PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Challenging the Bayesian confidence hypothesis in perceptual decision-making

Kai Xue[a,1], Medha Shekhar[a], and Dobromir Rahnev[a]

Edited by Richard Shiffrin, Indiana University Bloomington, Bloomington, IN; received May 29, 2024; accepted October 16, 2024

The Bayesian confidence hypothesis (BCH), which postulates that confidence reflects the posterior probability that a decision is correct, is currently the most prominent theory of confidence. Although several recent studies have found evidence against it in the context of relatively complex tasks, BCH remains dominant for simpler tasks. The major alternative to BCH is the confidence in raw evidence space (CRES) hypothesis, according to which confidence is based directly on the raw sensory evidence without explicit probability computations. Here, we tested these competing hypotheses in the context of perceptual tasks that are assumed to induce Gaussian evidence distributions. We show that providing information about task difficulty gives rise to a basic behavioral signature that distinguishes BCH from CRES models even for simple 2-choice tasks. We examined this signature in three experiments and found that all experiments exhibited behavioral signatures in line with CRES computations but contrary to BCH ones. We further performed an extensive comparison of 16 models that implemented either BCH or CRES confidence computations and systematically differed in their auxiliary assumptions. These model comparisons provided overwhelming support for the CRES models over their BCH counterparts across all model variants and across all three experiments. These observations challenge BCH and instead suggest that humans may make confidence judgments by placing criteria directly in the space of the sensory evidence.

visual metacognition | perceptual decision-making | confidence computation | signal detection theory

Significance

Humans can judge decision accuracy using confidence ratings, but the underlying computations remain debated. One of the most prominent theories of confidence, the Bayesian confidence hypothesis (BCH), posits that confidence reflects the probability of being correct. We tested BCH in three perceptual tasks that are assumed to induce Gaussian evidence distributions. Specifically, we examined whether humans correctly use knowledge of task structure when rating confidence. We found that, when informed about task difficulty, humans barely shift their confidence criteria. Extensive model comparisons revealed that the data were fit best by models which assume that confidence in perceptual tasks is given in sensory evidence space. Our findings thus challenge one of the most prominent theories of confidence in perceptual decision-making.

Humans have the metacognitive ability to evaluate the quality of their own decisions using confidence ratings (1). However, the computations behind confidence continue to be a matter of debate (2–4). Perhaps the most prominent theory of how people rate confidence is the Bayesian Confidence Hypothesis (BCH), according to which confidence reflects the probability of being correct (5). More formally, BCH is defined as a model according to which people give confidence judgments that reflect the probability of being correct by accurately combining the correct statistical structure of the task and the present sensory evidence (4, 6–11). We note that this definition includes three underlying assumptions: 1) people are aware of the statistical structure of the task, 2) people estimate the probability of being correct, and 3) people perform this computation correctly. This formal definition is usually simplified to just "confidence reflects the probability of being correct," which implicitly includes the assumptions of knowing the structure of the task and performing correct computations.

The popularity of BCH is due to at least three factors. First, BCH is intuitive. Researchers typically implicitly or explicitly instruct subjects to use confidence ratings as indices of the probability that their decisions are correct, so it only makes sense to assume that subjects are reporting exactly what the experimenters are asking them to. Second, BCH is general. For every decision that humans make, it is possible, at least in principle, to estimate the probability that that decision is correct and report confidence accordingly. This fact means that BCH is applicable to virtually all tasks. Third, the hypothesis has been promoted by several high-profile opinion articles (5, 10, 11) that go as far as to define confidence in terms of BCH. Together, these considerations have propelled BCH into the status of, in the words of a recent article, the "leading theory of confidence" (6).

Given the prominence of BCH, it is perhaps surprising that empirical support for the hypothesis is rather limited. Most prominently, Sanders et al. (12) formulated and found empirical support for several signatures of "statistical" confidence (defined equivalently to BCH). However, it should be noted that none of the signatures constitute sufficient conditions for BCH (8, 13) and some signatures have since been shown not to always hold in the empirical data (14, 15). Several other papers examined qualitative predictions of BCH and typically found support for these predictions only for some subjects (14) or conditions (16). Only a single paper found support for BCH, which came in the context of a simple 2-choice perceptual decision-making task (7). The authors fit three different models to two datasets and found that the BCH model provided the best fit in one of the

Author affiliations: [a]School of Psychology, Georgia Institute of Technology, Atlanta, GA 30332

Author contributions: K.X. and D.R. designed research; K.X. and D.R. performed research; K.X. contributed new reagents/analytic tools; K.X. and M.S. analyzed data; M.S. and D.R. reviewed the paper; and K.X. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: kxue33@gatech.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2410487121/-/DCSupplemental.

Published November 22, 2024.

experiments and tied with another model in the second experiment. Contrary to these results with a simple perceptual task, several papers have rejected BCH in the context of more complex manipulations or tasks (4, 6, 8).

At this point, it is important to make a distinction between BCH and the broader Bayesian decision-making framework. The Bayesian framework is a conceptual and philosophical approach, according to which decision-making is conceptualized as probabilistic reasoning that normatively integrates prior knowledge with new evidence. As has been argued in the philosophy of science literature (17), the core of a scientific framework is often not directly falsifiable and instead should be judged based on its usefulness for generating novel insights. A similar argument has been explicitly made in the case of the neuroconnectionist framework (18). In this view, the Bayesian framework is a way of thinking that can be used to generate specific explanations and hypotheses, but the core of the framework is not subject to falsification. In contrast to the broader Bayesian framework, BCH is a specific, testable, and falsifiable model that has been extensively investigated in numerous empirical studies (4, 6–8, 11, 12, 16, 19). This is because, unlike the broader Bayesian framework, BCH assumes both accurate knowledge of the statistical structure of the task and correct computations. In fact, as noted above, BCH could be formulated outside of the Bayesian framework as a "statistical confidence" model (12). As such, falsifying BCH would not falsify the broader Bayesian framework, and supporting BCH does not support the broader Bayesian framework (though either result may have implications about the usefulness of the framework in the context of confidence judgments). Critically, here we focus narrowly on BCH and examine the implications for the broader Bayesian framework in the *Discussion*.

The brief review above demonstrates that, the status of BCH remains uncertain despite the prominence of this hypothesis. It is thus of no surprise that comparing BCH to alternatives was voted as one of the central medium-term goals in the field of visual metacognition (20). Critically, given that BCH computations become computationally expensive—and sometimes intractable—for complex tasks and manipulations (21), BCH should first be tested in the context of the simplest possible 2-choice tasks.

The main alternative to the idea that confidence reflects the probability of being correct is that confidence instead reflects the strength of the raw sensory evidence (22–26). We refer to this alternative as the confidence in raw evidence space (CRES) model. The idea behind CRES is that humans give confidence by placing criteria directly on the sensory output from perceptual areas of the brain. To illustrate the distinction between CRES and BCH, consider a left/right motion direction discrimination and imagine that this task is performed by comparing the outputs of two sets of neurons selective to left and right motion direction. CRES postulates that the confidence criteria are placed directly on the difference in activation between these two sets of neurons. In contrast, BCH postulates that the raw evidence (i.e., the difference in activation between the two sets of neurons) is used to first compute the likelihood of each choice and then the probability that a specific choice is correct. In practice, it is difficult to determine what "raw evidence" is and where to find that in the brain. Nevertheless, in the context of simple 2-choice perceptual tasks used here, we follow the long tradition of modeling sensory evidence as Gaussian distributions on an abstract internal evidence axis (27). Unlike CRES, which does not require that the decision-maker has knowledge of the statistical structure of the task, BCH necessitates that the decision-maker has an accurate representation of the distributions of activations in these two sets of neurons given each stimulus category and uses them correctly to compute likelihoods. Thus, CRES and BCH make vastly different assumptions about the computations performed by the decision-maker when making a confidence judgment. Nevertheless, while these two hypotheses propose very different computations, it has remained challenging to distinguish between them empirically.

Importantly, a previously unappreciated fact about 2-choice tasks is that they can provide a decisive qualitative signature that can be used to test the validity of BCH. If confidence indeed reflects the probability of being correct and subjects know the task difficulty on each trial, then confidence criteria in conditions of varying difficulty should have constant log odds but vary in evidence space (Fig. 1 *A* and *B*). Alternatively, if confidence is given by setting criteria directly in the sensory evidence space (22–26), then confidence criteria in conditions of varying difficulty should be fixed in evidence space but have different log-likelihood ratio (i.e., log odds). The reason for the dissociation between the criterion patterns in evidence space vs. log-likelihood ratio space is that the same criterion location in evidence space maps onto different log odds depending on the distance between the distributions of internal evidence (Fig. 1*A*). Because easier tasks are associated with larger separations of the internal evidence distributions, simple manipulations of difficulty allow us to examine this straightforward qualitative signature of criterion location to distinguish between the BCH and CRES hypotheses.

Here, we examine the qualitative signature of criterion location in three experiments. Experiment 1 is a previously published dataset with simple 2-choice tasks that feature several uncued, interleaved difficulty levels (26). Experiment 2 is a new experiment where different difficulty levels are explicitly cued but remain interleaved across trials. Finally, Experiment 3 is also a new experiment where stimulus difficulty remained constant throughout each run of 120 trials, providing subjects with an extended period to adjust their confidence criteria. All three experiments involved perceptual tasks where sensory evidence is assumed to follow Gaussian distributions. We find that the criteria in log-likelihood ratio space (*LO*) change substantially across conditions for all three experiments, whereas the criteria in evidence space (*c*) stay comparatively more constant, thus contradicting the prediction of the BCH model. In addition to these qualitative results, we fit both BCH and CRES models with varying auxiliary model assumptions, resulting in 16 different model fits for each experiment. In line with the qualitative pattern, we find decisive support for the CRES over the BCH models. These results challenge BCH even in the context of the simplest 2-choice tasks where the BCH computations are most straightforward. Our results suggest that despite its many advantages, BCH does not describe human confidence ratings, which appear to be the product of setting confidence criteria directly in the space of raw sensory evidence.

Results

We compared CRES and BCH computations by examining qualitative patterns in the data and fitting 16 models with varying auxiliary assumptions to three experiments.

Experiment 1. We reanalyzed the data from Shekhar and Rahnev (26). Twenty subjects completed 2,800 trials each from a simple left/right orientation discrimination task and provided confidence ratings (Fig. 2*A*). The experiment included stimuli from three different contrast levels (easy, medium, and hard). The confidence ratings were originally given on a continuous scale but were rescaled to a 6-point confidence scale before analyses as in the original publication. This process resulted in five confidence criteria for a "left" choice and five confidence criteria for a "right" choice. Note that this experiment did not explicitly inform subjects about the

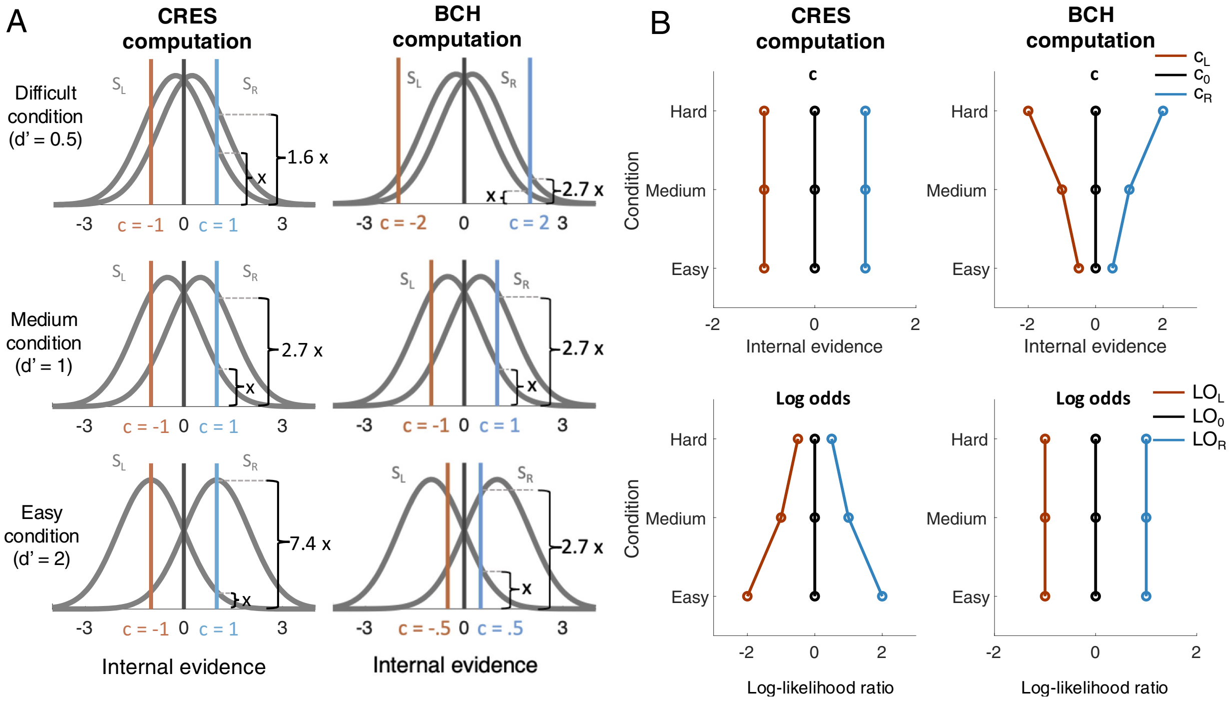

Fig. 1. A graphical depiction of the predictions of CRES and BCH computations. (*A*) Confidence criterion placement for CRES and BCH computations in a 2-choice task that involves discriminating between left and right stimuli (S_L and S_R). Each of the two stimulus categories produces a Gaussian distribution of internal evidence (plotted in gray) with easier conditions characterized by greater separation between the Gaussian distributions. The CRES hypothesis predicts that the confidence criterion, *c*, defined in internal evidence space stays constant across conditions (e.g., *c* = 1 in the figure), which leads to large differences in log odds, *LO*, defined in log-likelihood ratio space (odds = 1.6, 2.7, and 7.4 in the figure). On the other hand, BCH predicts that *LO* should stay constant (e.g., odds = 2.7 in the figure), which leads to large differences in *c* (e.g., *c* = 2, 1, and 0.5 in the figure). Note that the prediction for BCH holds only if subjects know the difficulty level on each trial. The orange and blue vertical lines indicate the confidence criteria for left and right decisions, respectively. (*B*) The change of the location of criterion in internal evidence space (measured by *c*) and in log-likelihood ratio space (measured by *LO*) across contrasts for the CRES and BCH models from the example in (*A*). Note that these model predictions would slightly change in the presence of metacognitive noise or lapse rate. Note that the odds (1.6, 2.7, and 7.4) from the example in (*A*) are converted to log odds (0.5, 1, and 2) here.

difficulty of each trial, but subjects could still guess the difficulty as they could observe the changing contrast of the stimulus (24).

We first tested the qualitative predictions made by the BCH and CRES hypotheses. Namely, BCH predicts that confidence criteria should have stable log-likelihood ratio values but vary in evidence space, whereas CRES predicts that confidence criteria should be stable in evidence space but vary in log-likelihood ratio space (Fig. 1*B*). We quantified the confidence criterion location in evidence and log-likelihood ratio space as the signal detection measure *c* and log odds (*LO*), respectively. We then computed the absolute values of the change in each quantity from the hard to the easy condition using the formulas $c_{change} = \left|\frac{c_{hard}-c_{easy}}{c_{hard}+c_{easy}}\right|$ and $LO_{change} = \left|\frac{LO_{hard}-LO_{easy}}{LO_{hard}+LO_{easy}}\right|$. We found that while the criteria in evidence space (measured by *c*) had an average change score of 0.09 across all 10 confidence criteria ($c_{L5}, \ldots, c_{L1}, c_{R1}, \ldots, c_{R5}$), the criteria in log-likelihood ratio space (measured by *LO*) had an average change score of 0.44. Critically, the change in *LO* was significantly larger than the change in *c* for all 10 criteria (all *P*'s $< 7.9 \times 10^{-5}$). These results match better the signature of the CRES computation (which predicts that *c* should stay constant across conditions, whereas *LO* should change; Fig. 1*B*) but diverge strongly from the signature of BCH computation (which predicts that *LO* should stay constant across conditions, whereas *c* should change; Fig. 1*B*).

The qualitative patterns of results in Fig. 2 strongly suggest that confidence in Experiment 1 follows CRES rather than BCH computations. However, our results showed that confidence criteria defined in internal evidence space (i.e., measure *c*) move slightly toward zero from the hard to the easy condition, which may be interpreted as reflecting a partial BCH computation. Nevertheless, another possibility is that this pattern emerges from CRES computations in the presence of other factors such as lapses or metacognitive noise. Therefore, to formally compare the two types of computation, we formulated and fit several CRES and BCH models to the empirical data.

To ensure our conclusions do not depend on auxiliary assumptions made in the fitting process, we formulated eight CRES and eight BCH models by systematically varying three sets of assumptions related to 1) the symmetry in confidence criteria, 2) the presence of lapse rate, and 3) the presence of metacognitive noise. First, half of the models included the assumption of symmetric confidence criteria, whereas the other half were more flexible and fit the confidence criteria for "left" and "right" choices separately. Second, half of the models included lapses (where a proportion of trial subjects give a random choice and a random confidence rating), whereas the other half did not. Third, half of the models included Gaussian metacognitive noise, whereas the other half did not. We fit each model to the empirical data using maximum likelihood estimation and compared the model fits based on their Akaike Information Criterion (AIC) values. Note that corresponding CRES and BCH models always have the same number of free parameters, and therefore the exact measure used for model comparison would not affect the comparison between the corresponding models of each type (e.g., using the Bayesian Information Criterion would lead to the same relative fits between corresponding CRES and BCH models).

We found that every one of the eight CRES models fit significantly better than their corresponding BCH models (Fig. 3*A*).

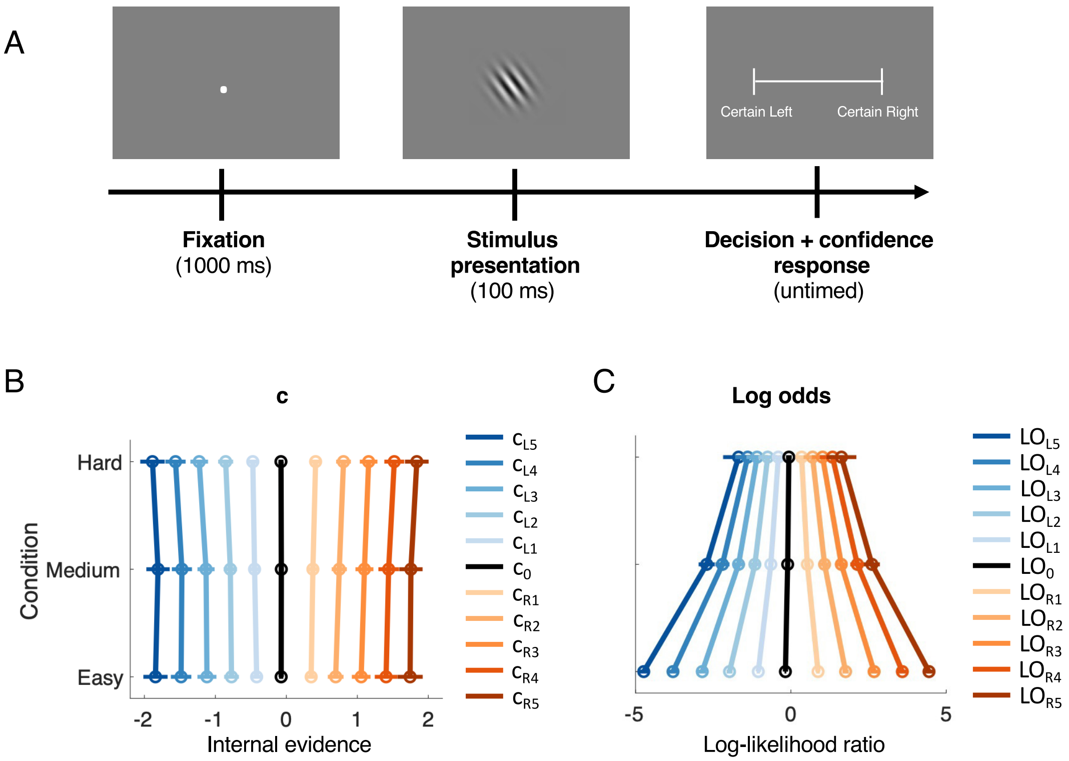

Fig. 2. Experiment design and behavioral results for Experiment 1. (*A*) Task. Each trial included the presentation of a noisy Gabor patch tilted 45° either to the left or right of the vertical. Subjects indicated the tilt of the Gabor patch while simultaneously rating their confidence on a continuous scale from 50 to 100 (transformed into a six-point scale for analysis). The Gabor patches had three different contrast levels resulting in easy, medium, and difficult conditions. (*B* and *C*) Empirical criterion values in internal evidence space (*c*) and log-likelihood ratio space (*LO*). Each colored line represents a confidence criterion, and the central black line represents the decision criterion. The variables c_{Lk} and c_{Rk} (as well as LO_{Lk} and LO_{Rk}) represent the confidence criteria for response S_L and S_R separating confidence ratings k and k + 1. Horizontal error bars depict SEM.

Indeed, the CRES models had summed AIC scores that were lower (indicating better fit) than their corresponding BCH models by an average of 2,122 points (range: 1,747 to 2,524). Further, even the worst CRES model slightly outperformed the best BCH model (summed AIC difference = 1,071), demonstrating the robust advantage of the CRES models over their BCH counterparts. Finally, we compared the best CRES and the best BCH models (both models had the same auxiliary assumptions of confidence criteria not being symmetric, and included both lapses and metacognitive noise). We found that the best CRES model outperformed the best BCH model for all 20 subjects (Fig. 3*B*), demonstrating that the advantage of the CRES models is robust across individual subjects too.

To obtain better intuition about the best fitting CRES and BCH models, we assessed how well each model could capture the observed criterion signature from the empirical data (Fig. 2 *B* and *C*). We found that the best CRES model correctly reproduced the pattern of criterion values across the three difficulty conditions (Fig. 3*C*), as *c* stayed relatively constant (average change score of 0.06; empirical value = 0.09) whereas *LO* changed substantially (average change score of 0.45; empirical value = 0.44). The change in *c* across conditions in the model is driven by the inclusion of metacognitive noise and lapse rate (note that the lack of this effect in the predictions shown in Fig. 1*B* comes from not including either of these assumptions). Indeed, both of these parameters introduce randomness in the responses and thus bring the measured criteria closer to zero, with the effect being larger for easier conditions.

Contrary to the good qualitative reproduction of the criterion data by the best CRES model, the best BCH model performed more poorly (Fig. 3*C*). According to that model's fits, confidence criteria in evidence space, *c*, moved further toward zero from the difficult to the easy conditions compared to the empirical data (average change score of 0.17 compared to 0.09 in the empirical data), whereas the confidence criteria in log-likelihood ratio space, *LO*, moved less than in the empirical data (average change score of 0.03 compared to 0.44 in the empirical data).

Beyond the central comparison between CRES and BCH models, the model fits also allowed us to determine the validity of the three auxiliary assumptions. We found strong evidence against the assumption of symmetric confidence criteria, with each model assuming symmetric confidence criteria performing worse than its corresponding model that did not make that assumption (average summed AIC difference = 792). In addition, we found that all models assuming both metacognitive noise and lapse rate performed better than their corresponding models that did not include these parameters (average summed AIC advantage for models with metacognitive noise = 165; average summed AIC advantage for models with lapse rate = 198).

Experiment 2. Experiment 1 provided strong evidence for the notion that confidence is based on CRES rather than BCH computations. However, one issue in that experiment is that subjects were not informed about the difficulty level on each trial. Because BCH computations assumed here require that subjects are aware of the expected sensitivity for each condition, not explicitly providing this information may have prevented people from performing BCH computations. Therefore, we conducted a new, preregistered experiment (Experiment 2) where we informed subjects about the difficulty level on each trial. We also provided subjects with practice that included trial-by-trial feedback to enable them to build an internal model of their performance for each difficulty level (Fig. 4*A*). Beyond these features, Experiment 2 had the same design as Experiment 1 and included a total of 20 subjects each completing 750 trials.

We repeated all the analyses we performed on Experiment 1, and again found robust evidence that subjects performed CRES rather than BCH computations. The empirical data exhibited the signature of CRES computation where confidence criteria in evidence space, *c*, shifted less than the confidence criteria in log-likelihood ratio space, *LO* (average change score for *c*: 0.17; average change score for *LO*: 0.46; Fig. 4 *B*, *Left*). The change score for *LO* was larger than for *c* for all 10 criteria (all *P*'s < 0.04). Examining the best fitting CRES and BCH models (Fig. 4 *B*, *Middle* and *Right*) showed that they both reproduced the *c* values relatively well (change scores: CRES model = 0.06, BCH model = 0.14), but only the CRES model

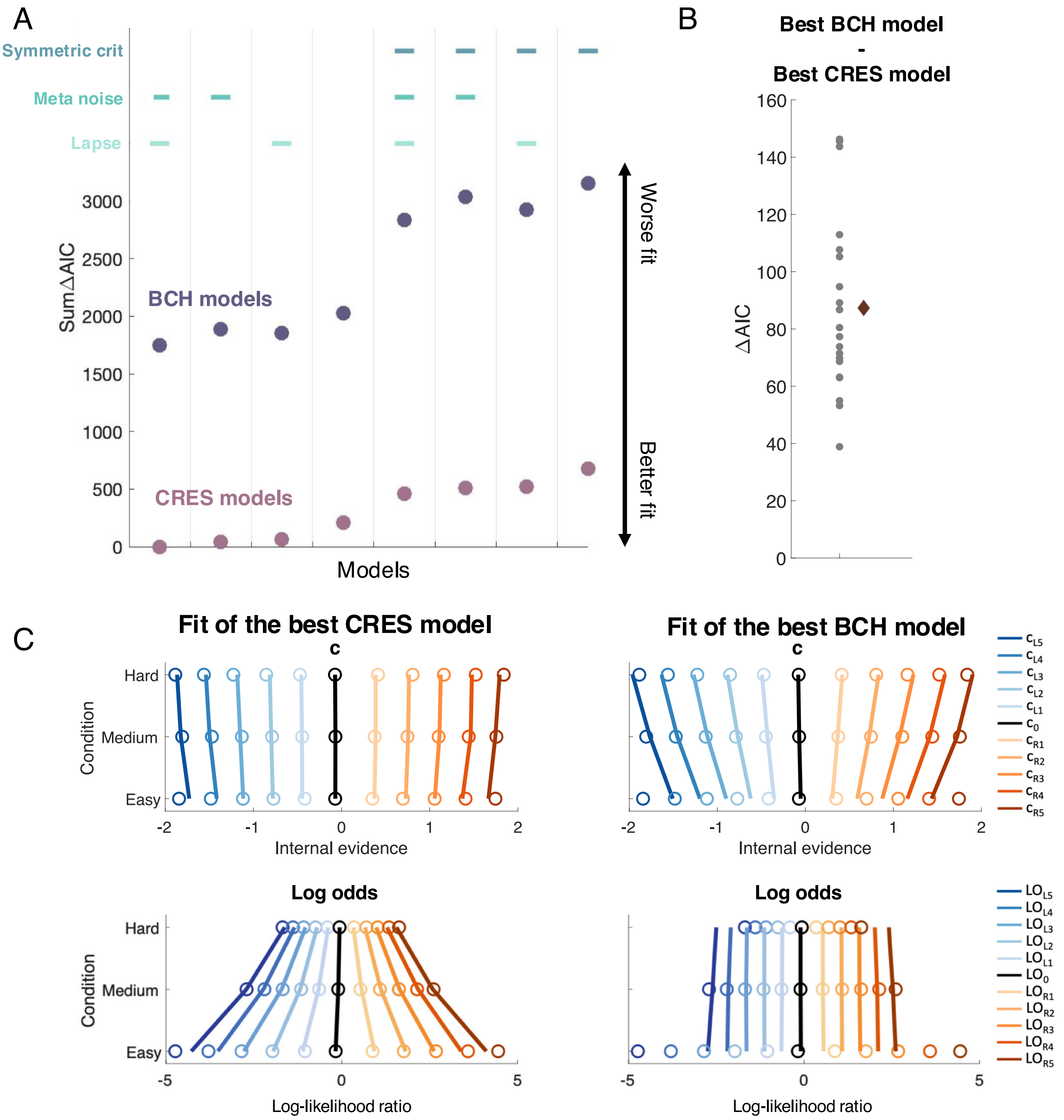

Fig. 3. Model fitting results for Experiment 1. (*A*) The CRES models significantly outperformed the corresponding BCH models regardless of auxiliary assumptions. The figure shows the summed differences in AIC values between each model and the best fitting model. The type of computation (CRES vs. BCH) is indicated by the color of the dots. The auxiliary assumptions are indicated at the *Top*. The order of the models is determined by the rank (from the best to the worst-fitting) of CRES models, and then each of the CRES models is paired up with BCH model that has the same sets of auxiliary assumptions. (*B*) The AIC difference between the best CRES and the best BCH model for individual subjects. A positive value indicates that the CRES model is preferred. The CRES model outperformed BCH model for all 20 subjects. The brown diamond shows the mean value of the AIC differences across subjects. (*C*) Fits of the best CRES and BCH models. Colored lines show model fits, whereas circles depict the empirical *c* and *LO* values from Fig. 2 *B* and *C*.

reproduced the *LO* values well (change scores: CRES model = 0.43, BCH model = 0.12).

Critically, we found that all eight CRES models had a lower summed AIC value compared to the corresponding BCH models (average summed AIC difference = 591.18, range: 244.95 to 906.46; Fig. 4*C*). We also replicated the previous findings regarding the auxiliary assumptions. Namely, our results supported models that confidence criteria are not symmetric (average summed AIC difference = 165.16), and that included both metacognitive noise (average summed AIC difference = 30.95) and lapse rate (average summed AIC difference = 143.59, Fig. 4*C*).

Experiment 3. Experiments 1 and 2 provided strong evidence for the notion that confidence is based on CRES rather than BCH computations. However, the difficulty levels were interleaved in both experiments. This design choice may have made it difficult for subjects to correctly compute the probability of being correct because this computation must be updated on trial-by-trial basis depending on the expected sensitivity. Thus, we conducted another preregistered experiment (Experiment 3) where the difficulty level was fixed within each run of 120 trials. In addition, we provided subjects with trial-by-trial feedback and trained them with long blocks of trials with fixed difficulty. These design features were chosen to make it as easy as possible for subjects to implement the computations required by BCH. Experiment 3 was otherwise identical to Experiments 1 and 2 and included a total of 20 subjects each completing 720 trials (Fig. 5*A*).

We repeated all the analyses from Experiments 1 and 2 and found consistent evidence favoring CRES rather than BCH computations. As in the previous experiments, the confidence criteria in evidence space, *c*, changed less than the confidence

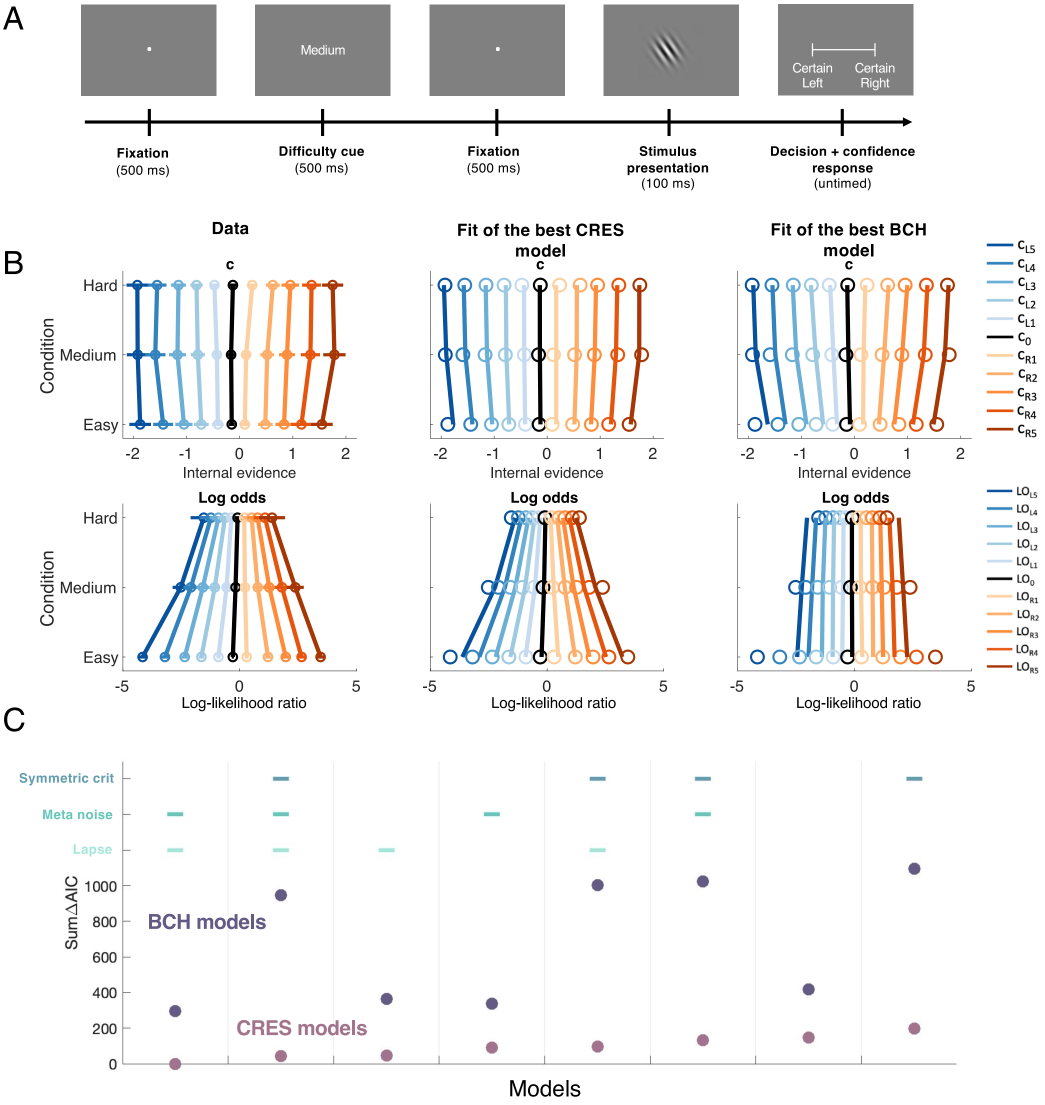

Fig. 4. Design and results for Experiment 2. (*A*) Experimental design. We used the same design as in Experiment 1 with one critical change: each trial began with a cue that indicated the difficulty level on that trial (easy, medium, or difficult). (*B*) Empirical and fitted criterion values in internal evidence space (*c*) and log-likelihood ratio space (*LO*). Colored lines represent confidence criteria, and the central black line represents the decision criterion. (*C*) Summed AIC difference scores between each model and the best fitting model. The CRES models provided better fits compared to the corresponding BCH models across all sets of auxiliary assumptions.

criteria in log-likelihood ratio space, *LO*, (average change score for *c*: 0.19; average change score for *LO*: 0.51; Fig. 5 *B*, *Left*). The change score for *LO* was larger than for *c* for all 10 criteria (all *P*'s < 0.02). Examining the best fitting CRES and BCH models (Fig. 5 *B*, *Middle* and *Right*) showed that they both reproduced the *c* values relatively well (change scores: CRES model = 0.10, BCH model = 0.18), but only the CRES model reproduced the *LO* values well (change scores: CRES model = 0.60, BCH model = 0.06).

Critically, we found that all eight CRES models had a lower summed AIC value compared to the corresponding BCH models (average summed AIC difference = 361.32, range: 68.43 to 651.02; Fig. 5*C*). We also replicated the previous findings regarding the auxiliary assumptions. Namely, our results supported models that confidence criteria are not symmetric (average summed AIC difference = 229.56), and that included both metacognitive noise (average summed AIC difference = 10.65) and lapse rate (average summed AIC difference = 40.19, Fig. 5*C*).

Finally, as preregistered, we examined whether the results change based on the length of exposure to a given difficulty level. Specifically, we repeated these analyses on the first vs. second of the data within each run. We found that on average CRES models outperformed their BCH counterparts for both the first and second half of each run (average summed AIC differences: first half of each run: 178.97, second half of each run: 90.46; *SI Appendix*, Fig. S1). However, the summed AIC difference decreased from the first half to the second half of each run for all eight pairs of models, indicating that the computations used by our subjects did shift toward BCH computation over the course of a run of 120 trials with a difficulty level.

To further investigate potential training effects, we performed an exploratory analysis where we separately examined the data from the first and second half of the entire experiment (i.e., the first three and the last three runs). We then analyzed the qualitative patterns of the measures *c* and log odds (*LO*) for each half. Compared to the first half, the change score of measure *c* increased while the change score of *LO* decreased in the second half of the experiment (First half: change score of *c* = 0.24, *LO* = 0.77; Second half: change score of *c* = 0.47, *LO* = 0.52; *SI Appendix*, Fig. S2). This trend suggests a shift toward a more BCH-like computation over the course of the experiment. However, even in the latter half of the experiment, the

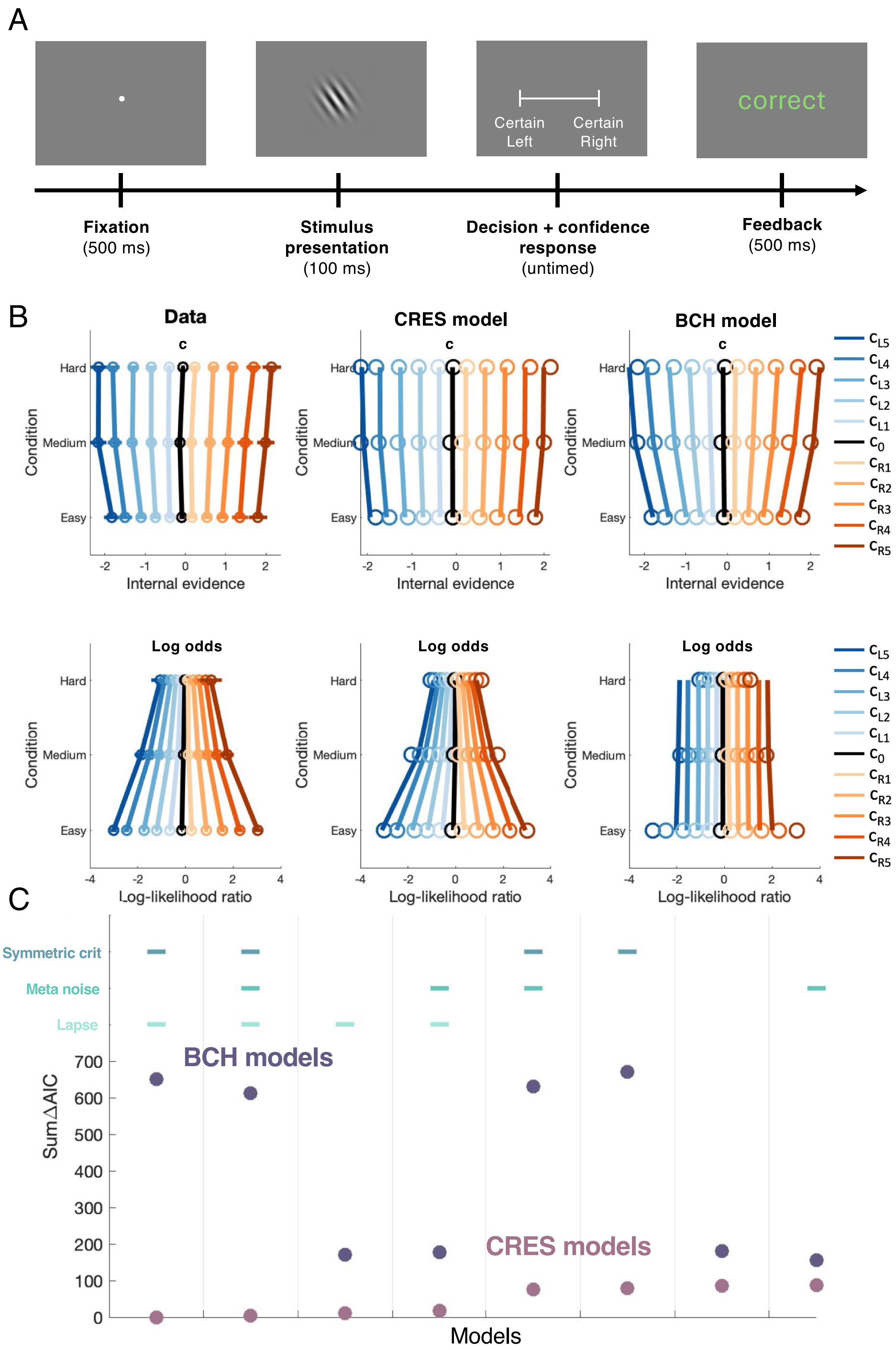

Fig. 5. Design and results for Experiment 3. (*A*) Experimental design. We used a similar design as in Experiments 1 and 2, but each difficulty level was fixed within each run of 120 trials and subjects were given trial-by-trial feedback. (*B*) Empirical and fitted criterion values in internal evidence space (*c*) and log-likelihood ratio space (*LO*). Colored lines represent confidence criteria, and the central black line represents the decision criterion. (*C*) Summed AIC difference scores between each model and the best fitting model. The CRES models provided better fits compared to the corresponding BCH models across all sets of auxiliary assumptions.

qualitative patterns of measures *c* and *LO* still exhibited characteristics indicative of CRES computation.

Similar Results When Response Time Is Used as a Proxy for Confidence. Our analyses focus on self-reported confidence. However, in a control analysis, we considered reaction time (RT) as a proxy for confidence and examined whether it showed the signatures of BCH or CRES computations. Specifically, we converted raw RT values into a measure we call "RT-derived confidence" by coding the faster 25% of RTs as RT-derived confidence of 4, the next 25% of RTs as RT-derived confidence of 3, the next 25% of RTs as RT-derived confidence of 2, and the slowest 25% of RTs as RT-derived confidence of 1. Then, we repeated the analysis we conducted on confidence ratings to determine whether these RT-derived confidence ratings followed the signature consistent with CRES or BCH models. We found that these RT-based analyses produced similar results to confidence (*SI Appendix*, Fig. S3), such that as the RT-derived confidence criteria remained stable in internal evidence space (average change score for *c*: 0.10, 0.13, 0.23 for Experiments 1, 2, and 3) but shifted substantially in log-likelihood ratio space (average change score for *LO*: 0.55, 0.51, 0.63 for Experiments 1, 2, and 3), with the change score for *LO* larger than for *c* for all

10 criteria in all three experiments (all *P*'s < 0.03). Therefore, the CRES computations describe well not only the explicitly provided confidence ratings but also the implicit, RT-based proxy for one's level of certainty.

Discussion

The BCH is often considered the most prominent theory of confidence. Here, we show that there is a behavioral signature that distinguishes BCH from CRES computations in simple 2-choice tasks, namely whether confidence criteria are relatively more stable in log-likelihood ratio vs. internal evidence space. Contrary to BCH, reanalyses of one published experiment, as well as two new experiments that informed subjects about the difficulty of each trial, showed that confidence criteria were much more stable in internal evidence than log-likelihood ratio space. Further, by varying three sets of auxiliary assumptions, we compared eight BCH and corresponding eight CRES models, finding that the CRES models performed better in every pair of models in all three experiments. These results strongly challenge BCH and suggest that human subjects may make confidence judgments by placing criteria directly on the sensory evidence.

The Status of BCH. Our results add to previous work that found evidence against BCH in relatively complex tasks (4, 6, 8). These studies used a combination of external and internal noise (8), a 3-choice task (6), or varied both information quantity (presenting between 2 and 5 dots) and quality (manipulating the variability of the dots' locations) (4). In each case, BCH was outperformed by alternative models. However, one potential issue with these studies is that their relatively complex designs can make BCH computations too computationally demanding. By providing evidence against BCH even in simple 2-choice tasks, the current study goes well beyond previous work in challenging the viability of BCH.

Could our results be explained by BCH-adjacent models that differ minimally from BCH? As noted earlier, BCH involves three distinct assumptions: 1) people are aware of the statistical structure of the task, 2) people estimate the probability of being correct, and 3) people perform this computation correctly. However, for a theory to be Bayesian, only Assumption 2 is required. Thus, BCH is a specific model that makes stronger assumptions than Bayesian models in general. Below, we examine three possible BCH-adjacent models that drop each one of these assumptions.

First, one can consider models that remove Assumption 1, which states that people are aware of the statistical structure of the task. Without any knowledge, it becomes impossible to compute the probability of being correct. However, knowledge of the statistical structure of the task can exist but be partly inaccurate or incomplete. Exploring all possible beliefs regarding the structure of the task can lead to overly flexible and ultimately unfalsifiable models (28–32). That said, for the current experiments, it appears that the only way to change Assumption 1 to fit the behavioral data would be to consider the possibilities that subjects are not aware of the difficulty level on each trial or that they do not realize how different their performance is across difficulty levels. However, Experiments 2 and 3 make these possibilities extremely unlikely given that in these experiments, we explicitly cued the difficulty level and provided very long training periods with feedback.

Second, one can consider models that remove Assumption 2, which states that subjects estimate the probability of being correct. People may instead give confidence directly on the sensory evidence (as CRES does) or compute some other quantity altogether such as entropy (6). However, the assumption that people compute the probability of being correct is so central to BCH, that any model that drops it can no longer be meaningfully described even as Bayesian or "BCH-adjacent."

Finally, one can consider models that remove Assumption 3, which states that people perform the required computation correctly. Decision-makers may make errors in the computation, employ shortcuts, or have internal priors or payoffs that diverge from the experimenter-defined ones. Similar to relaxing Assumption 1, this approach can also lead to overly flexible and ultimately unfalsifiable models (28–32). In the context of the current experiments, to fit the data by changing Assumption 3, one could postulate that subjects are motivated to avoid giving overly high confidence ratings in the easy conditions and that this internal cost function is calibrated just so that criteria stay roughly stable in evidence space. Alternatively, one could postulate that subjects choose to ignore the information about difficulty despite this information being prominently featured in the experiment. Nevertheless, both explanations appear unlikely.

As the discussion above shows, our results could have several competing explanations based on which BCH assumption is most likely to be violated. We speculate that the assumption most likely to be false is Assumption 2 regarding people estimating the probability of being correct. This is because the CRES model—which does not make Assumption 2—parsimoniously fits the data without needing any additional parameters or mechanisms. In fact, given our qualitative results, any BCH-adjacent model would need to behave similarly to a CRES model to fit the data. Further, making BCH-adjacent models that relax Assumptions 1 or 3 requires postulating additional mechanisms that appear unlikely. Nevertheless, our results cannot definitively establish which BCH assumption is violated. What our results do show, however, is that BCH – defined as the combination of the three assumptions—is challenged.

Implications for Bayesian Framework for Decision-Making. While our findings strongly challenge BCH, they should not be interpreted as challenging the broader decision-making Bayesian framework. The Bayesian framework is a framework; that is, it provides a core set of assumptions but is not falsifiable itself (17, 18). Further, challenging BCH does not necessarily question the usefulness of the Bayesian framework, which is fundamentally about a careful consideration of how people can make the best possible decisions given the many constraints they face (33). From that perspective, CRES can be seen as a Bayesian model that incorporates heavy constraints on subjects' ability or willingness to shift their confidence criteria (defined in sensory evidence space) even when explicitly informed about the difficulty of the stimulus. Nevertheless, even a pure CRES model where criteria remain fully fixed in evidence space still allows confidence ratings to serve as proxies for the subjective probability of being correct. Therefore, the current results can be interpreted from a Bayesian perspective without having the abandon the perspective itself.

CRES Models. Our results show that CRES models provide a good explanation to the data, suggesting that confidence may be given by setting criteria directly on the raw evidence. This conclusion opens the question about which decision axes can be considered as "raw evidence" and which cannot. At present, it is unclear where in the brain this evidence may be on what form it might take. That said, we believe that "raw evidence" should be reserved for the output of the sensory cortex in perceptual tasks (and, similarly, the output of memory-related areas in memory tasks). For example, in the context of modern deep neural networks, raw evidence would be the output of the network in its last layer. In contrast, transformations of this output that represent the likelihood or log odds should not be considered raw evidence. Similarly, setting confidence criteria in stimulus space (19, 34, 35) is also not equivalent to setting them in raw evidence space, because the internal evidence space does not always correspond to the stimulus space (36, 37). Nevertheless, outside of standard 2-choice perceptual decision-making tasks where sensory evidence can be assumed to be Gaussian, more work needs to be done to precisely define in an a priori fashion what the raw evidence space is.

Our results show that CRES models provide a better description of the data than BCH models. However, the fits were still sometimes imperfect, suggesting that the current models may not fully capture the confidence computations. Many other models have been examined in the literature. For example, Rausch et al. (38) and Shekhar and Rahnev (39) compared the fits of eight and 15 different models, respectively, finding that different models perform differently on different datasets. The present CRES models were chosen to be as simple as possible and therefore do not reflect many proposals from the literature, such as that confidence computations may reflect stimulus visibility (38, 40, 41), exhibit lognormal rather than Gaussian noise (42, 43), and be influenced by a host of other factors (2, 42). Therefore, we believe that a full description of the computations underlying confidence would involve mechanisms beyond the ones assumed by the current CRES models.

The CRES models examined here assume that the decision and confidence criteria are fixed in internal evidence space. It has sometimes been argued that when different conditions are interleaved, subjects are unable to use different criteria due to working memory constraints (25, 44, 45). However, other work has demonstrated that subjects are in fact able to use different criteria for different conditions, at least when measured in the space of external stimulus features (8, 19, 34, 35, 46). In our three experiments, the criterion measured in internal evidence space, c, had a change score of 0.09, 0.17, and 0.19, respectively. While this change is relatively subtle, the fact that these values are consistently positive suggests that subjects can shift their criteria in the right direction when given the right conditions to do so (that said, the shift was always much smaller than what BCH predicts).

Finally, it is also worth noting that some researchers have proposed models where the decision evidence is first normalized by the estimated sensory noise (2, 4). However, here we modeled the difficulty manipulation as shifting the means of the internal distributions while keeping their SD fixed. In other words, in our modeling framework, the sensory noise was identical across all conditions, and thus models that do or do not assume normalization based on sensory noise cannot be distinguished using the data from the current experiments.

Implications for Asymmetric Confidence Criteria, Metacognitive Noise, and Lapse Rate. Fitting eight different versions of the BCH and CRES models allowed us to make conclusions about the confidence computations that go beyond the distinction between BCH and CRES models. First, we showed that confidence criteria tend to be asymmetric around the decision criterion. It is possible that this asymmetry stems from the fact that subjects reported their confidence on a continuous scale that was then converted into discrete categories for analysis. Nevertheless, we note that similar results were obtained in a previous study that analyzed confidence data given on a 4-point scale (2). The reason for this asymmetry is unclear and should be clarified by future research. Additionally, future research should explore how the use of different confidence scales might impact results in metacognitive studies. Previous research has reported inconsistent findings regarding the effect of scale granularity on confidence calibration, highlighting the need for continued examination in this area (47–49).

Second, we also found evidence that the inclusion of both metacognitive noise and lapse rate improves fits. It should be noted that we followed the standard practice of assuming that lapses lead to a random choice *and* a random confidence rating (2, 7, 8, 19). This assumption is potentially problematic because if a person experiences an attentional lapse during the stimulus presentation, they are likely to give a random choice followed by the lowest possible confidence (rather than a random confidence). In fact, modeling lapses as producing random confidence responses makes them mimic the inclusion of metacognitive noise. Due to the difficulty of a priori determining how subjects would give confidence during potential lapses and the fact that such lapses may not be all that common in human subjects (50), we suggest that adding lapse parameters should not be treated as a default in models of confidence and that more work needs to be done to provide evidence for the plausibility of assuming that subjects provide random confidence ratings during attentional lapses. On the other hand, the existence of metacognitive noise is both theoretically motivated and empirically established (26, 51, 52), and therefore should be included as a common feature in models of confidence.

RT and Confidence. Our implementations of BCH and CRES are static models that do not account for RT. In contrast, a substantial body of research has examined dynamic models that can jointly explain RT, choice, and confidence (53–60). These dynamic models can account for many important psychological phenomena, such as the speed-accuracy tradeoff, that static models cannot explain and also allow for decision-making strategies to potentially change adaptively over the course of a trial. However, dynamic models can make it difficult to formulate specific BCH or CRES model variants. This makes it challenging to compare between BCH and CRES when RT is taken into account in the modeling. Instead, we took a different approach and created RT-derived proxies for confidence and examined whether they showed the signatures of BCH or CRES computation. We found that RT-derived confidence is also better described by CRES computations, suggesting that the timing of decision-making is also more consistent with CRES than BCH computations.

Conclusion

In conclusion, we show that confidence computations deviate substantially from the predictions of BCH even in simple 2-choice tasks. These results strongly challenge BCH and suggest that confidence may involve placing criteria directly in the sensory evidence space.

Methods

Subjects. Experiment 1 was taken from Shekhar and Rahnev (26) and featured 20 subjects each completing a total of 2,800 trials over three sessions held on separate days. Experiment 2 was a new experiment where 21 subjects completed 750 trials each. One subject was excluded from Experiment 2 due to performance lower than 55%, which resulted in 20 subjects in total. The age range was 18 to 24, with an average age of 20.0. In Experiment 3, 23 subjects completed 720 trials each. Three subjects were excluded from Experiment 3 due to performance lower than 55%, which resulted in 20 subjects in total for Experiment 3. The age range was 18 to 21, with an average age of 19.3. All subjects had normal or corrected to normal vision and signed informed consent. Experimental procedures were approved by the Georgia Institute of Technology Institutional Review Board.

Experimental Design.

Experiment 1. The complete details of this experiment are available in the original publication (26). Briefly, each trial began with a fixation point at the center of the screen for a duration of 500 ms, followed by presentation of the stimulus for 100 ms. The stimulus consisted of a Gabor patch with a diameter of 3°, oriented either to the left (counterclockwise) or right (clockwise) of the vertical by 45°. These gratings were overlaid on a background containing random noise. Subsequent to the disappearance of the stimulus, a response screen became visible and was presented until the subjects provided a response. Subjects' task was to indicate the tilt direction (left/right) of the stimulus, while simultaneously rating their confidence using a continuous confidence scale. This scale ranged from 50 to 100% accuracy for each type of response, and subjects were required to indicate their confidence level through a single click of the mouse. The experiment included three different contrast levels (4.5%, 6%, and 8%) that were presented in an interleaved manner. Each subject completed a total of 350 training trials and 2,800 trials of the main experiment.

To encourage accurate confidence ratings, we used a method developed by ref. 61. For each trial, a random number I1 (ranging from one to 100), was generated by the computer. If the reported confidence level *P* exceeded I1, the participant gained a point for a correct response and lost a point for an incorrect one. In this way, the approach penalized overconfidence. Conversely, if *P* was less than or equal to I1, the computer selected another random number, I2, which ranged from one to 100 as well. One point was awarded if I2 was greater than I1 and subtracted otherwise, and thus deincentivized underreporting of confidence. Subjects were informed of the scoring rules and were presented with simulations of different strategies. They were also scored during their practice round to ensure that they were familiar with the scoring system. Furthermore, at the end of each main experiment block, participants were informed of their scores. After finishing three sessions, subjects were given a bonus based on their cumulative scores.

Experiment 2. We preregistered Experiment 2 (https://osf.io/df5ch) before the data collection and followed all the preregistered analyses. Experiment 2 followed the procedure of Experiment 1, except that it also included a cue that indicated the difficulty of the upcoming trial. Including a cue about the difficulty of the upcoming trial made subjects aware of the expected sensitivity and gave them the information needed to perform BCH computations outlined here. Subjects again gained points for reporting confidence based on the expected probability of being correct, but this time did not receive a monetary bonus. During the training blocks, subjects received both trial-by-trial and end-of-block feedback regarding their accuracy for each level of difficulty to ensure they are aware of their performance under different conditions. The cue was presented before the onset of the stimulus presentation for 500 ms. Before the beginning of the experiment, each subject went through five 30-trial blocks of training. In the first block of training, subjects experienced a fixed contrast level of 8% to familiarize themselves with the task and did not have to indicate confidence. Then, blocks 2-4 successively introduced the three contrasts (8%, 6%, 4.5%) while also introducing the confidence rating. In the fifth training block, subjects were familiarized with the design of the actual experiment in which contrast levels were interleaved. The main experiment was organized in three runs each consisting of five 50-trial blocks. Overall, each subject completed 150 trials of training and 750 trials of the main experiment.

Experiment 3. We preregistered Experiment 3 (https://osf.io/enrbc/) before the data collection and followed all the preregistered analyses. Experiment 3 followed most of the procedures of Experiments 1 and 2 with several exceptions. Experiment 3 was organized in six runs each containing four 30-trial blocks. There was only a single difficulty level in each run, meaning that the same difficulty level was presented for 120 consecutive trials. Each difficulty level was randomly repeated twice throughout the experiment, once in the first three runs and once in the last three runs. At the beginning of each block, subjects were informed about the level of difficulty in each block. As in Experiment 1, subjects again received a monetary bonus based on how closely their confidence ratings matched their accuracy. Unlike either of the previous two experiments, subjects also received trial-by-trial feedback during the entire experiment.

Before the beginning of the experiment, each subject first went through seven blocks of training. In the first block, subjects completed 15 trials with a fixed contrast level of 8% and without indicating confidence. Blocks 2 to 4 each included 60 trials each and successively introduced the three contrasts (8%, 6%, 4.5%). Feedback regarding the accuracy of the response was given both right after the response and at the end of each block. Cues regarding the difficulty level were given at the beginning of each block. Finally, blocks 5 to 7 included 15 trials each (one contrast level per block) and introduced the confidence ratings. Overall, each subject completed 240 trials of training and 720 trials of the main experiment.

CRES vs. BCH Computations. Following the signal detection theory (27), we assume that each stimulus presentation generates a sensory response, r, which is corrupted by Gaussian sensory noise, such that $r_{sens} = N\left(-\mu, \sigma_{sens}^2\right)$, when the stimulus belongs to the first category, S$_1$, and $r_{sens} = N\left(\mu, \sigma_{sens}^2\right)$, when the stimulus belongs to the second category, S$_2$. The parameter μ_{sens} represents the distance between the two evidence distributions, whereas σ_{sens} is the SD of each sensory distribution. For simplicity, and without the loss of generality, we set the σ_{sens} to always equal to one.

The CRES models assume that both decision and confidence are generated by placing a set of confidence and decision criteria, $[c_{-n}, c_{-n+1}, \ldots, c_{-1}, c_0, c_1, \ldots, c_{n-1}, c_n]$ on the evidence axis, where n is the number of ratings on the confidence scale. The criteria c_i are monotonically increasing with $c_{-n} = -\infty$ and $c_n = \infty$. When the primary response is "S_2," confidence is generated using the criteria $\left[c_0, c_1, \ldots, c_n\right]$ such that r_{sens} falling within the interval $[c_i, c_{i+1})$ results in a confidence of $i+1$. When the primary response is "S_1," confidence is generated using the criteria $[c_{-n}, c_{-n+1}, \ldots, c_0]$ such that r_{sens} falling within the interval $[c_i, c_{i+1})$ results in a confidence of $-i$.

In contrast, BCH models assume that confidence ratings reflect the probability one's decision is correct. In other words, confidence rating are based on the value of $P\left(S_2|x\right)$ where x represents the amount of sensory evidence. Similar to CRES model, confidence is computed by placing a set of confidence and decision thresholds, $[t_{-n}, t_{-n+1}, \ldots, t_{-1}, t_0, t_1, \ldots, t_{n-1}, t_n]$ However, instead of placing the confidence thresholds on evidence axis like what CRES models assume, BCH models define confidence thresholds in the posterior probability space. Following the BCH principle, we obtain the following for the posterior probability of correctly choosing "S_2" given sensory evidence x (using that the prior probabilities of S_1 and S_2 are equal to 0.5):

$$P\left(S_2|x\right) = \frac{P\left(x|S_2\right) * P\left(S_2\right)}{P(x)} = \frac{P\left(x|S_2\right) * P\left(S_2\right)}{P\left(x|S_1\right) * P\left(S_1\right) + P\left(x|S_2\right) * P\left(S_2\right)}$$
$$= \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-(-\mu))^2}{2}} + \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}} = \frac{1}{e^{-2x\mu} + 1}.$$

Critically, the posterior probability $P\left(S_2|x\right)$ has a one-to-one mapping with the likelihood ratio, $\frac{P(x|S_1)}{P(x|S_2)}$, such that

$$P\left(S_2|x\right) = \frac{P\left(x|S_2\right) \times P\left(S_2\right)}{P\left(x|S_1\right) \times P\left(S_1\right) + P\left(x|S_2\right) \times P\left(S_2\right)}$$
$$= \frac{P\left(x|S_2\right)}{P\left(x|S_1\right) + P\left(x|S_2\right)} = \frac{1}{1 + \frac{P(x|S_1)}{P(x|S_2)}} = \frac{1}{1 + LR},$$

where LR is the likelihood ratio. Thus, we have the formula for LR as

$$LR = \frac{1}{P\left(S_2|x\right)} - 1.$$

By taking the logarithm of the two sides of the formula, we obtain the formula for log odds, LO, which is simply the logarithm of the likelihood ratio:

$$LO = \ln(LR) = \ln\left(\frac{1}{P\left(S_2|x\right)} - 1\right).$$

Therefore, placing confidence thresholds on the log odds is equivalent to placing them on the probability of being correct. Similarly, placing confidence thresholds on the log odds is also equivalent to placing them on the odds. So, placing one's criteria in any of these spaces (log odds, odds, or probability of being correct) results in the same decision behavior.

Qualitatively Different Predictions of BCH and CRES Models. While the BCH models define confidence thresholds t_i in the log-likelihood ratio space, CRES models define confidence criteria c_i in the evidence space. Therefore, it is helpful to translate the confidence thresholds t_i from the log-likelihood ratio space to the evidence space. There is such straightforward correspondence: each threshold t_i in the log-likelihood ratio space corresponds to a c_i in evidence space, such that

$$t_i = \ln\left(\frac{1}{P(S_2|c_i)} - 1\right) = \ln\left(\frac{1}{\frac{1}{e^{-2c_i\mu}+1}} - 1\right) = \ln\left(e^{-2c_i\mu} + 1 - 1\right)$$
$$= \ln\left(e^{-2c_i\mu}\right) = -2\mu c_i = -d' c_i.$$

Thus, for the same confidence criterion c_i across conditions of increasing difficulty (i.e., lower d'), the absolute value of the confidence thresholds t_i decrease linearly with d'. Similarly, for the same confidence threshold t_i across conditions of increasing difficulty (i.e., lower d'), the absolute value of the confidence criteria c_i increase linearly with d'. Therefore, CRES models assume that the confidence criteria c_i stay constant across conditions of increasing difficulty, and lead to confidence thresholds t_i that decrease linearly with d'. On the contrary, BCH models assume that the confidence threshold t_i stay constant across conditions of increasing difficulty, and lead to confidence criteria c_i that increase linearly with d'. The CRES and BCH models thus predict the opposite pattern, providing a very strong test for adjudicating between the two classes of models (Fig. 1*B*).

It is important to note that the divergent predictions made by BCH and CRES models depend on variations in d′ across experimental conditions. This difference in model predictions can only be observed in studies with multiple difficulty levels. Conversely, in experiments with a single difficulty level, BCH and CRES models become indistinguishable.

Computing Empirical Criterion Locations. We computed all confidence and decision criteria, c_i, following the equation from the signal detection theory:

$$c_i = -0.5 * \left(\phi^{-1}(HR_i) + \phi^{-1}(FAR_i)\right),$$

where i goes from $-(n-1)$ to $n-1$ for confidence ratings collected on a n-point scale,

$\phi^{-1}(HR_i)$ represents the z-score of the hit rate associated with the criterion c_i, and $\phi^{-1}(FAR_i)$ represents the z-score of the false alarm rate associated with the criterion c_i. Again, following signal detection theory, we computed log odds (LO_i) as

$$LO_i = d' * c_i,$$

where d' is the stimulus sensitivity measure using the formula:

$$d' = \phi^{-1}(HR_0) - \phi^{-1}(FAR_0),$$

where HR_0 and FAR_0 are the hit and false alarm rates associated with the decision criterion.

We compared how much the criteria measured in internal evidence and log-likelihood ratio space moved between conditions. To do so, we first calculated a change score between the criteria in the hardest and the easiest conditions using the formulas $c_{change} = \left|\frac{c_{hard} - c_{easy}}{c_{hard} + c_{easy}}\right|$ and $LO_{change} = \left|\frac{LO_{hard} - LO_{easy}}{LO_{hard} + LO_{easy}}\right|$. To compare whether the percent change of confidence criteria in the log-likelihood ratio space is statistically larger than that in the internal evidence space, we performed two-tailed *t* tests for the change of each confidence criterion.

Auxiliary Model Assumptions. We tested three auxiliary assumptions: the symmetry of confidence criteria, the existence of metacognitive noise, and the existence of lapses. Varying these three auxiliary assumptions led to eight CRES and eight BCH models where each of those three assumptions was present or absent.

Symmetry of confidence criteria. We fit half of the models with the assumption that the confidence criteria for each choice are symmetric around the decision criterion (the assumption made in (7)), whereas in the other half of models we fit the confidence criteria for each choice separately. For CRES models with the symmetric confidence criteria assumption, confidence criteria $[c_{-n}, c_{-n+1}, \ldots, c_{-1}]$ are computed by flipping confidence criteria $[c_1, \ldots, c_{n-1}, c_n]$ across the decision criterion c_0 such that $c_{-k} = c_0 - c_k$. In this way, the confidence criteria for "S_1" is symmetrical to the confidence criteria for "S_2" around the decision criterion c_0. Thus, given that confidence was given on a 6-point scale in all both experiments, the models that assume symmetric confidence criteria include six free parameters for the criteria (for $c_0, c_1, \ldots, c_5$), whereas the models that do not assume symmetry include 11 free parameters (for $c_{-5}, c_{-4}, \ldots, c_5$).

Presence vs. absence of metacognitive noise. We fit half of the models with the assumption of the presence of metacognitive noise [an assumption made in a lot of previous research, including (24, 26, 62, 63), whereas in the other half of the models we fit models without metacognitive noise. Here, we conceptualized the metacognitive noise as the variability in the confidence criteria, such that the confidence criteria c_i follow a Gaussian distribution, $N(c_i, \sigma^2_{meta})$, centered on the location of the confidence criterion and having constant variability σ_{meta}. Models with the assumption of the presence of metacognitive noise include one extra free parameter (σ_{meta}).

Presence vs. absence of lapse rate. We fit half of the models with the assumption of the presence of lapse rate [an assumption made in a lot of previous research (2, 7, 8, 19), whereas in the other half of the models we fit models without lapse rate. The inclusion of lapse rate is meant to account for trials in which subjects make errors unrelated to the tasks. We made the standard assumption that a lapse led to subjects giving both the perceptual decision and confidence rating randomly. Models that assume the presence of lapse rate include one extra free parameter.

Model Fitting and Model Comparison. Model fitting was based on a maximum likelihood estimation procedure that searches for the set of parameters that maximize the log-likelihood associated with the full probability distribution of responses, using established procedures from our lab (25, 26, 64). The log-likelihood, log$L$, was computed using the following formula:

$$\log L = \sum_{i,j,k} \log(p_{ijk}) * n_{ijk},$$

where p_{ijk} and n_{ijk} are the response probability and number of trials, respectively, associated with the stimulus class $i = \{1, 2\}$, confidence response $j = \{-6, -5 \ldots -1, 1, \ldots 6\}$ (where negative confidence responses correspond to S_1 responses), and stimulus contrast level, $k = \{1, 2, 3\}$. The parameter search was conducted using the Bayesian Adaptive Direct Search (BADS) toolbox, version 1.0.5 (65). To ensure good model fits, we ran the fitting procedure for each model two times and selected the fitted parameters associated with the highest log-likelihood values.

We evaluated how closely the model fits the observed data using the Akaike Information Criterion (AIC). AIC measure the goodness-of-fit of data generated by a certain model, while penalizing the use of additional free parameters. AIC was computed using the standard formula: $\mathrm{AIC} = -2\log L + 2k$, where k indicates the total number of free parameters of a model and n refers to the number of trials in the data. A lower value of AIC indicates better quality of fits. To access whether AIC differences between models are significant, we obtained bootstrapped 95% CI on the AIC differences between models, summed across all subjects. The bootstrapped intervals were computed from 100,000 data samples. CI that do not contain zero are indicative of a significant difference in the AIC values of the models being compared.

To further assess model fit, we computed G^2 (likelihood ratio chi-square) statistics for all three experiments (*SI Appendix*, Table S1). G^2 allows for an assessment of the absolute fit of our models by providing a true zero for the log-likelihood and thus allows the fit to be interpreted as Kullback–Leibler divergence between the model and data. The results of the G^2 statistics are consistent with those obtained using AIC, as all CRES models showed lower G^2 values, indicating a better fit.

Model and Parameter Recovery. Finally, to rule out the possibility that the lack of good fit for the BCH model is primarily due to model recoverability issues, we performed model and parameter recovery analyses for all three experiments. We found that CRES and BCH models have high recoverability in all experiments (*SI Appendix*, Fig. S4) and that the model parameters could be recovered with high fidelity (*SI Appendix*, Fig. S5). These results validate our fitting procedure and the conclusions based on it.

Data, Materials, and Software Availability. Data and code for analysis and model fitting for both experiments are available at https://osf.io/xqgyz/. Previously published data were used for this work (26). All other data are included in the manuscript and/or *SI Appendix*.

ACKNOWLEDGMENTS. This work was supported by the NIH (award: R01MH 119189) and the Office of Naval Research (award: N00014-20-1-2622).

1. J. Metcalfe, A. P. Shimamura, Eds., *Metacognition: Knowing About knowing [Internet]* (The MIT Press, 1994). https://direct.mit.edu/books/book/3931/metacognitionknowing-about-knowing.
2. Z. M. Boundy-Singer, C. M. Ziemba, R. L. T. Goris, Confidence reflects a noisy decision reliability estimate. *Nat. Hum. Behav.* **7**, 142–154 (2023).
3. M. Shekhar, D. Rahnev, How do humans give confidence? A comprehensive comparison of process models of metacognition. *J. Exp. Psychol. Gen.* **153**, 656–688 (2024).
4. S. M. Locke, M. S. Landy, P. Mamassian, Suprathreshold perceptual decisions constrain models of confidence. *PLoS Comput. Biol.* 18, e1010318 (2022).
5. A. Pouget, J. Drugowitsch, A. Kepecs, Confidence and certainty: Distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19**, 366–374 (2016).
6. H. H. Li, W. J. Ma, Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. *Nat. Commun.* **11**, 2004 (2020).
7. L. Aitchison, D. Bang, B. Bahrami, P. E. Latham, Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Comput. Biol.* 11, e1004519 (2015).
8. W. T. Adler, W. J. Ma, Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLoS Comput. Biol.* 14, e1006572 (2018).
9. W. T. Adler, W. J. Ma, Limitations of proposed signatures of Bayesian confidence. *Neural Comput.* **30**, 3327–3354 (2018).
10. A. Kepecs, Z. F. Mainen, A computational framework for the study of confidence in humans and animals. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 1322–1337 (2012).
11. F. Meyniel, M. Sigman, Z. F. Mainen, Confidence as Bayesian probability: From neural origins to behavior. *Neuron* **88**, 78–92 (2015).
12. J. I. Sanders, B. Hangya, A. Kepecs, Signatures of a statistical computation in the human sense of confidence. *Neuron* **90**, 499–506 (2016).
13. B. Hangya, J. I. Sanders, A. Kepecs, A mathematical framework for statistical decision confidence. *Neural Comput.* **28**, 1840–1858 (2016).
14. J. Navajas *et al.*, The idiosyncratic nature of confidence. *Nat. Hum. Behav.* **1**, 810–818 (2017).
15. M. Rausch, M. Zehetleitner, The folded X-pattern is not necessarily a statistical signature of decision confidence. *PLoS Comput. Biol.* 15, e1007456 (2019).
16. A. Bertana, A. Chetverikov, R. S. van Bergen, S. Ling, J. F. M. Jehee, Dual strategies in human confidence judgments. *J. Vis.* **21**, 21 (2021).
17. I. Lakatos, "Falsification and the methodology of scientific research programmes" in *Can Theories Be Refuted? [Internet]*, S. G. Harding, Ed. (Springer Netherlands, Dordrecht, 1976), pp. 205–259, http://link.springer.com/10.1007/978-94-010-1863-0_14.
18. A. Doerig *et al.*, The neuroconnectionist research programme. *Nat. Rev. Neurosci.* **24**, 431–450 (2023).
19. R. N. Denison, W. T. Adler, M. Carrasco, W. J. Ma, Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 11090–11095 (2018).
20. D. Rahnev *et al.*, Consensus goals in the field of visual metacognition. *Perspect. Psychol. Sci.* **17**, 174569162210756 (2022).
21. J. Kwisthout, I. van Rooij, Computational resource demands of a predictive Bayesian brain. *Comput. Brain Behav.* **3**, 174–188 (2020).
22. Y. Jang, T. S. Wallsten, D. E. Huber, A stochastic detection and retrieval model for the study of metacognition. *Psychol. Rev.* **119**, 186–200 (2012).
23. P. Mamassian, V. de Gardelle, Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychol. Rev.* **129**, 976–998 (2022), 10.1037/rev0000312.
24. B. Maniscalco, H. Lau, The signal processing architecture underlying subjective reports of sensory awareness. *Neurosci. Conscious* **2016**, niw002 (2016), 10.1093/nc/niw002/2757122.
25. D. Rahnev *et al.*, Attention induces conservative subjective biases in visual perception. *Nat. Neurosci.* **14**, 1513–1515 (2011).
26. M. Shekhar, D. Rahnev, The nature of metacognitive inefficiency in perceptual decision making. *Psychol. Sci.* **128**, 45–70 (2021).
27. D. M. Green, J. A. Swets, *Signal Detection Theory and Psychophysics* (John Wiley, Oxford, England, 1966), **vol. xi**, p. 455.
28. J. S. Bowers, C. J. Davis, Bayesian just-so stories in psychology and neuroscience. *Psychol. Bull.* **138**, 389–414 (2012).
29. J. S. Bowers, C. J. Davis, Is that what Bayesians believe? Reply to Griffiths, Chater, Norris, and Pouget (2012). *Psychol. Bull.* **138**, 423–426 (2012).
30. F. Eberhardt, D. Danks, Confirmation in the cognitive sciences: The problematic case of Bayesian models. *Minds Mach.* **21**, 389–410 (2011).
31. M. Jones, B. C. Love, Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behav. Brain Sci.* **34**, 169–188 (2011).
32. D. Rahnev, R. N. Denison, Suboptimality in perceptual decision making. *Behav. Brain Sci.* **41**, e223 (2018).
33. F. Lieder, T. L. Griffiths, Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* **43**, e1 (2020).
34. J. L. Lee, R. Denison, W. J. Ma, Challenging the fixed-criterion model of perceptual decision-making. *Neurosci. Conscious* **2023**, niad010 (2023).
35. D. Rahnev, A robust confidence–accuracy dissociation via criterion attraction. *Neurosci. Conscious* **2021**, niab039 (2021).
36. M. L. Green, M. Hu, R. N. Denison, D. Rahnev, Using artificial neural networks to relate external sensory features to internal decisional evidence. PsyArXiv [Preprint] (2023). https://doi.org/10.31234/osf.io/b8ah2 (Accessed 20 September 2024).
37. M. Shekhar, D. Rahnev, Human-like dissociations between confidence and accuracy in convolutional neural networks. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.02.01.578187 (Accessed 20 September 2024).
38. M. Rausch, S. Hellmann, M. Zehetleitner, Modelling visibility judgments using models of decision confidence. *Atten. Percept. Psychophys.* **83**, 3311–3336 (2021).
39. M. Shekhar, D. Rahnev, How do humans give confidence? A comprehensive comparison of process models of perceptual metacognition. *J. Exp. Psychol. Gen.* **153**, 656–688 (2024).
40. S. Hellmann, M. Zehetleitner, M. Rausch, Simultaneous modeling of choice, confidence and response time in visual perception. *Psychol. Rev.* **130**, 1521–1543 (2023), 10.1037/rev0000411.
41. M. Rausch, S. Hellmann, M. Zehetleitner, Confidence in masked orientation judgments is informed by both evidence and visibility. *Atten. Percept. Psychophys.* **80**, 134–154 (2018).
42. M. Shekhar, D. Rahnev, Sources of metacognitive inefficiency. *Trends Cogn. Sci.* **25**, 12–23 (2021).
43. K. Xue, M. Shekhar, D. Rahnev, Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. *Conscious Cogn.* **95**, 103196 (2021).
44. A. Gorea, D. Sagi, Failure to handle more than one internal representation in visual detection tasks. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 12380–12384 (2000).
45. A. Gorea, D. Sagi, Disentangling signal from noise in visual contrast discrimination. *Nat. Neurosci.* **4**, 1146–1150 (2001).
46. I. Zak, M. Katkov, A. Gorea, D. Sagi, Decision criteria in dual discrimination tasks estimated using external-noise methods. *Atten. Percept. Psychophys.* **74**, 1042–1055 (2012).
47. N. Weber, N. Brewer, The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *J. Appl. Psychol.* **88**, 490–499 (2003).
48. C. S. Dodson, D. G. Dobolyi, Confidence and eyewitness identifications: The cross-race effect, decision time and accuracy. *Appl. Cogn. Psychol.* **30**, 113–125 (2016).
49. S. Jin, P. Verhaeghen, D. Rahnev, Across-subject correlation between confidence and accuracy: A meta-analysis of the confidence database. *Psychon. Bull. Rev.* **29**, 1405–1413 (2022).
50. M. W. Schurgin, J. T. Wixted, T. F. Brady, Psychophysical scaling reveals a unified theory of visual memory strength. *Nat. Hum. Behav.* **4**, 1156–1172 (2020).
51. S. T. Mueller, C. T. Weidemann, Decision noise: An explanation for observed violations of signal detection theory. *Psychon. Bull. Rev.* **15**, 465–494 (2008).
52. D. Rahnev, A. Koizumi, L. Y. McCurdy, M. D'Esposito, H. Lau, Confidence leak in perceptual decision making. *Psychol. Sci.* **26**, 1664–1680 (2015).
53. R. Ratcliff, J. J. Starns, Modeling confidence and response time in recognition memory. *Psychol. Rev.* **116**, 59–83 (2009).
54. C. Voskuilen, R. Ratcliff, Modeling confidence and response time in associative recognition. *J. Mem. Lang.* **86**, 60–96 (2016).
55. S. Chen, D. Rahnev, Confidence response times: Challenging postdecisional models of confidence. *J. Vis.* **23**, 11 (2023).
56. T. J. Pleskac, J. R. Busemeyer, Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychol. Rev.* **117**, 864–901 (2010).
57. D. Vickers, *Decision Processes in Visual Perception* (Academic Press, 1979).
58. R. Ratcliff, J. J. Starns, Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychol. Rev.* **120**, 697–719 (2013).
59. S. Hellmann, M. Zehetleitner, M. Rausch, Simultaneous modeling of choice, confidence, and response time in visual perception. *Psychol. Rev.* **130**, 1521–1543 (2023).
60. R. Moran, A. R. Teodorescu, M. Usher, Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognit. Psychol.* **78**, 99–147 (2015).
61. S. M. Fleming, S. Massoni, T. Gajdos, J. C. Vergnaud, Metacognition about the past and future: Quantifying common and distinct influences on prospective and retrospective judgments of self-performance. *Neurosci. Conscious* **2016**, niw018 (2016).
62. J. W. Bang, M. Shekhar, D. Rahnev, Sensory noise increases metacognitive efficiency. *J. Exp. Psychol. Gen.* **148**, 437–452 (2019).
63. M. Shekhar, D. Rahnev, Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. *J. Neurosci.* **38**, 5078–5087 (2018).
64. D. A. Rahnev, B. Maniscalco, B. Luber, H. Lau, S. H. Lisanby, Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *J. Neurophysiol.* **107**, 1556–1563 (2012).
65. L. Acerbi, W. J. Ma, Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. arXiv [Preprint] (2017). https://doi.org/10.48550/arXiv.1705.04405 (Accessed 16 August 2024).