboard). Confidence decision variables used by researchers in their modelling can be categorised into three main metric types (Table 1): Probability, Evidence-Strength, and Heuristic.

Probability confidence metrics are consistent with a common definition that confidence is the probability that the perceptual decision is correct [9]. This is also referred to as statistical confidence, or Bayesian confidence, if it is computed from a Bayesian posterior probability [10–12]. To compute a probability metric, the observer must consider the probability of all the possible states of the stimulus consistent with their perceptual choice. For example, if they reported a motion direction as clockwise of vertical, this would be the probability of the stimulus being anywhere between $0 - 180^{\circ}$ clockwise of the decision boundary. Thus, computation of a full probability distribution is necessary for this confidence decision variable. Bayesian Probability-metric models are typically used as an upper benchmark of confidence

Table 1. Some implementations of the three most common confidence metric types used for modelling perceptual confidence and example study that used this model.

Probability	Evidence-Strength	Heuristic
• **Bayesian confidence**: posterior probability of being correct [11] • **Log-Probability-Ratio**: of the posterior probability or likelihood [11] • **Two-best**: posterior probability difference of the top two choice alternatives [12] • **Entropy**: the uncertainty across the posterior distribution of all choice alternatives [12]	• **Extended SDT**: distance between measurement & criterion [6, 13] • **Drift Diffusion**: diffusion-process state & elapsed time w./w.o. additional accumulation [8] • **Balance-of-Evidence**: relative final state of separate accumulators [16] • **Supporting-evidence**: strength of decision-consistent evidence only [18, 19]	• **Stimulus Variability**: over- or under-weighting of external noise as a cue [14] • **Reaction-time**: between stimulus onset and response [15] • Other stimulus cues to task difficulty [17]

https://doi.org/10.1371/journal.pcbi.1010318.t001

performance, as they consider all available information to assess the probability of being correct for reporting their confidence.

In contrast, *Evidence-Strength* confidence metrics are derived from point estimates in the Type 1 decision process and so do not require the representation of full probability distributions. One common Evidence-Strength metric is the *Distance-From-Criterion* (DFC) metric typically used in extended Signal Detection Theory (SDT) models of confidence [13, 20]. In the extended SDT framework, confidence is monotonically related to the unsigned distance between the sensory measurement and the decision criterion. In the case of categorical confidence judgements (e.g., binary report, scale, etc.), additional confidence criteria delineate the mapping between distance and confidence [6, 7, 21]. Most DFC-metric models also allow this sensory measurement to differ from that used for the perceptual decision [22], either by additional confidence noise corrupting the measurement [23–25] and/or from altering the measurement with parallel decision processes [7, 26]. Variants of the DFC metric have been proposed to allow a point estimate of sensory uncertainty to remap the confidence criteria [11, 27, 28], or allow biases in the distance measures to reflect consideration of only choice-congruent evidence [19].

Another common Evidence-Strength metric is the *Accumulator* metric. In accumulation-to-bound models, the Type 1 decision process is represented with one or more decision variables that are updated according to incoming sensory evidence [29]. In this manner, accumulation-to-bound models capture both the final choice and the temporal dynamics of that choice (e.g., reaction time). Thus, both the distance of the Type 1 decision variable from the initial decision state and total decision time are factored into the computation of the confidence decision variable. Often a mapping that corresponds to the probability of being correct is selected by experimenters [5, 30]. That is, the observer uses a point estimate of the decision process and a well-calibrated mapping function to extract the same confidence decision variable they would have got if they computed it from a full probability distribution. Other confidence decision variables of this type consider the relative states of two or more separate accumulators at the time of the decision [16] or partially interacting accumulators [18]. Extended versions of the accumulation-to-bound models allow the final state of the decision variable for the perceptual judgement and the confidence judgement to differ based on additional evidence accumulation between the Type 1 and Type 2 reports [8, 31].

The third metric type is the *Heuristic* metric, an ever-expanding category of confidence decision variables that have been created to capture behaviour beyond that predicted by the favoured standard models mentioned above. While in some cases, the heuristic label has been applied to variants of the Evidence-Strength metric [12, 19, 32], the majority of "heuristic"

confidence models have focused on the use of stimulus cues to infer decision uncertainty. Support for this method of computing confidence comes from a series of studies demonstrating that observers over- or under-weight external noise in the stimulus when reporting confidence [14, 33–35]. That is, they display a dissociation between Type 1 and Type 2 performance such that, if perceptual performance is matched for two stimuli with different levels of external noise, confidence is not equated. Other identified heuristic cues are reaction time [15] and task-difficulty variables [17]. Importantly, Heuristic metrics do not require access to the Type 1 decision process, and reflect learned associations between stimulus features or difficulty cues and confidence, which may or may not be well calibrated to reality. However, we note that for all the evidence of heuristic cue use, participants do not always rely on stimulus-uncertainty cues when they are available [36, 37].

These variations in the confidence computation alter how stimulus strength and sensory uncertainty influence confidence. To take an example from our task, consider observers trying to infer the mean of an invisible generating distribution from some generated dot samples, to judge if the mean is left or right of centre. Here, the position of the mean is the stimulus strength, and the sensory uncertainty is inversely related to the number of dot samples drawn. To understand the effect of sensory uncertainty on the confidence decision variable, we will consider two dot-cloud stimuli, one with 2 dots generated from a mean on the left (i.e., high sensory uncertainty) and the other with 3 dots generated from a mean on the right (i.e., low sensory uncertainty). In both cases, the means are equidistant from the centre and the observer correctly identifies the laterality of the generating distribution from the presented dot cloud. But how does their sense of confidence compare in these two scenarios?

Fig 1A depicts how an observer using a probability metric would assign higher confidence in the 3-dot scenario, as the probability of being correct (i.e., the shaded region) is greater due to the difference in spread (i.e., sensory uncertainty). Note that the two probability distributions are centred on the two sensory measurements, which, in this case, are the most likely sensory measurements and so are equidistant from the unbiased decision criterion (i.e., are matched in terms of stimulus strength). The behaviour of the Probability metric for different stimulus strengths also depends on whether any transformations are applied. As shown in Fig 1D, the probability metric approaches a ceiling of 100% probability of being correct if stimulus strength is increased. Even with higher sensory uncertainty, this ceiling is approached relatively quickly with slightly greater stimulus strengths. However, the probability of being correct for a binary Type 1 decision could also be expressed as a Log-Probability Ratio (LPR), which is the log of the ratio of the probability of being correct for the selected Type 1 choice to the probability of being correct if you had selected the other Type 1 option [11, 13, 38]. In this scenario, the confidence metric is unbounded, and the difference in the metric between low and high uncertainty increases for larger stimulus strengths (see Fig 1E). So while the relationship of greater confidence for lower uncertainty is typically preserved, the underlying confidence decision variables can differ dramatically depending on transformations in the computation.

An observer who uses the standard DFC Evidence-Strength metric would not behave the same as an observer using the Probability metric. As both measurements have the same DFC, the observer would have the same degree of confidence for the 3-dot and 2-dot scenarios. It is only if the observer scales this measurement according to the degree of uncertainty (i.e., computes the signal-to-noise ratio (SNR)) using a point-estimate of sensory uncertainty, would confidence be greater for the 3-dot scenario (Fig 1B). The effect of this scaling means the sensory measurement is now further from the decision boundary for the the 3-dot scenario compared to the 2-dot scenario. A similar reasoning would apply for an accumulation-to-bound Evidence-Strength metric, where the sensory measurement (i.e., final decision state) and decision time (proportional to uncertainty) jointly determine the degree of confidence. The value

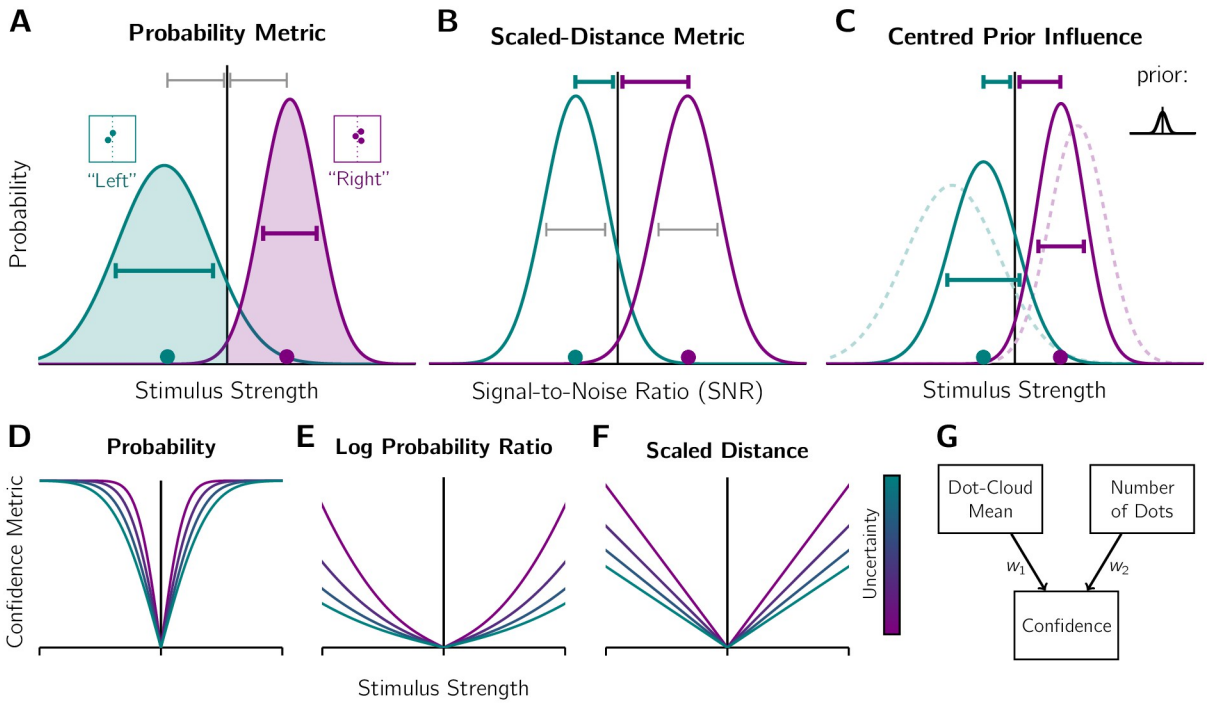

Fig 1. Sensory uncertainty and stimulus strength in the confidence computation. A-C) Extracting the confidence decision variable from the Type 1 decision process. Scenarios contrasted: dot cloud with 2 dot samples (teal; high uncertainty; see inset for stimulus) and a 3-sample dot cloud (purple; low uncertainty), with both dot-clouds' means equidistant from the screen centre (dotted line of inset). The observer correctly identifies the lateralisation of the generating distribution's mean, as shown by their response in quotation marks. Vertical line: decision boundary. Horizontal lines: similarities (grey) or differences (coloured) between the distributions. A) Probability metric. Curves: normalised likelihood functions of the distribution mean, given the sensory measurement (marker; the most likely measurement selected for illustrative purposes). Shaded region: probability of the judgement being correct. The shaded region is greater for the 3-dot scenario, so the observer has higher confidence in this judgement. B) Scaled-distance metric. A signal-to-noise ratio transformation is applied to the distributions in A (i.e., rescaled to units of standard deviation while the areas under the curve on either side of the Type 1 criterion are preserved). The rescaled sensory measurement in the 3-dot scenario has a greater Distance-from-Criterion (DFC) and is judged as more confident. C) Influence of a centred prior (see inset). The posterior distributions (continuous curves), computed according to Bayes' Rule, are differentially shifted towards the centre from the likelihood function locations (dashed). The 2-dot scenario is shifted more because of its higher uncertainty. Consequently, Probability metrics and DFC metrics yield higher confidence for the 3-dot scenario. D-F) How the confidence metric is affected by stimulus strength and sensory uncertainty (low to high uncertainty represented by colours ranging from purple to teal). The greater the stimulus strength (i.e., distance of mean from centre), the larger the confidence metric. However, raw probability values asymptote at 100% confident (D), whereas the confidence decision variable is unbounded if a Log-Probability-Ratio transformation is applied to the probability of being correct (E) or an Unscaled- or Scaled-Distance metric is used (F). Sensory uncertainty affects the rate of change of the confidence metric in response to changes in stimulus strength. Note *y*-axes have been rescaled for illustrative purposes. G) Heuristic confidence metric computed from estimates of stimulus strength (dot-cloud mean) and sensory uncertainty (number of dots) without consideration of the Type 1 process. The observer sets the weights on these factors, w_1 and w_2.

https://doi.org/10.1371/journal.pcbi.1010318.g001

of a DFC Evidence-Strength metric scales linearly with stimulus strength. In the case of a SNR scaling, the sensory uncertainty affects the slope (Fig 1F).

Fig 1C depicts yet another way sensory uncertainty can influence confidence. If the prior expectation about the stimulus is concentrated at the decision boundary (e.g., the generating mean is likely to be at or near the screen centre), the current observation can be biased towards this expected stimulus. This enhances the difference between the 3-dot and 2-dot scenarios for both a Probability metric and a DFC Evidence-Strength metric. In fact, even an unscaled DFC Evidence-Strength metric would now reflect greater confidence for the 3-dot scenario due to the prior biasing effect being stronger for measurement from the 2-dot scenario. Note that we are defining different states of the Type 1 decision process in Fig 1A–1C, with the metric type reflecting different degrees of access or use of the decision-process information by the

metacognitive system (i.e., full probability distributions or only point estimates). Thus, there is no conflict when an observer, who uses the prior for their Type 1 decision, only uses a point estimate of a biased sensory measurement to compute an unscaled DFC metric for confidence.

Finally, an observer computing a heuristic confidence metric considers the number of dots as well as mean position of all dots as independent inputs without considering an underlying Type 1 process. They then combine these inputs with some weighting scheme of their choice (Fig 1G). If they correctly apply the rule of lower confidence for fewer dots, even if it imperfectly captures the relationship between number of dots and external noise, there will be an effect of sensory uncertainty on confidence.

Despite the diversity of potential confidence models, only a few studies have directly compared models of different metric types. One approach has been to compare the behavioural signatures of competing models. Studies using this approach have tended to compare a Probability metric with a Heuristic metric, finding support for the probability metric [10, 36] or differing support for each metric depending on the perceptual task [35]. Formal model comparisons, on the other hand, have been focused on Probability versus Evidence-Strength metrics with mixed results (note that this is according to our metric definitions not those of the authors who used the term "heuristic" more liberally). Aitchison et al. [32] investigated simultaneous versus sequential Type 1 and Type 2 reports, finding the Bayesian-confidence probability metric best fits sequential reports, but this metric and an Evidence-Strength metric could explain simultaneous reports equally well. Adler & Ma [11] investigated categorisation behaviour when category means differed but variances were matched versus when category variances differed but means were matched. They found that an Evidence-Strength metric with quadratic sensory-uncertainty-dependent bounds best fits the behaviour in both tasks. Lisi et al. [39] tested if confidence could be used in a subsequent decision where the perceptual evidence depended on the correctness of previous perceptual choice, and found that the model with a discrete Evidence-Strength metric outperformed the Bayesian Probability-metric model. Finally, Li & Ma [12] investigated various Probability and Evidence-Strength metrics in the context of a three-alternative forced-choice task. They found a Probability metric that compares the posterior probabilities of the two best alternatives outperformed other metrics. Thus, even with direct comparison of the models, the evidence is mixed for the different metric types and further work is needed to understand the nature of the confidence decision variable. This need was highlighted recently by visual metacognition researchers who stated that determining how confidence is computed with detailed and falsifiable models is important for the field [40].

The aim of the present study was to compare the fit of confidence models of all three metric types to the behaviour of humans performing a visual decision-making task. Specifically, observers had to infer if the mean of a dot-generating distribution was left or right of centre (Fig 2A). We employed a mixed-difficulty design, varying the position of the mean (i.e., stimulus strength), as well as the spread of the dot-generating distribution (the *quality* manipulation) and the number of dots drawn (the *quantity* manipulation). Both of the quantity and quality manipulations affected sensory uncertainty. This design allowed us to assess how the observer took sensory uncertainty into account in the computation of confidence. We used the confidence forced-choice method [26, 41]: after two consecutive perceptual decisions, the observer reports whether they had greater confidence in the first or second decision. This allowed us to investigate suprathreshold perceptual decisions where confidence is typically high without being concerned with ceiling effects (e.g., always reporting high confidence or the highest scale rating). We targeted this difficulty regime because the confidence metrics are particularly divergent for more extreme stimulus strengths (Fig 1D–1F). We selected seven base models (twelve models total considering variants in the prior distribution) that captured

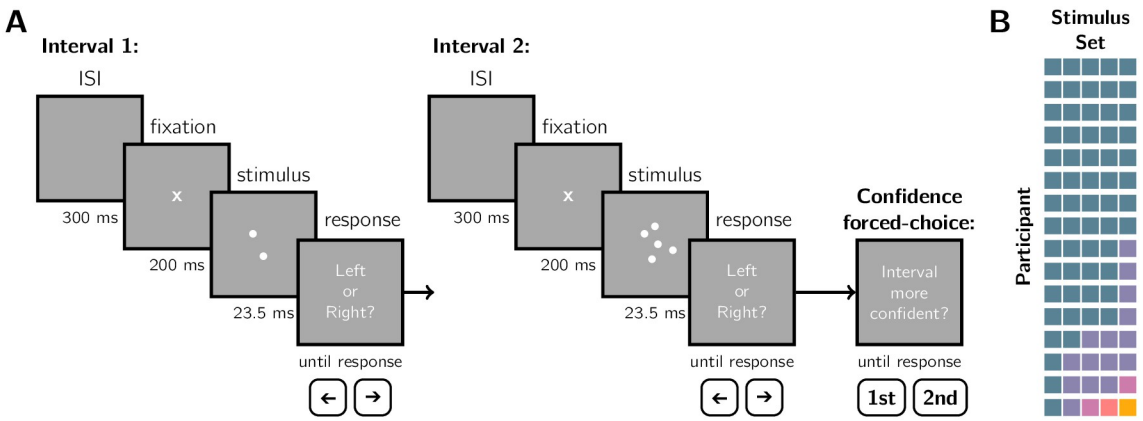

Fig 2. Experimental methods. A) Task design. Observers were shown dots drawn from a Gaussian generating distribution with seven possible horizontal spatial offsets between ±4˚ from the screen centre. They judged if the distribution mean was left or right of centre. After each pair of perceptual decisions (Interval 1 and Interval 2), they reported the interval in which they had higher confidence in their decision. There were six levels of stimulus uncertainty, defined by information quantity (number of dots: 2 or 5) and quality (sampling distribution SD: 1.5, 2, or 2.5˚), presented in an interleaved design. B) Occurrence of unique stimulus sets (colour coded). Squares: stimulus set in a single session. Sets are ordered by type, not session order, and participants by frequency of stimulus set repeats (5-pass to 1-pass).

https://doi.org/10.1371/journal.pcbi.1010318.g002

the diversity of potential confidence computations illustrated in Fig 1 (models described in Table 2). In addition to formal model comparison, we also investigated a new, qualitative, behavioural signature of confidence: *confidence agreement*. Due to a coding error, many sessions were identical repeats, in some cases as many as 5 repeats (Fig 2B). We compared the confidence agreement of the observers to the confidence agreement of model simulations using the best-fitting parameter values as a benchmark of model fit.

To preview the results, we found that the confidence judgements were affected by both the quantity and quality manipulations and all three metric types were supported by at least one

Table 2. Summary of the seven base models. Models spanned all three metric types and considered different implementations of confidence noise, prior distributions, and confidence-variable transformations. Standard unbiased Gaussian confidence noise is used unless noted otherwise. There are twelve distinct models for comparison when including the prior variants.

Model	Type	Description
Ideal Conf. Observer	Probability	Compares the p(correct) for each decision, computed from the posterior distribution. Only the centred-prior variant considered for this model.
Basic Probability	Probability	Compares the p(correct), but with early confidence noise (beta, constrained [0, 1]) applied before the comparison, interval bias, and two prior variants considered (flat & centred).
Probability Difference	Probability	Compares the p(correct), but with late confidence noise applied after the comparison, interval bias, and two prior variants considered (flat & centred).
Log Probability Ratio	Probability	Compares the p(correct), but with a LPR transformation applied, confidence noise, interval bias, and two prior variants considered (flat & centred).
Unscaled Distance	Evidence Strength	Compares the DFC of point-estimates of the Type 1 decision process, with confidence noise and interval bias. Two prior variants considered (flat & centred).
Scaled Distance	Evidence Strength	Compares the DFC of SNR-scaled point-estimates of the Type 1 decision process, with confidence noise and interval bias. Two prior variants considered (flat & centred).
Heuristic	Heuristic	Compares a weighted sum of the estimated difference in stimulus strength and estimate difference in sensory-uncertainty factors, with late confidence noise and interval bias.

https://doi.org/10.1371/journal.pcbi.1010318.t002

observer. The overall best-fitting model at the group level was the Heuristic model, followed closely by a DFC Evidence-Strength model where evidence strength was scaled by the sensory uncertainty. Both of these models best-fit an equal number of participants. Together, our results indicate heterogeneity in confidence strategy; observers were unlikely to compute full probability distributions for confidence and were more likely to rely on stimulus-based heuristic cues or summary statistics of the decision process. For confidence agreement, the best-fitting model, on a per-participant basis, almost always underestimated the observers' confidence agreement. This finding suggests there is still room for improvement in the modelling of perceptual confidence.

Results

Confirming that the stimulus manipulations affected confidence

First, we examined whether the confidence reports meaningfully distinguished accuracy in the Type 1 spatial task (Fig 3A). Observers had an overall high level of accuracy of 0.91 ± 0.002 (mean ± SEM, $\mu_{cloud} = 0^{\circ}$ trials excluded from calculation). Despite this near-ceiling performance, the interval chosen as more confident was on average more likely to be correct than the declined interval in the pair: 0.95 ± 0.003 versus 0.85 ± 0.004 ($t_{15} = 14.96$, $p < 0.01$). The low variance in performance across observers was due to two factors: 1) the exclusion of the ambiguous $\mu_{cloud} = 0^{\circ}$ trials leaving only the easy near-ceiling conditions; and 2) error trials being largely due to the sampled dots favouring the side opposite to the mean and therefore common responses were given across repeated sessions and different observers. The latter factor is evident from the even better performance results if responses are scored according to the mean of the dots displayed on the screen (i.e., the centroid), removing the factor of external

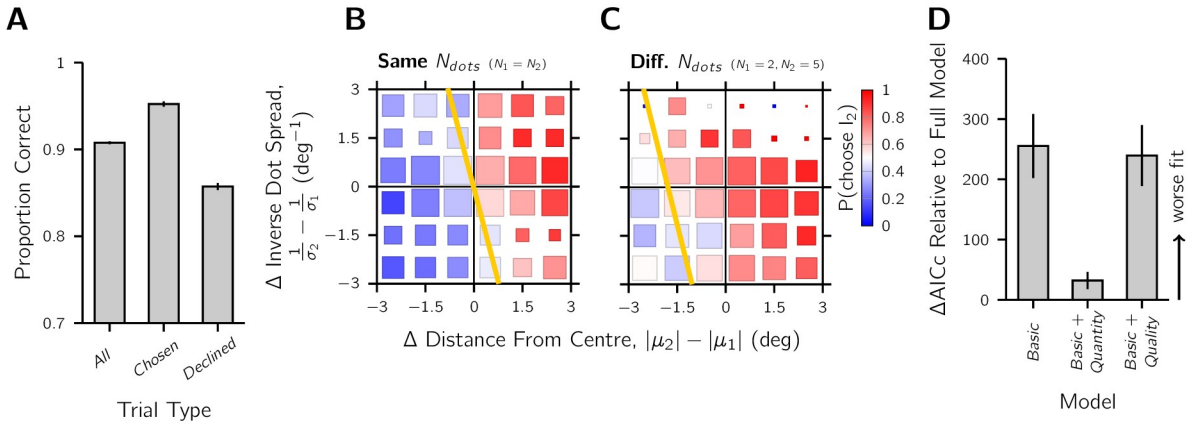

Fig 3. Confidence manipulation checks. A) Suprathreshold Type 1 task performance. The mean proportion of correct spatial judgements across observers is shown for different trial categories. The left-most bar shows performance for all trials, and the next two bars for trials sorted by confidence, either being in the interval chosen as more confident or declined. B-C) Raw confidence choices, sorted by stimulus properties across the two intervals of a confidence pair. The colour code represents the proportion of "Interval 2" as more confident choices averaged across observers. Confidence choices are plotted as a function of the difference in distance from centre across intervals and difference in inverse dot spread. The distance from centre and dot spread were calculated using the empirical mean and SD of the dots displayed, and binned in the range ±3˚ for plotting. Gold line: the confidence-indifference contour, where the observer is equally likely to report Interval 1 or 2, calculated from the Full model in the nested logistic regression analysis. B) Comparisons where the number of dots was the same in each interval. C) Comparisons where the number of dots differed, with stimulus information and confidence selectively flipped so that Interval 2 has more dots for plotting purposes. D) Model comparison for the nested logistic regression analysis. AICc scores are reported relative to the Full model (winner) that contained both quantity and quality predictors. The Basic + Quantity and Basic + Quality models only contained one of these predictors and the Basic model contained neither. The results show that both the quantity and quality manipulations affected confidence. Larger positive scores indicate a worse fit. Error bars: ±SEM.

https://doi.org/10.1371/journal.pcbi.1010318.g003

noise on performance: 0.94 ± 0.003 for all, 0.98 ± 0.003 for chosen, and 0.90 ± 0.004 for declined ($t_{15} = 19.55$, $p < 0.01$). In sum, these results indicate that observers made meaningful confidence forced-choice judgements.

The effect of the quantity and quality manipulations on confidence can be seen in the raw response data (Fig 3B and 3C). In the plots, positive difference values favour Interval 2 as more confident choices (red) and negative differences Interval 1 choices (blue). For all trials, the further the dot cloud was from the screen centre, compared to the stimulus of the other interval, the more likely it was to be chosen as more confident. When the two intervals had a different number of dots (Fig 3C), the interval that contained the greater quantity of dots was more likely to be chosen as more confident. Note in Fig 3C, the interval with more dots was coded as Interval 2 for plotting purposes. Finally, when the two intervals differed in dot-cloud spread, computed as the horizontal standard deviation in the presented dots, an interval was more likely to be chosen as more confident if it had the smaller spread. Smaller spread corresponds to larger inverse spread ($1/\sigma$) in Fig 3B and 3C.

To quantitatively confirm both the quantity and quality manipulations affected confidence, we performed a nested logistic regression analysis. The full model, which had the difference in distance from the screen centre, inverse dot spread, and number of dots as predictors, outperformed simpler models without the quantity and/or quality predictors (Fig 3D). See Fig B in S1 Text for more details on the models and model comparison. The gold confidence-indifference lines in Fig 3B provide a visualisation of the fit of the full model.

Finally, we investigated if the set-repetition affected behaviour. For each set, we computed the proportion correct and the difference in proportion correct for chosen versus declined trials to reflect discrimination performance and metacognitive sensitivity respectively. Values were normalised using a participant-specific z-score and then sets grouped by their repeat number. There were fewer sets in higher repeat bins: 28, 16, 15, 13, and 8 sets for 1–5 repetitions respectively. Unbalanced 1-way ANOVA tests revealed a significant effect of repetition number on discrimination behaviour ($F_{4,75} = 3.41$, $p < 0.05$) but not metacognitive sensitivity ($F_{4,75} = 1.9$, $p > 0.05$). A multiple pairwise comparison of group means for normalised percent correct reveal that the discrimination performance in the fifth repetition was significantly lower than the second and third repetitions. Given that fifth repetitions were always the final testing session, we interpret this effect as related to motivation in the task being lower on the final day of testing rather than set-repetition specifically.

Together, these results suggest observers' confidence computations were affected by the stimulus strength as well as both sources of sensory uncertainty in this easy perceptual task. To better understand the computation of the confidence decision variable, we next examine several process models that capture the full decision process (i.e., both perceptual and confidence judgements).

Type 1 model comparison results

Four Type 1 models were considered. All models had three free parameters: a perceptual decision criterion (k_1), per-dot sensory noise (σ_{dot}), and a lapse rate (λ). The models differed in two ways: 1) whether the posterior distribution was computed with a flat or centred prior, and 2) if the Type 1 decision variable was the probability that the dot-cloud was on the left/right or the point estimate of the posterior's mode relative to the decision criterion (both metrics are always in 100% agreement). The two prior-variants are also often in agreement about the perceptual choice, except for cases close to a biased decision boundary. Thus, it was unlikely to see large differences in the Type 1 model-comparison results. As expected, we found identical fits for the probability and signed-distance metrics (Fig 4A), and near-identical fits for the prior

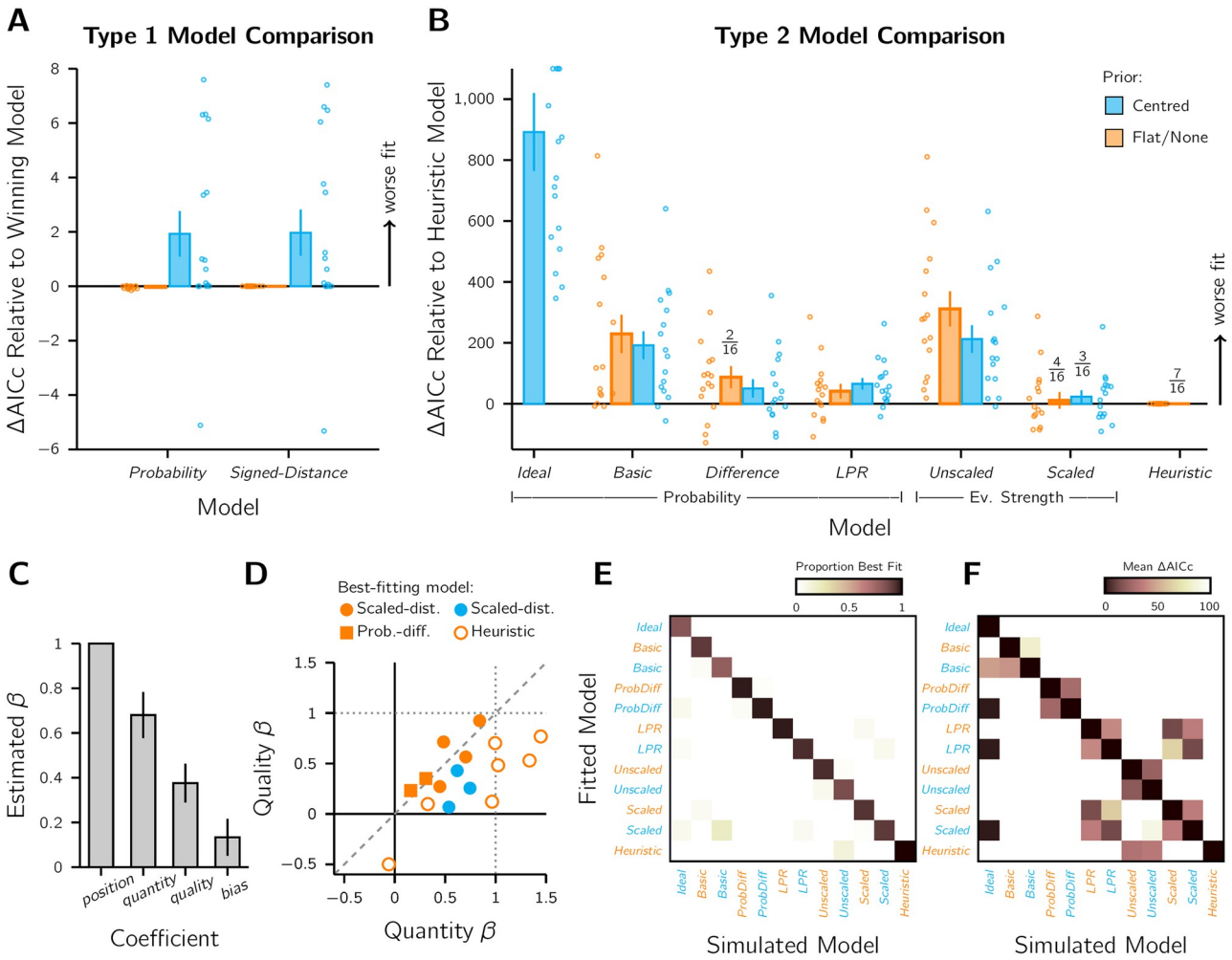

Fig 4. Model fit results (n = 16). A) Relative AICc scores for the Type 1 models. Scores were compared to the winning flat-prior variant models. Bars: average relative AICc score. Markers: Individual participant results. Colour: flat-prior (orange) or centred-prior (blue) variant. Error bars: ±SEM. B) Relative AICc scores for the Type 2 models, with fraction best-fit annotated. Models are grouped by metric type. Note the different y-axis scales between panels A and B. C) Average best-fitting Heuristic-model coefficients across all observers. Note that the position coefficient was always fixed at 1 in the model, but is graphed to illustrate the relative coefficient weights. Error bars: ±SEM. D) The best-fitting quantity and quality coefficients per observer. Marker colour and type indicates the best-fitting model for that observer. Dotted lines: the position coefficient value for comparison. Dashed line: equality line. E) Model recovery results. For each model and observer, 10 data sets were simulated using the participant's best-fitting parameters (1920 data sets total) and then fit by each of the 12 models. The best-fitting model was tallied per dataset per simulated model. Dark squares along the downward diagonal indicate high model recovery success. F) The relative AICc scores compared to the simulated model (downward diagonal of 0).

https://doi.org/10.1371/journal.pcbi.1010318.g004

variants. The purpose of fitting all four models was to ensure that the Type 1 parameters could be fixed accordingly based on the Type-1 responses for the Type 2 model fits.

Type 2 model comparison results

Twelve Type 2 models were compared, with confidence computed with a Probability metric in seven models, a DFC Evidence-Strength metric in four models, and a Heuristic metric in one model (Fig 4B). A summary of the models can be found in Table 2, the model equations in Table 3, and the mean and spread of the best-fitting parameter estimates per model in Table A in S1 Text. The Heuristic model had the best AICc score averaged across observers, with 7 of the 16 observers best fit by this model. The Scaled-Distance models were a strong competitor,

Table 3. Equations for the seven base Type 2 Confidence models. The models consider the decision evidence, Ev(), in favour of the perceptual choice, r, given the dot measurements, X, and Type 1 decision criterion, k_1. The DFC Evidence-Strength models take a point estimate of the posterior's mode ($\hat{\mu}$), with or without scaling by the posterior's spread ($\hat{\sigma}$), and compare its unsigned distance from k_1. Confidence forced-choice judgements involve comparing the relative confidence evidence for the intervals (w_1 versus w_2), with an influence of a confidence interval bias, k_2.

Model	Confidence Evidence	Decision Rule	Confidence Noise	Free Parameters
Ideal Conf. Observer	$w = \text{Ev}(r\|X, k_1)$	$w_2 > w_1$	none	none
Basic Probability	$w \sim \text{Beta}(vp, v(1-p)); p = \text{Ev}(r\|X, k_1)$	$w_2 + k_2 > w_1$	$0 < v < \infty$	v, k_2
Probability Difference	$w = \text{Ev}(r_2\|X_2, k_1) - \text{Ev}(r_1\|X_1, k_1) + \epsilon$	$w + k_2 > 0$	$\epsilon \sim N(0, \sigma^2_{\text{conf}})$	σ_{conf}, k_2
Log Probability Ratio	$w = \lvert\log\left(\frac{p}{1-p}\right) + \epsilon\rvert; p = \text{Ev}(r\|X, k_1)$	$w_2 + k_2 > w_1$	$\epsilon \sim N(0, \sigma^2_{\text{conf}})$	σ_{conf}, k_2
Unscaled Distance	$w = \lvert\hat{\mu} - k_1 + \epsilon\rvert$	$w_2 + k_2 > w_1$	$\epsilon \sim N(0, \sigma^2_{\text{conf}})$	σ_{conf}, k_2
Scaled Distance	$w = \lvert\frac{\hat{\mu} - k_1}{\hat{\sigma}} + \epsilon\rvert$	$w_2 + k2 > w_1$	$\epsilon \sim N(0, \sigma^2_{\text{conf}})$	σ_{conf}, k_2
Heuristic	$w = \Delta\lvert\mu_c - k_1\rvert + \beta_1\Delta N - \beta_2\Delta\frac{1}{\sigma_{\text{emp}}} + \epsilon$	$w + k_2 > 0$	$\epsilon \sim N(0, \sigma^2_{\text{conf}})$	β_1, β_2, σ_{conf}, k_2

https://doi.org/10.1371/journal.pcbi.1010318.t003

with 4 observers best fit by the flat-prior variant and 3 by the centred-prior variant. The average AICc scores for these variants were respectively 10.8 ± 26.9 and 23.1 ± 20.7 higher than the Heuristic model. The flat-prior variant of the Probability-Difference model best fit 2 observers and had an average relative AICc score of 41.0 ± 23.8. Overall, the Ideal, Basic-Probability, and Unscaled-Distance models fit poorly. Model results were relatively unchanged with alternative prior distributions and likelihood functions (see Fig D in S1 Text). A qualitative comparison of the different models in the style of Fig 3B and 3C is shown in Fig C in S1 Text. The comparison reveals that the Heuristic model, Ideal-Confidence-Observer model, and the Unscaled-Distance model have distinct patterns of confidence choice-probabilities.

Examining the Heuristic model fits

We used the best-fitting parameters of the Heuristic model to investigate choice behaviour further. Fig 4C shows the estimated coefficients, which can be compared directly because the predictors were z-scored. Confidence was most strongly determined by the stimulus strength (i.e., dot-cloud position). Of the two sensory-uncertainty manipulations, the quantity of information (i.e., number of dots) was given more weight than the quality of the information (i.e., dot-cloud spread), as can be seen by the coefficients, $\beta_{quantity} = 0.68 \pm 0.10$ versus $\beta_{quality} = 0.38 \pm 0.09$ ($t_{15} = 3.75$, $p < 0.01$) respectively, which are also both significantly different from 0 ($t_{15} = 6.61$, $p < 0.01$ and $t_{15} = 4.34$, $p < 0.01$ respectively). These are contrasted per participant in Fig 4D. There appears to be clustering of coefficient values according to best-fitting model (examined in more detail in Fig E in S1 Text), with only Heuristic-best-fit participants giving more weight to dot quantity than stimulus strength. According to the best-fitting parameters for the Heuristic model across all participants, confidence noise was $\sigma_{conf} = 1.20 \pm 0.17$ and there was a slight but not significant confidence interval bias to choose Interval 2 as more confident ($\beta_{bias} = k_2 = 0.13 \pm 0.08$; $t_{15} = 1.63$, $p > 0.05$; Fig 4C). In the confidence forced-choice paradigm, a confidence bias for one interval could indicate a memory effect (e.g., selecting Interval 2 more frequently because that decision is better remembered), so a non-significant confidence interval bias suggests our task was well-paced enough to avoid memory constraints in the confidence comparison. The reader can see the confidence interval bias of the other model fits in Table A in S1 Text.

We then examined if the Type 1 performance differed between the Heuristic best-fit participants and the others. The estimated sensory noise parameter σ_{dot} was higher in the Heuristic group versus the non-Heuristic group, 1.05 ± 0.04 versus 0.90 ± 0.06, but not significantly ($t_{14} = 1.89$, $p > 0.05$). For a model-independent way of comparing Type 2 performance

between these two groups, we used the difference in accuracy between chosen and declined trials (Fig 3A). The Heuristic group had a significantly lower accuracy difference from the non-Heuristic group (7.94 ± 1.21% versus 10.70 ± 0.28%, $t_{14} = -2.51$, $p < 0.05$), reflecting lower metacognitive sensitivity.

Model recovery analysis

A model recovery analysis further supported our model-comparison results. Data were simulated according to a particular model, with identical experiment structure and parameters consistent with our participants' behaviour. Our simulated data were almost always best-fit by the model that generated them (Fig 4E). This indicates that the models are distinguishable from one another in model comparison. Thus, differences in the best-fitting model across participants likely reflects idiosyncrasies in the confidence computation, detectable due to the high number of trials per participant, rather than measurement noise one would expect in a low-powered design. The centred-prior variant of the Basic-Probability model had the lowest recovery success with 78.75% datasets recovered. The Heuristic model was the highest with a 100% recovery rate. On average the recovery rate was 91.72 ± 6.54% (mean ± SD).

A comparison of the relative AICc scores in Fig 4F shows that some model-simulation fit pairs are more similar in model-fit quality than others, even though the simulated model was almost always correctly recovered. First, flat- and centred-prior variants tended to have similar AICc scores. A similar pattern emerges for the Scaled-Distance and LPR models. Then there are the unidirectional similarities. Datasets generated by simulating the Ideal-Confidence-Observer model could often be well fit by the centred-prior variants of other models. This was to be expected if the model converges to the Ideal-Confidence-Observer model when confidence noise and interval bias approach 0. But, by penalising model complexity, the Ideal-Confidence-Observer model is often recovered. The other unidirectional similarity is between the Unscaled-Distance models and the Heuristic model. This is because the Heuristic model with no weight given to the quantity and quality predictors is the likelihood-variant Unscaled-Distance model. However, penalising the extra complexity of the Heuristic model helps to ensure the recovery of the Unscaled-Distance models.

Confidence agreement

Confidence agreement was quantified by counting the number of the most consistent confidence response on a per-trial basis (e.g, 4 "Interval 1" responses out of 5 passes is 80% agreement). The pattern of confidence agreement averaged across observers shows that 1) observers had high overall levels of confidence agreement, and 2) comparisons close to the confidence-indifference line are less consistent than those far from this line (Fig 5A). We then investigated the predicted confidence agreement according to each of the twelve models, simulated using the best-fitting parameters on a per-participant basis (results of an example participant shown Fig 5B and all participants in Fig F in S1 Text). As expected, the Ideal-Confidence-Observer model always had the highest confidence agreement, because there is no confidence noise in this model. The Basic-Probability models then had the second and third highest levels of confidence agreement, followed by the remaining models, which tended to have more similar levels of confidence agreement. The higher confidence agreement of the Basic-Probability models, however, was not robust. As expected, many of the confidence comparisons had values close to 1 (Fig 1D), with the computer simulations detecting very small differences in probability (e.g., 0.985 versus 0.998). A human observer is unlikely capable of such comparisons, and so to get a more realistic prediction of confidence agreement for the Basic-Probability model, we

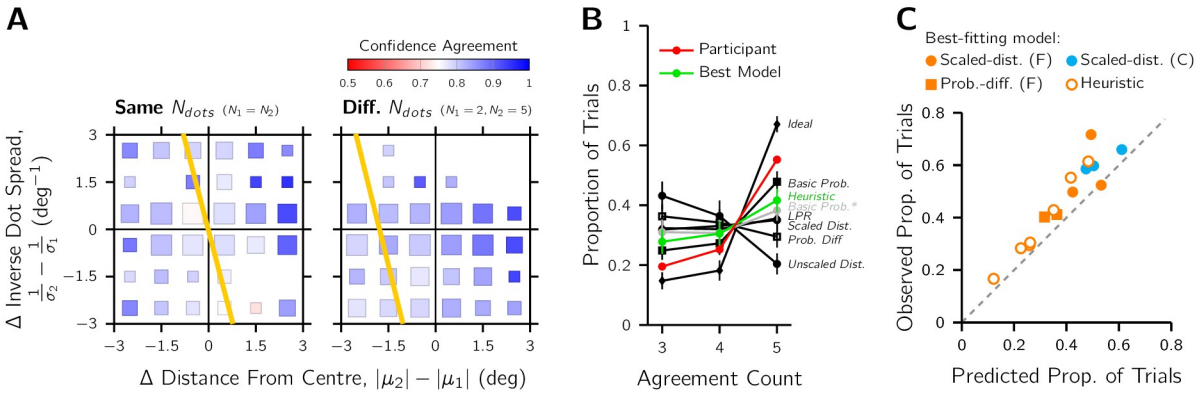

Fig 5. Confidence agreement results ($n = 15$). Confidence agreement was calculated per trial as the proportion of the most-selected confidence choice for the participants who did a 3-, 4-, or 5-pass version of the experiment (Fig 2B). A) Heatmaps of average confidence agreement according to the properties of the two stimuli displayed, pooled across observers. Gold: the indifference lines where each interval is equally likely to be selected as more confident according to the preliminary analyses (Fig 3B). B) Comparing the 5-pass confidence agreement of a representative example participant (red; #11) with the predicted confidence agreement of the models. Green: the best-fitting model for this observer (Heuristic model). Black: other models (flat- and centred-prior variants had similar confidence agreement counts so only the flat-prior variant is shown). Grey: the Basic-Probability model with additional late noise (1% SD). Model predictions calculated from 100 simulated datasets using the participant-specific best-fitting parameters. Error bars: ±2 SD. C) A comparison of the predicted and the observed proportion of trials for the highest agreement count. Each marker is an individual participant, where marker style indicates their best-fitting model ("F" refers to the flat-prior variant and "C" the centred-prior variant). The best-fitting model per observer was used for the confidence-agreement prediction. Dashed line of equality is also shown.

https://doi.org/10.1371/journal.pcbi.1010318.g005

also simulated a version of the model that included 1% SD late noise (grey lines in Fig 5B). The confidence agreement of this model was much closer to the other non-ideal models.

To investigate if the best fitting model captured the confidence agreement of observers, we compared the proportion of trials for the highest possible agreement count. Note that the minimum and maximum agreement depended on whether the participant did a 3-, 4-, or 5-pass version of the task. As the minimum number of passes was 3 (e.g., 2/3 and 3/3 agreement), selecting only a single count value to compare was the only appropriate choice for a statistical comparison. For 14 of the 15 observers, their confidence behaviour was in closer agreement across the repeats than predicted by their best-fitting model (see Fig 5C). A Wilcoxon signed-rank test confirmed that this difference in confidence agreement is significant ($n = 15$, $z = 3.35$, $p < 0.01$).

Discussion

We fit confidence models of all three metric types to confidence forced-choice responses in a suprathreshold spatial-discrimination task. Each metric type was supported by at least one observer, with 44% observers best fit by the Heuristic model (Heuristic metric), 44% by a Scaled-Distance model (DFC Evidence-Strength metric with a SNR transformation), and 12% by the Probability-Difference model (Probability metric with late confidence noise). The Heuristic model was considered the winner overall at the group level. The modelling results suggest there is no universal computational strategy for taking sensory uncertainty into account when judging confidence. All four possibilities (i.e, a probability metric, SNR-scaling, use of a centred-prior, heuristic cue use; see Fig 1) were supported by at least one observer. Taken together, these results suggest that the computation of confidence is highly idiosyncratic in environments where decision uncertainty is influenced by multiple factors, further supporting the hypothesis of a highly individual nature to perceptual confidence [42–44], which may also depend on the task at hand [32, 35]. However, within the metric types, there was a preference

for some computations over others, which has implications for our understanding of the computation of perceptual confidence more generally.

Evidence-Strength metrics

The Unscaled-Distance model performed significantly worse than the Scaled-Distance model, which took into account sensory uncertainty by applying a SNR transformation to the point estimate used to compute the DFC. In other words, a single point estimate of the decision process was insufficient to capture confidence behaviour and a secondary point estimate of sensory uncertainty is used by the metacognitive system. This finding is particularly relevant for the Extended-SDT framework of perceptual confidence. It is often assumed that the sensory measurement is simply compared to static confidence criteria [6, 22], which is no issue due to many SDT tasks being of a fixed difficulty level (note that this is not true of tasks that staircase difficulty, e.g., [25, 45]). In mixed-difficulty designs, however, it has been proposed that confidence criteria are updated according to the level of sensory uncertainty [11, 27, 46]. The reason an observer would do this is to avoid a confidence paradox of more readily assigning high confidence to stimuli with large sensory noise that are more likely to have the measurement fall far from the perceptual decision criterion. Yet, human observers do not shift their criteria appropriately to avoid this paradox [27, 46]. However, without an incentive structure for confidence, confidence ratings are essentially meaningless and there is little motivating accurate shifts, which could explain these results. Though this is not true for the confidence forced-choice judgements as relative comparisons will always have a sensible interpretation even in the absence of an incentive structure.

From a modelling perspective, shifting the criteria to account for sensory uncertainty and the scaled DFC measure have the same effects on the confidence computation. Previous studies likely favoured the shifting-criterion description simply because criterion plasticity is commonly accepted [47], whereas in the current study there were no confidence criteria applied to the confidence decision variable because the two confidence decision variables are directly compared [41]. Our results suggest shifting of confidence criteria may not be necessary as observers can scale DFC measures, however the "stickiness" of criteria has been useful for explaining suboptimal behaviour in decision contexts requiring shifting criteria [27, 48, 49]. To explain suboptimal confidence judgements in this manner within the confidence forced-choice paradigm, one would need to apply noise to the interval bias term (i.e., k_2). However, we interpret a changing propensity to report one interval over another as distinct from traditional criterion biases, which reflect under- and over-confidence. This highlights the need to better understand the metacognitive mechanism for accounting for uncertainty. If researchers wish to apply SNR scaling to an Extended SDT model, we propose using the log-likelihood-ratio representation ($\ln \beta$), which rescales the decision axis by the amount of sensory noise for a given stimulus level [50], as placing perceptual and confidence criteria in this space removes the need to model multiple shifted criteria.

Heuristic metrics

The Heuristic model in this task had to consider three stimulus-based cues when assigning confidence: the position of the dot cloud, the number of dots, and the dot spread. The distance of the centroid from the centre was the predictor given the most weight, followed by the number of dots, then dot spread. Heuristic observers tended to give more weight to the number of dots displayed than the other observers, with 5/7 observers giving almost equal or more weight to this predictor than the position of the dot cloud (i.e., stimulus strength). It is unclear why the number of dots was particularly salient to these observers in this task, given the amount of

previous evidence suggesting that variability, here dot-cloud spread, is a strong heuristic cue [14, 33–35]. However, to our knowledge, no other perceptual confidence task has jointly manipulated both the quantity and quality of sensory information. This suggests that future work interested in investigating heuristic cues in confidence should consider targeting the quantity of information provided.

From a computational perspective, it may not always be possible to distinguish a Heuristic observer from an Evidence-Strength observer. Take, for example, a task that only manipulates stimulus strength. With the same noise sample, the estimate of stimulus-strength magnitude by the Heuristic observer is identical to the Evidence-Strength observer's distance-from-criterion metric as long as the observer is unbiased. Thus, despite different underlying process models, these two observers are likely to display similar confidence behaviour. This could be seen in our model recovery analysis, where simulations of the Unscaled-Distance model (which only considered stimulus strength) could be somewhat well fit by the Heuristic model, though penalising model complexity almost always led to the correct model being recovered. As such, manipulations of sensory uncertainty help disambiguate these metric types, as will strong biases in the Type 1 criterion (e.g., by manipulating stimulus priors or rewards, see [21]). However, varying sensory uncertainty is only helpful if the Evidence-Strength model is constrained appropriately, either by principally relating the relevant stimulus factors to the likelihood uncertainty as we did here, or enforcing the corresponding relationship to the d' of different difficulty levels for Extended-SDT. Otherwise, if likelihood uncertainty or d' values are fit independently for each unique difficulty level, the Evidence-Strength and Heuristic models may appear to fit more similarly.

Probability metrics

In addition to the Probability-Difference model, which best-fit two observers, the Probability metric category contained the Ideal Confidence Observer, the Basic-Probability model, and the LPR model. The centred-prior variants of these models conform to the definition of "Bayesian Confidence" according to the Bayesian Confidence Hypothesis [10–12], and were not a best-fitting model of any participant. It is unclear if the flat-prior variants of these models also conform to the definition of Bayesian Confidence, which is unspecific about whether a true stimulus prior is required. Overall, our results do not support the Bayesian Confidence Hypothesis, in agreement with some previous studies [11, 28, 39] but not others [12, 32]. Despite this finding, the results of the current study should not be interpreted as against Bayesian computations in the brain more generally. The averaging of the perceived dot locations that is the basis of all models we tested is optimal Bayesian cue combination [51]. Additionally, the computation of the posterior mode in the centred-prior variants of the Unscaled- and Scaled-Distance models must also involve a Bayesian computation. This is ambiguous in the case of the flat-prior variants of Unscaled- and Scaled-Distance models, because these models are identical to a model that uses the point estimate from the un-normalised likelihood function.

Another consideration in the computation of confidence is the resolution of probability judgements for confidence evaluations. This can be best seen in the contrast of the Basic-Probability model and the LPR model. If the distribution means in both intervals are far from the centre, the raw probability values will be similarly close to 1 (Fig 1D), but if a LPR transformation is applied, small differences in the relative positions of the distribution means can translate into large differences in confidence (Fig 1E). Effectively, the LPR model has a very high resolution for extreme probability comparisons. Fortunately, this can be assessed in the confidence forced-choice method despite potential distortions in perceived probability [52, 53]

because any distortion would apply to both Interval 1 and 2 and thus not change the relative judgement in confidence, as long as the distortion function is monotonic. In contrast, traditional methods of measuring confidence are likely to suffer from probability distortions in probability judgements or ceiling effects of ratings at high levels of confidence [20]. Our results are not conclusive on this point. While the Probability-Difference model was third best in terms of the number of observers best fit, the LPR model was overall a superior model at the group level and in the model recovery analysis showed some fit similarity to the second best-fitting Scaled-Distance model (Fig 4F). Reviewing previous studies, a high resolution for extreme probabilities does not match with findings of compression of continuous confidence probabilities to a few discrete levels for perceptual confidence [39] or knowledge of motor uncertainty distributions (i.e., motor confidence, [54]). Further work is needed to understand the representation of confidence for extreme probabilities. It is possible that our Probability-metric models were overly simplistic, and models with intermediate resolution, such as by allowing the confidence noise to vary with signal strength or including probability distortions in the mapping function [55], would improve the fit. However, the increased flexibility of such models is also likely to pose a challenge in distinguishing between candidate models.

Use of a centred prior

We were unable to reach a strong conclusion about the use of centred priors in the computation of confidence. Only three observers were best-fit by a centred-prior variant model (Fig 4B), but this increased to six if the stimulus prior assumed perfect knowledge of the three levels of dot spread tested (see Fig D in S1 Text). It is well known that observers do not always adapt perfectly to the experimental environment, causing them to use incorrect priors or priors matching environmental statistics [56, 57], and this may have occurred in the current study.

Heterogeneity in the confidence computation

We found a variety of computational strategies in our sample population. The model-recovery results support the conclusion of model heterogeneity [58, 59] against other possible interpretations, such as high measurement noise. Our results are consistent with the previously reported heterogeneity in confidence modelling, within and across studies [11, 12, 32, 35, 39]. We consider the possibility that task demands influence the adopted strategy generally and in the present study.

First there is the use of mostly suprathreshold Type 1 decisions paired with the confidence forced-choice reporting method. The Heuristic and Evidence-Strength metrics are both good strategies for avoiding confidence indifference in the comparison of two certain choices. This is because often one dot cloud will be further from the centre, have more dots, or a smaller spread. In contrast, comparing near-ceiling probabilities of being correct is more likely to lead to indecision or indifference, which could make the participant feel like they are not doing well in the task. Thus, our task may have influenced the adoption of a specific strategy. However, it is important to consider that suprathreshold decisions and confidence forced-choice are not rare but a common part of real-world decision-making and should be considered for a complete description of confidence behaviour. For example, when one is in the supermarket attempting to select the freshest salad mix of two already decently fresh options, the Heuristic approach of selecting the bag without spinach because it's known to expire quickly will lead to a satisfactory result fast. Broadly, it would be advantageous to be able to switch from more optimal computations to something that can provide a quicker and less complex answer if the situation called for it [60]. The second concern regards the complexity of the confidence report method itself. A misconception about the confidence forced-choice method is that it is more

difficult for the decision-maker than single-trial methods for reporting confidence (e.g., ratings on discrete or continuous scales). Participants in a confidence forced-choice paradigm do have to hold in memory the confidence of the previous choice for the comparison. However, it is unclear if this is more taxing for the participant than keeping in memory multiple confidence criteria [6, 27] or a confidence-response mapping function [35, 61]. Thus, further research would be required to claim method-complexity as a reason for adopting strategies that make use of Heuristic or Evidence-Strength metrics.

If multiple confidence strategies are possible, across or within observers, what then should be the goal of the computational analysis? One important result would be to scope out the general limitations of the metacognitive system. For example, does the metacognitive system have access to the full posterior distribution? Our results suggest this may be possible (i.e., the Probability-Difference observers), but the use of mean and uncertainty point estimates were more common (i.e., the Scaled-Distance observers; also reported by [11]). What is clear is that answering this question involves testing large sets of distinguishable models in a range of decision contexts. Such testing could also be used to construct a systematic description of contextual effects on confidence strategy (e.g., choice difficulty, mixed-uncertainty environments, reward structure, attentional resources, etc.). Combining a multi-model approach with neural measures could also be fruitful for understanding the neural activity associated with the different confidence strategies [57]. For example, a classifier that could independently partition observers according to similar patterns of neural activity could then be compared to the grouping of observers by model best-fit, which could serve as a test of the strategy-heterogeneity hypothesis. Similarly, if activity in one particular area of the brain is correlated with a particular model parameter, we would expect the observers better-fit by another model to show a weaker relationship between that parameter and neural activation. Importantly, the multi-strategy framing of confidence could help researchers avoid unproductively attempting to reach a consensus as to the single model of confidence, if indeed humans employ more than one strategy.

Confidence agreement

The confidence agreement analysis showed that the best-fitting model per observer almost always underestimated the degree of confidence agreement in the N-pass designs. From a modelling perspective, there are three factors that could have limited confidence agreement in the present experiment: 1) Type 1 sensory noise, 2) Type 2 confidence noise, and 3) the resolution of the confidence decision variable. The effect of sensory noise can be observed in the predicted confidence agreement of the Ideal Confidence Observer (Fig 5B), as this was the only factor relevant in the simulation of this model. The remaining models have lower predicted confidence agreement, which is due to the influence of the confidence noise. The Basic-Probability model had higher confidence agreement than the others, likely due to the model using a different noise distribution (Table 3). Examining the model simulations in more detail, we found that many of the confidence comparisons for the Basic-Probability model were between two extremely high probabilities and that the model is not robust to even small amounts of late decision noise. It is unlikely that a human observer has such a high resolution for the confidence decision variable, which is why resolution may be a third factor, effectively discretising confidence into various levels [39]. However, lowering the confidence resolution will decrease confidence agreement, so this alone cannot explain why the confidence models with non-discretised confidence variables under-predicted confidence agreement. It is also not due to an interval response bias, as this was included in the model simulations.

What the Basic-Probability model does illustrate is that the choice of the confidence noise model affects confidence agreement. This suggests that future studies of confidence agreement could investigate different confidence-noise distributions [61, 62], assumptions about noise being independent between the two intervals, partially or fully parallel confidence decision processes [7], and serial-dependence effects on confidence [63]. In general, the N-pass technique offers an interesting additional benchmark for assessing confidence models, just as it has proved useful in better understanding the computations of perceptual decision-making [64–66]. It is important to keep in mind that the inputs to the confidence computation will never be identical due to sensory noise, so not all analyses developed for perceptual decision-making will be applicable to their confidence counterpart. Overall, confidence agreement appears to be a promising evaluation technique, with some researchers already leveraging the power of multiple presentations mostly in the form of trial "replays" to understand confidence [31, 67], so we expect that the use of multiple passes in confidence experiments will only increase.

Strengths and limitations

A key strength of this study was the variety of metric types tested. We succeeded in fitting models of the Probability, Evidence-Strength, and Heuristic types while maintaining reasonable model identifiability (see the model recovery analysis in Fig 4E). In part, this was due to our novel approach of pairing the confidence forced-choice technique [26, 41] with an easy perceptual task of sufficient complexity. Specifically, the models are more divergent for easy trials and the confidence forced-choice method allowed us to probe the confidence decision variable in this range.

A downside of the confidence forced-choice technique is that there is only one confidence report for every two perceptual decisions, doubling the number of perceptual trials needed per participant. In our study, participants completed five hours of testing each on trials with very brief stimulus presentations. Using stimuli that require long presentation times (e.g., random-dot motion) would have a serious impact on the data collection rate and introduce concerns over memory of the decision in the first interval. A consequence of using very brief stimulus presentations is that these stimuli are arguably less suitable for studying the accumulation of evidence. Typically, researchers who investigate the temporal dynamics of decision-making with accumulation-to-bound models use long presentation times or let the observer decide when to terminate viewing the stimulus [5, 30, 31]. For these reasons, we did not investigate accumulation-to-bound models in the present study, despite their better-established link to the neural basis of decision-making [29]. A consequence of this choice is that we neglected to measure the decision reaction times in our study, which in hindsight could have been a relevant cue in the Heuristic model. Future work could contrast accumulator metrics versus Heuristics and DFC-Evidence-Strength metrics. We highlight that very recent research on this topic [62] indicates that the two-stage accumulator model [8] does not out-compete the DFC Evidence-Strength model [61].

The unintentional repetition of sessions resulting in an N-pass design to the experiment had both strengths and limitations. Obvious concerns are a reduction in the number of unique stimuli and stimulus pairings in the task, which affects the quality of the dataset for model fitting. There is also the possibility of the observer recognising the repetition and repeating responses, though none reported noticing the N-pass nature of the task. The clear advantage was the ability to conduct an exploratory analysis of confidence agreement, which revealed an explanatory deficit in the best-fitting models. We recommend researchers in confidence agreement design the N-pass carefully. For example, one can avoid order effects by shuffling the order of repeated pairs [66]. However, with the repeated-sessions technique accidentally employed here, the precise sequence of stimulation was identical, as was the position of the

pair within the session. This could keep constant stimulus history effects [68, 69], response history effects to some extent in the case of a suprathreshold task, and fatigue or motivation changes during the session. However, the researcher would run the risk of the observer noticing the repetition, leading to response memorisation or surprise effects, if the number of trials was low. Regarding motivation or surprise effects, these are less of a concern for the confidence forced-choice method with exactly repeated sets as these effects should apply to both Interval 1 and 2 almost equally. Another consideration is the difficulty of the task. Difficult trials are more likely to lead to Type 1 judgements that differ between passes, whereas in easier designs, such as in the present study, the majority of Type 1 judgements are the same. Finally, we see no reason why confidence-agreement experiments cannot be used with other types of confidence reports besides confidence forced-choice. As such, including at least a 2-pass design into any confidence experiment should be feasible.

Conclusion

By using the confidence forced-choice method, we have shown that observers are not indifferent for easy perceptual choices. Almost half of the observers took sensory uncertainty into account by computing the signal-to-noise ratio (SNR) in the Type 1 decision process, while a similar number used stimulus-based heuristic cues to compute confidence. Overall, this suggests that observers use the Distance-From-Criterion (DFC) Evidence-Strength and Heuristic metrics over Probability metrics, the most notable of which follow the Bayesian Confidence Hypothesis. Furthermore, while heuristic cue use is likely to vary according to the specifics of the individual experiment, the Scaled-Distance model is applicable in any mixed-difficulty design and should be more widely considered. Our results suggest that relying on the simple unscaled metric typically used in extended Signal Detection Theory (SDT) could lead to worse model fits or an unnecessary proliferation of confidence criteria to account for this transformation. An accidental repetition of the presented stimuli also allowed us to capture the confidence agreement of observers; a novel measure of model fit. This analysis revealed that observers are much more consistent than would be predicted by any of the models except the Ideal Confidence Observer. We propose confidence agreement as a readily accessible model-validation benchmark for future efforts in modelling perceptual confidence.

Materials and methods

Ethics statement

This study was approved by the New York University Committee on Activities Involving Human Subjects (IRB-FY2016–595). All participants received details of the experimental procedures and provided written consent prior to participation.

Participants

Sixteen participants (21–43 years old, ten female) with normal or corrected-to-normal vision took part in the study. All participants but two were naive to the design of the experiment.

Apparatus

Stimuli were displayed on a Sony G400 CRT monitor (36 x 27 cm, 1024 x 768 pixel, 85 Hz). Participants sat 55 cm from the monitor with their head stabilised by a chin rest. All responses were entered on a standard computer keyboard. The experiment was conducted using custom-written code in MATLAB version R2014a (The MathWorks, Natick, MA), using Psychtoolbox version 3.0.11 [70–72].

Task

In this task, participants judged if the mean of an invisible dot-generating distribution, a 2D circular symmetric Gaussian, was left or right of centre (i.e., a *Type 1*, perceptual judgement). The distribution had seven possible spatial offsets of the mean (-4, -2, -1, 0, 1, 2, and 4 deg) and three possible standard deviations (1.5, 2, or 2.5 deg). The distribution mean did not deviate vertically from the half-height of the screen and the screen centre was indicated by a fixation cross prior to stimulus presentation. Participants were presented with either two or five independently sampled dots (Gaussian-blobs with 0.1 deg SD). The white dots were simultaneously presented on a mid-grey background for 23.5 ms. The number of dots (i.e., *quantity* manipulation) and the spread of the generating distribution (i.e., *quality* manipulation), produced six levels of sensory uncertainty. Fewer dots or larger spread made the distribution mean more difficult to localise. After every two stimulus presentations (denoted Interval 1 and Interval 2), the participant reported if they had greater confidence that their first decision or second decision was correct (i.e., a *Type 2*, metacognitive judgement). This confidence forced-choice technique avoids over- or under-confidence biases by having the observer report relative confidence [26, 41]. An example trial pair is shown in Fig 2A. The stimulus location, number of dots, and distribution spread were randomised at the level of individual trials in an interleaved design, with trial pairings left to chance. Consequently, confidence comparisons could be between any combination of stimulus strength and sensory uncertainty (42×42 possibilities). In each session, there were 20 presentations per unique combination of stimulus strength and sensory uncertainty. New dots were sampled for each repeat. Participants completed 5 one-hour sessions of 840 perceptual judgements and 420 confidence forced-choice judgements, resulting in a total of 4200 and 2100 judgements, respectively, per participant. During the experiment, no feedback was provided on the correctness of the perceptual decisions. Data from this experiment are available at https://osf.io/k2nhq/.

Stimulus set repetition

Due to a coding oversight, the experiment was conducted with the same random seed for every session. This, coupled with the practice of switching off the testing computer between sessions, led to many of the sessions displaying the exact same sequence of stimuli (Fig 2B). For example, eight participants saw the same identical sequence of dot clouds for all five sessions; three saw this sequence for four out of five sessions; and only one participant was given a unique stimulus set for each session. In the most frequent stimulus set, of the possible pairings of sensory uncertainty for the confidence comparison, the worst sampled combination had seven unique pairs of stimuli. If the interval order is ignored, this number increases to nine unique pairs of stimuli for two same-uncertainty pairings. In the experiment debriefing, no participant reported noticing the repetition of stimulus sets. While this random seed setting limited the richness of the collected dataset, we saw this as an opportunity to investigate the agreement of confidence reports for exactly identical dot-cloud stimuli and thus identical forced-choice comparisons. In our modelling, we were able to leverage these measurements of confidence agreement to perform predictive checks of the model fits.

Models

General modelling framework

We modelled our participants as Bayesian observers making a joint inference about the mean and precision of the dot generating distribution from a noisy observation of the dots presented onscreen. Fig 6A shows the prior, likelihood, and posterior components for an example trial. It

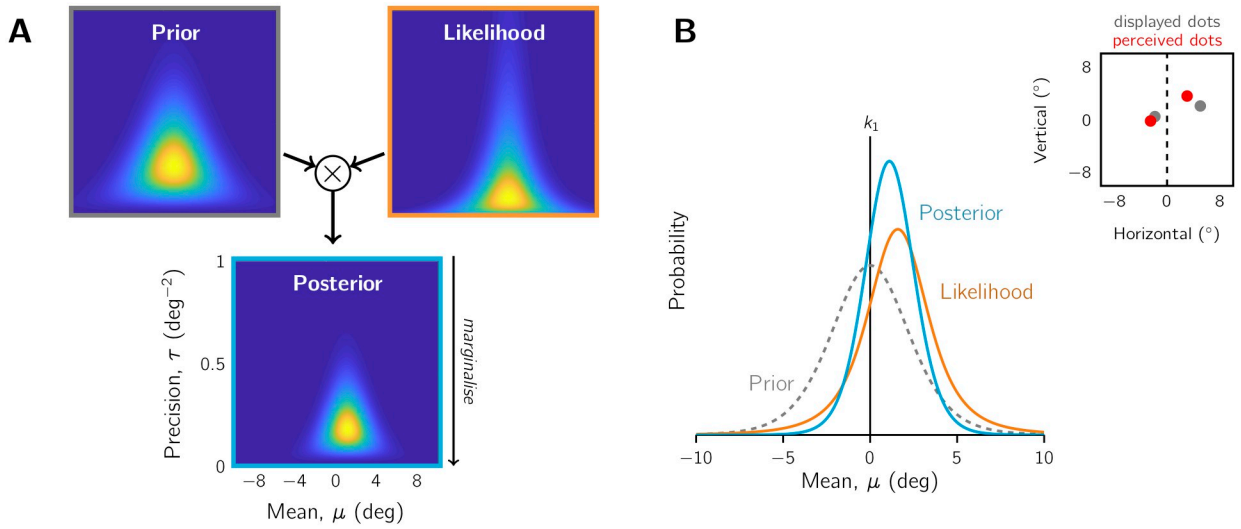

Fig 6. Elements of the decision models. Depicted is the estimated joint probability of the dot-cloud generating distribution mean and precision for an example stimulus presentation of 2 dots. A) Centred Normal-gamma prior distribution (based on the stimulus statistics of the experiment), likelihood function based on the noisy dot observations, and the resulting normal-gamma posterior distribution. All three distributions have the same axes as the ones shown in the posterior panel. B) The marginal prior (grey), likelihood (orange), and posterior (blue) for estimating the generating distribution mean. An unbiased Type 1 decision criterion, k_1, is also depicted. Inset: the displayed and perceived locations of the two sampled dots, with the dashed line indicating the screen midline.

https://doi.org/10.1371/journal.pcbi.1010318.g006

was important to include the inference about the precision of the generating distribution because uncertainty in the estimated mean depends on the precision (note that precision is the inverse variance, $1/\sigma^2$). This can be seen in the triangular shape of the prior, likelihood, and posterior. If the distribution's spread is small, precision is large, and the uncertainty in the distribution's mean is low. To make a spatial judgement, the observer marginalises over all possible precision values to get the probabilistic representation shown in Fig 6B. From this they decide whether the evidence favours a distribution mean to the left or the right, and subsequently their confidence. The advantage of using the same general framework for models of all three metric types (Probability, Evidence-Strength, and Heuristic) was that conclusions drawn from the model fits were more likely to reflect the confidence computation than "nuisance" differences between different modelling frameworks.

Decision context

There are two categories of stimuli in the perceptual task: left ($C = L$) and right ($C = R$). For each interval, the observer reports their belief about stimulus category (i.e., $\hat{C} = L$ or $\hat{C} = R$), as their Type 1 response based on their noisy sensory measurements of the sampled dots as well as any prior beliefs about the underlying generative sensory process. For the Type 2 confidence judgement, the observer compares the probability of their Type 1 perceptual decisions being correct in Interval 1 and 2: $p(\hat{C}_1 = C_1)$ versus $p(\hat{C}_2 = C_2)$.

In a single interval, a dot sampling distribution, defined by its horizontal mean (μ_{cloud}) and spread (σ_{cloud}), is drawn from one of the respective categories with equal probability: $P(C_L) = P(C_R) = 0.5$. The exception being the case where $\mu = 0$, which favours neither category and we treat as coming from either category randomly. From the sampling distribution, N dots are independently drawn, represented by the vector of horizontal dot locations, $\boldsymbol{D} = d_1, d_2, \ldots, d_N$. During the measurement process, additive sensory noise is applied per dot, $\epsilon \sim N(0, \sigma^2_{dot})$,

represented by the vector of noisy horizontal dot-location measurements, $\boldsymbol{X} = x_1, x_2, \ldots, x_N$. We assume the observer knows N (either 2 or 5) but not σ_{cloud} of the sampling distribution on any given trial. Thus, the observer's Type 1 task is to infer the category of stimulus with uncertain sampling distribution mean and spread based on the observed dot samples.

The likelihood function

The observer must consider two sources of uncertainty when choosing a likelihood function, the noisy draws of N dots from the sampling distribution, affected by σ^2_{cloud}, and the internal noise applied to each dot, σ^2_{dot}. These two noise sources are additive, resulting in the combined precision per dot of

$$\tau_{comb} = \frac{1}{\sigma^2_{cloud} + \sigma^2_{dot}} = \frac{1}{\tau^{-1}_{cloud} + \tau^{-1}_{dot}}. \tag{1}$$

Note that we have parameterised the variances in terms of precision: $\tau_{cloud} = \sigma^{-2}_{cloud}$ and $\tau_{dot} = \sigma^{-2}_{dot}$. The likelihood function for the observed dot cloud is

$$\begin{aligned} p(\boldsymbol{X}|\mu_{cloud}, \tau_{cloud}, \tau_{dot}) &= \prod_{i=1}^{N} p(x_i|\mu_{cloud}, \tau_{comb}) \\ &\propto \tau_{comb}^{\frac{N}{2}} \exp\left(\frac{-\tau_{comb}}{2}\sum_{i=1}^{N}(x_i - \mu_{cloud})^2\right). \end{aligned} \tag{2}$$

This likelihood function is centred on the dot-cloud centroid, the average of all the dot locations, because each dot is considered to be an equally reliable cue to the location of the generating distribution. An example likelihood function is shown in Fig 6A.

The prior distribution

In the centred-prior variants, we assumed the observer correctly inferred the joint distribution of the sampling distribution mean and precision through experience with the task. Even though the mean and precision were discretised in our experiment (7 and 3 possible values, respectively), we assumed that the observer could not have such a detailed representation and thus we consider that the prior is a continuous distribution. The conjugate prior for our likelihood function (i.e., inferring a normal distribution with unknown mean and precision) is a normal-Gamma distribution [73, 74]:

$$\begin{aligned} p(\mu_{cloud}, \tau_{cloud}) &= \mathcal{NG}(\mu_{cloud}, \tau_{cloud}|\mu_0, \kappa_0, \alpha_0, \beta_0) \\ &= \mathcal{N}(\mu_{cloud}|\mu_0, (\kappa_0\tau_{cloud})^{-1})\ \mathcal{G}a(\tau_{cloud}|\alpha_0, \beta_0). \end{aligned} \tag{3}$$

To find the parameters of the normal-Gamma prior, we calculated the values of μ_0, κ_0, α_0, and β_0 such that the marginal means and variances matched the true stimulus statistics from the experiment (derivation in S1 Text). This prior distribution, shown in Fig 6A, has $\mu_0 = 0$, $\kappa_0 = 0.68$, $\alpha_0 = 3.84$, and $\beta_0 = 13.48$. Note that κ_0 can be interpreted as a number of pseudo-observations, α_0 as related to degrees of freedom, and β_0 as related to the prior belief about pooled variance.

For the flat-prior variants, we applied a uniform distribution over all possible mean and spread values. The posterior distribution is thus equivalent to the normalised likelihood function. For the posterior distribution we detail next, we are referring to the centred-prior variant models.

The posterior distribution

Using Bayes' theorem, we can write the expression for the posterior that combines the prior beliefs about the joint probability of the distribution's mean and precision with the observed dot evidence. Because we have used the conjugate prior, the posterior is also a normal-Gamma distribution (derivation provided in S1 Text):

$$\begin{aligned} p(\mu_{cloud}, \tau_{cloud}|\boldsymbol{X}, \tau_{dot}) &\propto p(\boldsymbol{X}|\mu_{cloud}, \tau_{cloud}, \tau_{dot})p(\mu_{cloud}, \tau_{cloud}) \\ &= \mathcal{NG}(\mu_{cloud}, \tau_{cloud}|\mu_p, \kappa_p, \alpha_p, \beta_p, \tau_{dot}), \end{aligned} \tag{4}$$

with the following posterior parameters:

$$\mu_p = \frac{\kappa_0 \tau_{cloud} \mu_0 + N\tau_{comb}\bar{x}}{\kappa_0 \tau_{cloud} + N\tau_{comb}}, \tag{5}$$

$$\kappa_p = \frac{\kappa_0 \tau_{cloud} + N\tau_{comb}}{\tau_{cloud}}, \tag{6}$$

$$\alpha_p = \alpha_0 + \frac{N}{2}, \tag{7}$$

and

$$\beta_p = \beta_0 + \frac{N}{2\tau_{cloud}}\left(\frac{\kappa_0 \tau_{cloud}\tau_{comb}(\bar{x} - \mu_0)^2}{\kappa_0 \tau_{cloud} + N\tau_{comb}} + \tau_{comb}s^2 + \log\left(\tau_{dot}^{-1}\tau_{cloud} + 1\right)\right). \tag{8}$$

Where $\bar{x}$ is the sample mean of the observed dots,

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{9}$$

and s^2 is the observed sample variance in maximum-likelihood terms,

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2. \tag{10}$$

An example posterior distribution is shown in Fig 6C.

Taxonomy of confidence models

The seven base confidence models we considered differ in the level of access to the posterior distribution. They could access either 1) the full distribution, 2) only the mode, or 3) both mode and spread. Models that made use of the full distribution used Probability-metrics for computing confidence. Confidence computations that used the mode, with or without posterior spread, were DFC Evidence-Strength metrics. The Heuristic model combined separate estimates of stimulus strength and sensory uncertainty factors (i.e., number and spread of dots) in an idiosyncratic weighted sum. Model summaries are provided in Table 2 and model equations in Table 3.

We also considered two model variants when relevant, resulting in a total of twelve unique models. For the *centred-prior* variant, the observer uses the informative prior that matched the stimulus statistics (depicted in Fig 6). For the *flat-prior* variant, the observer uses a non-informative flat prior as if they had used the normalised likelihood function to make their perceptual and confidence decisions.

In regards to confidence interval bias and noise, as per its definition, the Ideal-Confidence-Observer model did not have either of these elements [41], but otherwise all confidence models had both. Confidence interval bias in the confidence forced-choice method is a preference for reporting Interval 2, for example and should not be confused with an over- or under-confidence bias [26, 41]. Confidence noise was dependent on the individual confidence model. Almost all models had additive Gaussian noise. However, for models that used raw probability values (i.e., the Basic-Probability models), we used Beta-distributed variability to keep the confidence decision variable between 0 and 1.

Ideal-Confidence-Observer model

The goal of the observer was to infer if the mean of the dot sampling distribution was left or right of centre. The ideal confidence observer would take into account the unknown precision of the sampling distribution by marginalising over all possible precision values

$$p(\mu_{cloud}|\boldsymbol{X}) = \int p(\mu_{cloud}, \tau_{cloud}|\boldsymbol{X}, \tau_{dot})d\tau_{cloud}, \tag{11}$$

which results in the marginal posterior distribution shown in Fig 6B. The evidence for each category, given the observed measurements and centred stimulus prior, can be computed by

$$p(C = R|\boldsymbol{X}, k_1) = p(\mu_{cloud} > k_1|\boldsymbol{X}, k_1) = \int_{k_1}^{\infty} p(\mu_{cloud}|\boldsymbol{X})d\mu_{cloud}, \tag{12}$$

where k_1 is the Type 1 discrimination boundary, and

$$p(C = L|\boldsymbol{X}, k_1) = 1 - p(C = R|\boldsymbol{X}, k_1). \tag{13}$$

For the Type 1 judgement, the inferred category of the stimulus, $\hat{C}$, is then determined by the relative probabilities of each category, with the observer selecting the most likely category. They select "right" if $p(C = R|\boldsymbol{X}, k_1) > p(C = L|\boldsymbol{X}, k_1)$, and "left" otherwise. The observer reports the more likely category with a keypress ($r = \hat{C}$), unless a lapse occurs, where the observer selects the unintended category, with the lapse rate of λ. As such,

$$p(\text{choose Right}) = \lambda + (1 - 2\lambda)p(p(C = R|\boldsymbol{X}, k_1) > p(C = L|\boldsymbol{X}, k_1)). \tag{14}$$

For the Type 2 confidence judgement, the Ideal Confidence Observer considers the relative strength of the evidence in two consecutive Type 1 judgements, selecting the interval with the response that is more likely to be correct as having higher confidence, where Interval 1 evidence, w_1, is

$$w_1 = \text{Ev}(r_1|\boldsymbol{X}_1, k_1) = \max\left[p(C_1 = R|\boldsymbol{X}_{\boldsymbol{1}}, k_1), p(C_1 = L|\boldsymbol{X}_{\boldsymbol{1}}, k_1)\right] \tag{15}$$

and Interval 2 evidence, w_2 is

$$w_2 = \text{Ev}(r_2|\boldsymbol{X}_2, k_1) = \max\left[p(C_2 = R|\boldsymbol{X}_{\boldsymbol{2}}, k_1), p(C_2 = L|\boldsymbol{X}_{\boldsymbol{2}}, k_1)\right]. \tag{16}$$

The observer reports Interval 1 for the confidence judgement if $w_1 > w_2$ and Interval 2 if $w_2 > w_1$. When a lapse in the Type 1 report occurs, the observer selects the unintended response. We assume the participant is aware of the lapse and reflects this in their confidence report. Mathematically, the max operation in the above equations is replaced by a min operation for instances of lapses. The confidence evidence computation and decision rule are shown in Table 3, along with those for all Type 2 confidence models we tested. As this is the Ideal-Confidence-Observer model, the observer does not incur any additional metacognitive noise in

their computation of confidence, but they are still subject to Type 1 decision bias through parameter k_1 [41]. Following the decision rule of the ideal confidence observer,

$$p(\text{choose } I_2) = \lambda + (1 - 2\lambda)p(w_2 > w_1). \quad (17)$$

We assumed the Type 2 lapse rate was identical to the Type 1 lapse rate, λ, and fixed it in the Type 2 model fits according to the best-fitting value from the Type 1 model fits. For both the Type 1 and 2 judgements, the choice probabilities were estimated by simulation. The Ideal-Confidence-Observer model had no free parameters after fixing the Type-1 parameters.

Probability-metric models

The remaining six Probability-metric models follow a similar logic to the Ideal Confidence Observer at the Type 1 and 2 levels, but included various forms of metacognitive noise (Table 3). We also considered flat- and centred-prior variants. The posterior distribution in Eq 11 differs between the two variants, but otherwise the model computations are unchanged. In all of the Probability-metric models, we included a confidence interval bias term for interval preferences, which was implemented in the decision rule as $-\infty < k_2 < \infty$. Thus, the Probability-metric models had two free parameters: ν or σ_{conf} (i..e., confidence noise), and k_2.

Basic-Probability models. The observer directly compares the evidence in favour of their choice, but these raw probability values have been corrupted by noise. As probabilities are constrained to the interval [0, 1], we implemented a beta-noise model, with ν as a concentration parameter (larger values: less confidence noise). Effectively, ν is a number of "internal" observations, of which a certain fraction are consistent with the chosen category and the rest with the unchosen category, with proportions in line with the observer's beliefs about being correct. Consequently, as $\nu \to \infty$, w becomes a Delta distribution at $\text{Ev}(r|\boldsymbol{X}, k_1)$.

Probability-Difference models. The decision evidence is first compared for the intervals and then late additive Gaussian noise is applied to their difference. The smaller this confidence noise, σ_{conf}, the closer the observer is to the ideal confidence observer.

Log-Probability-Ratio (LPR) models. The decision evidence of each interval is transformed onto a continuous scale of log probability and then Gaussian noise is applied before the intervals are compared.

DFC Evidence-Strength metric models

The Distance-From-Criterion (DFC) Evidence-Strength models rely on point estimates from the decision process. In both the flat- and centred-prior variants, the observer computes the mode of the posterior, $\hat{\mu}$. Using this point estimate in the Type 1 decision will lead to identical Type 1 choice behaviour as in Eq 14, for the same prior variant. This is because a posterior mode right of the Type 1 criterion will always corresponds to $p(C = R|\boldsymbol{X}, k_1) > p(C = L|\boldsymbol{X}, k_1)$ for the normal-gamma posterior distribution in Eq 4, and vice versa for a posterior mode to the left. However, the DFC Evidence-Strength models make different predictions for confidence. Similar to the Probability-metric models, we included a confidence interval bias, $-\infty < k_2 < \infty$. Thus, the DFC Evidence-Strength models also had two free parameters: σ_{conf} and k_2.

Unscaled-Distance model. For confidence evidence, this model considers the distance of $\hat{\mu}$ from the Type 1 discrimination boundary, k_1, with added Gaussian noise, $N(0, \sigma^2_{conf})$, per interval. The further $\hat{\mu}$ is from the criterion, the stronger the evidence for the spatial discrimination judgement.

Scaled-Distance models. The observer also considers the decision uncertainty by assessing the spread of the marginal posterior distribution in the form of a second point estimate, $\hat{\sigma}$.

The distance of the mode from k_1 is then computed units of standard deviation (i.e., signal-to-noise ratio; see S1 Text for more details).

Heuristic model

The *Heuristic* model uses an estimate of stimulus strength and estimates of each separate factor affecting sensory uncertainty (i.e, the number of dots and dot spread) as inputs to the confidence computation, without any constraint on the relative weighting of these inputs on confidence. In effect, this model is the best-fitting model from the preliminary logistic regression analysis (see Fig B in S1 Text), but fit in accordance with the discrimination behaviour (Table 3). This involved combining the noisy observations of the sampled dots to compute the position predictor (i.e., the centroid DFC $|\mu_c - k_1|$), the dot spread for the quality predictor (i.e., the empirical spread, σ_{emp}), and a count of the number of dots for the quantity predictor (N), as well as any confidence interval bias (k_2). To avoid equivalent best-fitting regressors, the position coefficient that was expected to be dominant was fixed at 1 in the model and the other two coefficients, β_1 (for the difference in the number of dots across intervals) and β_2 (for the difference in inverse dot spread), were free to vary. Confidence noise, σ_{conf}, was also free to vary. In total, the Heuristic model had 4 free parameters: β_1, β_2, σ_{conf}, and k_2.

Model fitting, comparison, and validation

Models were fit in a two-step procedure using custom-written Matlab code (available at https://osf.io/k2nhq/). Type 1 parameters were fit first (σ_{dot}, k_1, and λ) using the discrimination responses. Then the Type 2 parameters (v or σ_{conf}, k_2, and possibly the two β values) were estimated using the confidence forced-choice responses. Type 1 parameters were kept fixed at their best-fitting values for the Type 2 fits. Type 1 models were fit using a brute-force grid method, with response probabilities estimated by simulation of the observer. For the Type 2 model fits, only simulated sensory measurements consistent with the observer's Type 1 responses were used for fitting to ensure the calculated response probabilities were conditional on the discrimination choice. We used Bayesian Adaptive Direct Search (BADS) with the BADS toolbox [75], for the Type 2 model fits. Models were then compared in terms of their corrected Akaike information criterion (AICc) scores [76, 77]. The best-fitting model for an observer was the model that had the lowest AICc score. The ordering of model fit across observers was determined by the mean AICc score per model (i.e., the best-fitting model had the lowest average AICc score). Relative AICc scores are reported in text for the models best-fit by at least one observer and used for plotting purposes. Further details on the model fitting are provided in S1 Text. To confirm that the models are distinguishable for this experimental design, we performed a model recovery analysis. Using the MLE of the parameters, 10 data sets were simulated per observer per model. A successful model recovery indicates there is sufficient data to treat each participant as the replication unit (for more details on small-N designs, see [78]) and that different best-fitting models is less likely a result of measurement noise than model heterogeneity in the population [58, 59]. As such, the diversity in best-fitting model per participant is a meaningful result in the present study, reflecting idiosyncrasies in the confidence computation, and should be considered in conjunction with the group averages.

Supporting information

S1 Text. Supplementary information. 1) Model fitting: Derivation of the posterior for the Bayesian ideal observer; selecting the prior distribution parameters; computing the standard deviation of the marginal posterior distribution for the Scaled-Distance models; model-fitting

procedure; and parameter recovery. 2) Additional results: preliminary logistic analysis to confirm that the quantity and quality manipulations affected confidence; qualitative comparison of the models; results of model variants; examining the Heuristic-model coefficients from the simulated datasets; and confidence agreement behaviour and model predictions.
(PDF)

Acknowledgments

We would like to thank Wei Ji Ma for helpful early discussions on data analysis and Damaris Beutel for her help in collecting the research data.

Author Contributions

Conceptualization: Shannon M. Locke, Michael S. Landy, Pascal Mamassian.

Data curation: Shannon M. Locke.

Formal analysis: Shannon M. Locke.

Funding acquisition: Shannon M. Locke, Michael S. Landy, Pascal Mamassian.

Investigation: Shannon M. Locke.

Methodology: Shannon M. Locke, Michael S. Landy, Pascal Mamassian.

Project administration: Shannon M. Locke, Michael S. Landy, Pascal Mamassian.

Resources: Michael S. Landy, Pascal Mamassian.

Software: Shannon M. Locke.

Supervision: Michael S. Landy, Pascal Mamassian.

Validation: Shannon M. Locke.

Visualization: Shannon M. Locke, Michael S. Landy, Pascal Mamassian.

Writing – original draft: Shannon M. Locke.

Writing – review & editing: Shannon M. Locke, Michael S. Landy, Pascal Mamassian.

References

1. Clarke FR, Birdsall TG, Tanner WP. Two types of ROC curves and definitions of parameters. The Journal of the Acoustical Society of America. 1959; 31(5):629–630. https://doi.org/10.1121/1.1907764
2. Van den Berg R, Zylberberg A, Kiani R, Shadlen MN, Wolpert DM. Confidence is the bridge between multi-stage decisions. Current Biology. 2016; 26:3157–3168. https://doi.org/10.1016/j.cub.2016.10.021 PMID: 27866891
3. Meyniel F, Schlunegger D, Dehaene S. The sense of confidence during probabilistic learning: A normative account. PLoS Comput Biol. 2015; 11(6):e1004305. https://doi.org/10.1371/journal.pcbi.1004305 PMID: 26076466
4. Frömer R, Nassar MR, Bruckner R, Stürmer B, Sommer W, Yeung N. Response-based outcome predictions and confidence regulate feedback processing and learning. Elife. 2021; 10:e62825. https://doi.org/10.7554/eLife.62825 PMID: 33929323
5. Kiani R, Corthell L, Shadlen MN. Choice certainty is informed by both evidence and decision time. Neuron. 2014; 84(6):1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015 PMID: 25521381
6. Maniscalco B, Lau HC. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Consciousness and Cognition. 2012; 21:422–430. https://doi.org/10.1016/j.concog.2011.09.021 PMID: 22071269

7. Fleming SM, Daw ND. Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. Psychological Review. 2017; 124(1):91–114. https://doi.org/10.1037/rev0000045 PMID: 28004960
8. Pleskac TJ, Busemeyer JR. Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. Psychological Review. 2010; 117(3):864–901. https://doi.org/10.1037/a0019737 PMID: 20658856
9. Pouget A, Drugowitsch J, Kepecs A. Confidence and certainty: Distinct probabilistic quantities for different goals. Nature Neuroscience. 2016; 19(3):366–374. https://doi.org/10.1038/nn.4240 PMID: 26906503
10. Sanders JI, Hangya B, Kepecs A. Signatures of a statistical computation in the human sense of confidence. Neuron. 2016; 90(3):499–506. https://doi.org/10.1016/j.neuron.2016.03.025 PMID: 27151640
11. Adler WT, Ma WJ. Comparing Bayesian and non-Bayesian accounts of human confidence reports. PLoS Computational Biology. 2018; 14(11):e1006572. https://doi.org/10.1371/journal.pcbi.1006572 PMID: 30422974
12. Li HH, Ma WJ. Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. Nature Communications. 2020; 11(1):1–11. https://doi.org/10.1038/s41467-020-15581-6 PMID: 32332712
13. Galvin SJ, Podd JV, Drga V, Whitmore J. Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. Psychonomic Bulletin & Review. 2003; 10(4):843–876. https://doi.org/10.3758/BF03196546 PMID: 15000533
14. De Gardelle V, Mamassian P. Weighting mean and variability during confidence judgments. PLoS One. 2015; 10(3):e0120870. https://doi.org/10.1371/journal.pone.0120870 PMID: 25793275
15. Patel D, Fleming SM, Kilner JM. Inferring subjective states through the observation of actions. Proceedings of the Royal Society B: Biological Sciences. 2012; 279(1748):4853–4860. https://doi.org/10.1098/rspb.2012.1847 PMID: 23034708
16. Vickers D. Decision processes in visual perception. New York, NY: Academic Press; 1979.
17. Mole CD, Jersakova R, Kountouriotis GK, Moulin CJA, Wilkie RM. Metacognitive judgements of perceptual-motor steering performance. Quarterly Journal of Experimental Psychology. 2018; 71(10):2223–2234. https://doi.org/10.1177/1747021817737496 PMID: 30226435
18. Zylberberg A, Barttfeld P, Sigman M. The construction of confidence in a perceptual decision. Frontiers in Integrative Neuroscience. 2012; 6:79. https://doi.org/10.3389/fnint.2012.00079 PMID: 23049504
19. Maniscalco B, Peters MAK, Lau H. Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. Attention, Perception, and Psychophysics. 2016; 78(3):923–937. https://doi.org/10.3758/s13414-016-1059-x PMID: 26791233
20. Fleming SM, Lau HC. How to measure metacognition. Frontiers in Human Neuroscience. 2014; 8:00443. https://doi.org/10.3389/fnhum.2014.00443 PMID: 25076880
21. Locke SM, Gaffin-Cahn E, Hosseinizaveh N, Mamassian P, Landy MS. Priors and payoffs in confidence judgments. Attention, Perception, & Psychophysics. 2020; 82(6):3158–3175. https://doi.org/10.3758/s13414-020-02018-x PMID: 32383111
22. Mamassian P. Visual confidence. Annual Review of Vision Science. 2016; 2(1):459–481. https://doi.org/10.1146/annurev-vision-111815-114630 PMID: 28532359
23. Barrett AB, Dienes Z, Seth AK. Measures of metacognition on signal-detection theoretic models. Psychological Methods. 2013; 18(4):535–552. https://doi.org/10.1037/a0033268 PMID: 24079931
24. Maniscalco B, Lau H. The signal processing architecture underlying subjective reports of sensory awareness. Neuroscience of Consciousness. 2016; 2016(1):1–17. https://doi.org/10.1093/nc/niw002 PMID: 27499929
25. Bang JW, Shekhar M, Rahnev D. Sensory noise increases meta-cognitive efficiency. Journal of Experimental Psychology: General. 2019; 148(3):437–452. https://doi.org/10.1037/xge0000511
26. Mamassian P, de Gardelle V. Modeling perceptual confidence and the confidence forced-choice paradigm. Psychological Review. 2021. Advanced online publication available from: https://doi.org/10.1037/rev0000312 PMID: 34323580
27. Zylberberg A, Roelfsema PR, Sigman M. Variance misperception explains illusions of confidence in simple perceptual decisions. Consciousness and Cognition. 2014; 27:246–253. https://doi.org/10.1016/j.concog.2014.05.012 PMID: 24951943
28. Denison RN, Adler WT, Carrasco M, Ma WJ. Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. Proceedings of the National Academy of Sciences. 2018; 115(43):11090–11095. https://doi.org/10.1073/pnas.1717720115 PMID: 30297430

29. Gold JI, Shadlen MN. The neural basis of decision making. Annual Review of Neuroscience. 2007; 30:535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038 PMID: 17600525

30. Zylberberg A, Fetsch CR, Shadlen MN. The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. Elife. 2016; 5:e17688. https://doi.org/10.7554/eLife.17688 PMID: 27787198

31. Balsdon T, Wyart V, Mamassian P. Confidence controls perceptual evidence accumulation. Nature Communications. 2020; 11(1):1–11. https://doi.org/10.1038/s41467-020-15561-w PMID: 32273500

32. Aitchison L, Bang D, Bahrami B, Latham PE. Doubly Bayesian analysis of confidence in perceptual decision-making. PLoS Computational Biology. 2015; 11(10):e1004519. https://doi.org/10.1371/journal.pcbi.1004519 PMID: 26517475

33. Spence ML, Dux PE, Arnold DH. Computations underlying confidence in visual perception. Journal of Experimental Psychology: Human Perception and Performance. 2015; 42(5):671–682. PMID: 26594876

34. Boldt A, de Gardelle V, Yeung N. The impact of evidence reliability on sensitivity and bias in decision confidence. Journal of Experimental Psychology: Human Perception and Performance. 2017; 43 (8):1520–1531. https://doi.org/10.1037/xhp0000404 PMID: 28383959

35. Bertana A, Chetverikov A, van Bergen RS, Ling S, Jehee JF. Dual strategies in human confidence judgments. Journal of Vision. 2021; 21(5):21–21. https://doi.org/10.1167/jov.21.5.21 PMID: 34010953

36. Barthelmé S, Mamassian P. Flexible mechanisms underlie the evaluation of visual confidence. Proceedings of the National Academy of Sciences. 2010; 107(48):20834–20839. https://doi.org/10.1073/pnas.1007704107 PMID: 21076036

37. Locke SM, Mamassian P, Landy MS. Performance monitoring for sensorimotor confidence: A visuomotor tracking study. Cognition. 2020; 205:104396. https://doi.org/10.1016/j.cognition.2020.104396 PMID: 32771212

38. Peirce CS, Jastrow J. On small differences in sensation. Memoirs of the National Academy of Science. 1884; 3:73–83.

39. Lisi M, Mongillo G, Milne G, Dekker T, Gorea A. Discrete confidence levels revealed by sequential decisions. Nature Human Behaviour. 2021; 5(2):273–280. https://doi.org/10.1038/s41562-020-00953-1 PMID: 32958899

40. Rahnev D, Balsdon T, Charles L, de Gardelle V, Denison RN, Desender K, et al. Consensus goals for the field of visual metacognition. PsyArXiv [Preprint]. 2021 [cited 2022 June 23]. Available from: https://psyarxiv.com/z8v5x.

41. Mamassian P. Confidence forced-choice and other metaperceptual tasks. Perception. 2020; 49 (6):616–635. https://doi.org/10.1177/0301006620928010 PMID: 32552488

42. Graziano M, Sigman M. The spatial and temporal construction of confidence in the visual scene. PLoS One. 2009; 4(3):e4909. https://doi.org/10.1371/journal.pone.0004909 PMID: 19290055

43. Ais J, Zylberberg A, Barttfeld P, Sigman M. Individual consistency in the accuracy and distribution of confidence judgments. Cognition. 2016; 146:377–386. https://doi.org/10.1016/j.cognition.2015.10.006 PMID: 26513356

44. Navajas J, Hindocha C, Foda H, Keramati M, Latham PE, Bahrami B. The idiosyncratic nature of confidence. Nature Human Behaviour. 2017; 1(11):810–818. https://doi.org/10.1038/s41562-017-0215-1 PMID: 29152591

45. Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G. Relating introspective accuracy to individual differences in brain structure. Science. 2010; 329(5998):1541–1543. https://doi.org/10.1126/science.1191883 PMID: 20847276

46. Rahnev D. A robust confidence–accuracy dissociation via criterion attraction. Neuroscience of Consciousness. 2021; 2021(1):niab039. https://doi.org/10.1093/nc/niab039 PMID: 34804591

47. Rahnev D, Denison RN. Suboptimality in perceptual decision making. Behavioral and Brain Sciences. 2018; 41:e223. https://doi.org/10.1017/S0140525X18000936 PMID: 29485020

48. Gorea A, Sagi D. Failure to handle more than one internal representation in visual detection tasks. Proceedings of the National Academy of Sciences. 2000; 97(22):12380–12384. https://doi.org/10.1073/pnas.97.22.12380 PMID: 11050253

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

50. Macmillan NA, Creelman CD. Detection theory: A user's guide. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.; 2005.

51. Landy MS, Banks MS, Knill DC. Ideal-observer models of cue integration. In: Sensory cue integration. Oxford University Press; 2011. p. 5–29.
52. Kahneman D, Tversky A. Prospect theory: An analysis of decision under risk. Econometrica. 1979; 4:263–291. https://doi.org/10.2307/1914185
53. Fox CR, Poldrack RA. Prospect theory and the brain. In: Glimcher PW, Camerer CF, Fehr E, Poldrack RA, editors. Neuroeconomics: decision making and the brain. Elsevier, New York, NY; 2009. p. 145–173.
54. Zhang H, Daw ND, Maloney LT. Human representation of visuo-motor uncertainty as mixtures of orthogonal basis distributions. Nature Neuroscience. 2015; 18(8):1152–1158. https://doi.org/10.1038/nn.4055 PMID: 26120962
55. Zhang H, Maloney LT. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. Frontiers in Neuroscience. 2012; 6:1. https://doi.org/10.3389/fnins.2012.00001 PMID: 22294978
56. Ma WJ. Organizing probabilistic models of perception. Trends in Cognitive Sciences. 2012; 16 (10):511–518. https://doi.org/10.1016/j.tics.2012.08.010 PMID: 22981359
57. Gardner JL. Optimality and heuristics in perceptual neuroscience. Nature Neuroscience. 2019; 22 (4):514–523. https://doi.org/10.1038/s41593-019-0340-4 PMID: 30804531
58. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. Neuroimage. 2009; 46(4):1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025 PMID: 19306932
59. Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group studies—Revisited. NeuroImage. 2014; 84:971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065 PMID: 24018303
60. Gigerenzer G, Gaissmaier W. Heuristic decision making. Annual Review of Psychology. 2011; 62:451–482. https://doi.org/10.1146/annurev-psych-120709-145346 PMID: 21126183
61. Shekhar M, Rahnev D. The nature of metacognitive inefficiency in perceptual decision making. Psychological Review. 2021; 128(1):45–70. https://doi.org/10.1037/rev0000249 PMID: 32673034
62. Shekhar M, Rahnev D. How do humans give confidence? A comprehensive comparison of process models of metacognition. PsyArXiv [Preprint]. 2022 [cited 2022 June 23]. Available from: https://psyarxiv.com/cwrnt.
63. Rahnev D, Koizumi A, McCurdy LY, D'Esposito M, Lau H. Confidence leak in perceptual decision making. Psychological science. 2015; 26(11):1664–1680. https://doi.org/10.1177/0956797615595037 PMID: 26408037
64. Burgess AE, Colborne B. Visual signal detection. IV. Observer inconsistency. Journal of the Optical Society of America A. 1988; 5(4):617–627. https://doi.org/10.1364/JOSAA.5.000617 PMID: 3404312
65. Li RW, Klein SA, Levi DM. The receptive field and internal noise for position acuity change with feature separation. Journal of Vision. 2006; 6(4):2–2. https://doi.org/10.1167/6.4.2 PMID: 16889471
66. Hasan BAS, Joosten E, Neri P. Estimation of internal noise using double passes: Does it matter how the second pass is delivered? Vision Research. 2012; 69:1–9. https://doi.org/10.1016/j.visres.2012.06.014
67. Charles L, Chardin C, Haggard P. Evidence for metacognitive bias in perception of voluntary action. Cognition. 2020; 194:104041. https://doi.org/10.1016/j.cognition.2019.104041 PMID: 31470186
68. Fischer J, Whitney D. Serial dependence in visual perception. Nature Neuroscience. 2014; 17(5):738–743. https://doi.org/10.1038/nn.3689 PMID: 24686785
69. Gekas N, McDermott KC, Mamassian P. Disambiguating serial effects of multiple timescales. Journal of Vision. 2019; 19(6):24. https://doi.org/10.1167/19.6.24 PMID: 31251808
70. Brainard DH. The Psychophysics Toolbox. Spatial Vision. 1997; 10(4):433–436. https://doi.org/10.1163/156856897X00357 PMID: 9176952
71. Pelli DG. The Video Toolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision. 1997; 10:437–442. https://doi.org/10.1163/156856897X00366 PMID: 9176953
72. Kleiner M, Brainard D, Pelli D, Ingling A, Murray R, Broussard C. What's new in psychtoolbox-3. Perception. 2007; 36(14):1–16.
73. Bishop CM. Pattern recognition and machine learning. Springer. New York, NY; 2006.
74. Murphy KP. Conjugate Bayesian analysis of the Gaussian distribution. KP Murphy's faculty webpage, The University of British Columbia [Technical report]. 2007 [cited 2022 June 23]. Available from https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf.
75. Acerbi L, Ma WJ. Practical Bayesian optimization for model fitting with Bayesian Adaptive Direct Search. Advances in Neural Information Processing Systems. 2017; 30:1834–1844.

76. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19(6):716–723. https://doi.org/10.1109/TAC.1974.1100705

77. Cavanaugh JE. Unifying the derivations for the Akaike and corrected Akaike information criteria. Statistics & Probability Letters. 1997; 33(2):201–208. https://doi.org/10.1016/S0167-7152(96)00128-9

78. Smith PL, Little DR. Small is beautiful: In defense of the small-N design. Psychonomic Bulletin & Review. 2018; 25(6):2083–2101. https://doi.org/10.3758/s13423-018-1451-8