Cognition 272 (2026) 106490

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full Length Article

Stimulus reliability but not boundary distance manipulations violate the folded-X pattern of confidence

Kai Xue [*,1], Herrick Fung [**,1], Dobromir Rahnev

School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA

ARTICLE INFO

Keywords:
Confidence
Perceptual decision making
Visual metacognition
Confidence-accuracy dissociation
Folded-X pattern
Artificial neural networks

ABSTRACT

The folded-X pattern has been identified as a critical signature of confidence: as conditions become easier, confidence increases for correct trials but decreases for error trials. However, recent work has identified violations of the folded-X pattern where easier conditions lead to increased confidence for both correct and error trials (double-increase pattern). Nevertheless, it remains unclear which stimulus manipulations produce each pattern. Here we test the hypothesis that the double-increase pattern emerges for manipulations of the quality of the sensory input (stimulus reliability), whereas the folded-X pattern emerges for manipulations of the distance between the relevant sensory feature and the decision boundary (boundary distance). Across two experiments (N = 78) using orientation judgment with either Gabor patches or moving dots, we first replicate previous findings that boundary distance manipulations have a stronger effect on accuracy whereas stimulus reliability manipulations have a stronger effect on confidence. Critically, boundary distance manipulations produced the classic folded-X pattern, whereas stimulus reliability manipulations yielded the double-increase pattern. Artificial neural networks (ANNs) trained on the same tasks exhibited folded-X patterns for both manipulations, suggesting that human confidence judgments do not simply reflect the statistical nature of the task and stimuli. Reaction time (RT) patterns largely mirrored confidence, though with some notable exceptions, underscoring both the utility and limitations of RT as a confidence proxy. These results demonstrate that different stimulus manipulations have dissociable effects on the signatures of confidence and suggest that human confidence is influenced by mechanisms nor present in standard ANNs.

1. Introduction

Confidence judgments provide crucial insights into how the brain evaluates the quality of its own decisions, representing a fundamental aspect of metacognition. To identify neural correlates of decision confidence, researchers have relied on statistical signatures that characterize confidence patterns across different stimulus conditions (Kepecs & Mainen, 2012).

The most widely used statistical signature is the “folded-X" pattern, wherein easier conditions lead to confidence increases for correct trials but confidence decreases for error trials (Rausch & Zehetleitner, 2019; Sanders, Hangya, & Kepecs, 2016). This pattern has played a central role in confidence research because it has been proposed to reflect a fundamental computational property of how the brain derives confidence from decision evidence. The folded-X pattern has been used as a criterion for identifying neural correlates of confidence in both humans and non-human animals, with studies demonstrating its presence across domains including auditory discrimination, general knowledge tasks, and certain visual discrimination paradigms (Moran, Teodorescu, & Usher, 2015; Sanders et al., 2016). This widespread adoption as a diagnostic signature has made it important for interpreting confidence-related neural signals: neural activity that tracks the folded-X pattern is often interpreted as directly representing confidence computations (Kepecs et al., 2008a; Sanders et al., 2016).

However, recent work has identified systematic violations of the folded-X pattern where easier trials increase confidence for both correct and error trials (“double-increase” pattern). For instance, higher motion coherence in random-dot motion tasks produced the double-increase pattern (Kiani, Corthell, & Shadlen, 2014; Van Den Berg et al., 2016). Similarly, in a masked orientation task, smaller stimulus-onset

* Corresponding author at: J.S. Coon Building, Room B76, 654 Cherry St NW, Atlanta, GA 30332, USA.
* Corresponding author.
E-mail address: kxue33@gatech.edu (K. Xue).
[1] K.X. and H.F. contributed equally to this work as joint first authors.

https://doi.org/10.1016/j.cognition.2026.106490
Received 4 August 2025; Received in revised form 11 February 2026; Accepted 15 February 2026
Available online 24 February 2026

asynchronies (leading to stronger masking) also produced a double-increase pattern (Rausch, Hellmann, & Zehetleitner, 2018). However, it remains an open question what types of stimulus manipulations lead to the folded-X vs. double-increase patterns.

Understanding what manipulations lead to the folded-X vs. the double-increase pattern is critical for evaluating computational models of confidence. Indeed, multiple studies have demonstrated that many models of confidence cannot account for the double-increase pattern of confidence (Hellmann et al., 2023b; Rausch et al., 2018; Rausch & Zehetleitner, 2019). Similarly, a recent paper showed that model comparison results change dramatically when models are used to explain datasets that show the folded-X vs. the double-increase patterns (Shekhar & Rahnev, 2024). As such, building models that robustly capture confidence across a range of paradigms without becoming overly flexible requires clear understanding of what stimulus manipulations produce the folded-X vs. the double-increase patterns.

In perceptual decision making, researchers have manipulated numerous stimulus features to investigate how these changes affect accuracy and confidence. For example, a task can be made harder by reducing stimulus contrast (Cannon, 1979; Lesica et al., 2007; Nauhaus, Busse, Carandini, & Ringach, 2009), increasing the level of noise in the stimulus (Adler & Ma, 2018a; Magliero, Bashore, Coles, & Donchin, 1984), decreasing motion coherence (Williams, Phillips, & Sekuler, 1986; Yuille & Grzywacz, 1988), or increasing the spread of stimulus features over a set of stimuli (de Gardelle & Summerfield, 2011; Desender, Boldt, & Yeung, 2018). Other common manipulations include decreasing the angular distance of dot motion from the choice boundary (Bang & Fleming, 2018; de Gardelle & Mamassian, 2015), reducing the mean distance of a cloud of dots from the center of the screen in a left/right discrimination task (Locke, Landy, & Mamassian, 2022), and using colors closer to purple in a red/blue discrimination task (Boldt, De Gardelle, & Yeung, 2017; de Gardelle & Summerfield, 2011). However, despite the proliferation of different manipulations, much less work has been done on classifying these manipulations into different categories that share fundamental similarities.

One classification that has emerged in recent years is that the various stimulus manipulations can be separated into two broad categories: (1) manipulations of stimulus reliability and (2) manipulations of boundary distance (Bang & Fleming, 2018; de Gardelle & Mamassian, 2015; de Gardelle & Summerfield, 2011). In this classification, "stimulus reliability" refers to any manipulation that changes subjects' ability to identify in absolute terms the critical feature of the stimulus that has to be judged. These manipulations are typically easily noticeable and affect stimulus quality without changing the visual feature on which the decision is based. Consequently, they are also sometimes referred to as "auxiliary" or "non-task-defining" manipulations (Fung, Shekhar, Xue, Rausch, & Rahnev, 2025). For example, decreasing the contrast of a Gabor patch or increasing its noise level make it more difficult to identify its tilt even though the tilt itself (the task-defining feature) has not been changed. Similarly, decreasing the coherence of dot motion or increasing the spread of color values in a group of circles make it more difficult to identify the direction of motion or average color. In general, low stimulus reliability leads to ambiguous stimuli where the critical feature to be judged is difficult to identify.

In contrast, "boundary distance" refers to any manipulation that changes the ability to decide the category of the stimulus without affecting the ability to identify the critical feature. These manipulations directly affect the stimulus feature that determines the categorization decision and have therefore been described as "task-defining" manipulations (Fung, Shekhar, et al., 2025). For example, a Gabor patch of given contrast and noise level is more difficult to categorize as tilted clockwise or counterclockwise from 45° when its true orientation is 44° than when it is 34°. In this case, the reliability of the stimulus itself does not change because orientations of 44° and 34° are equally difficult to identify in absolute terms (i.e., if subjects were asked to identify the true orientation, they would have similar levels of precision for 44° and 34° orientations). Instead, the difficulty in these manipulations comes from the critical feature being either close or far from the decision boundary. To give a different example, it is easier to determine the gender for male and female faces (far from the decision boundary) than for a male-female morph (close to the decision boundary), even if the morph image is as clear and easy to perceive as the original images. As a rule of thumb, stimulus reliability manipulations make the stimuli ambiguous and difficult to perceive, whereas boundary manipulations do not affect the clarity of the stimulus itself but make it harder to categorize as belonging to one class or another, thereby affecting categorization difficulty.

Previous research has found differences in how these manipulation types affect confidence and accuracy. For example, Spence, Dux, and Arnold (2016) and Boldt et al. (2017) found that stimulus reliability manipulations had a stronger effect on confidence than boundary distance manipulations when accuracy was matched. However, no study has systematically investigated whether these manipulation types might systematically produce either the folded-X or double-increase patterns. Previous modeling work suggests that when subjects have information about the difficulty of the task, they can use this information to adjust their overall confidence ratings (Rausch & Zehetleitner, 2019). For example, if a subject infers that a trial is easy, they may use higher confidence. Such confidence increase would occur for both correct and error trials, thus resulting in a double-increase pattern of confidence. Critically, since it is easier to infer the difficulty of a trial for stimulus reliability than for boundary distance manipulations (Fung, Shekhar, et al., 2025), a natural prediction is that stimulus reliability manipulations would lead to the double-increase pattern, whereas boundary distance manipulations would lead to the folded-X pattern. Nevertheless, the effect of these two types of manipulations on the folded-X vs. double-increase patterns has not been tested within the same experimental paradigm.

Here we examine whether the stimulus reliability and boundary distance manipulations produce different confidence patterns (folded-X vs. double-increase). We performed two large experiments (total $N = 78$) using orientation judgments with either Gabor patches or moving dots. We manipulated stimulus reliability by varying the contrast of Gabor patches and the coherence of dot motion. We manipulated boundary distance by varying the degree of offset from 45°. We first replicated previous findings that, compared to stimulus reliability manipulations, boundary distance manipulations more strongly affect accuracy, while stimulus reliability manipulations have a greater impact on confidence. Critically, we discovered that boundary distance manipulations consistently produced the standard folded-X pattern, while stimulus reliability manipulations produced the double-increase pattern. This dissociation, replicated across both experiments, demonstrates that stimulus reliability and boundary distance manipulations have dissociable effects on the signatures of confidence. Importantly, artificial neural networks (ANNs) demonstrated a folded-X pattern for both stimulus reliability and boundary distance manipulations, suggesting that human confidence judgments reflect additional metacognitive mechanisms beyond the statistical properties of the task and stimuli. These findings show that different stimulus manipulations produce distinct effects on confidence patterns, indicating that human confidence computation relies on processes that go beyond the mechanisms present in conventional neural networks.

2. Methods

2.1. Subjects

We recruited 78 subjects from the Georgia Institute of Technology community for either monetary compensation ($10/h) or course credit equivalent. All subjects reported to have normal or corrected-to-normal vision and signed informed consent before the experimental procedure. Experiment 1 had 57 subjects and Experiment 2 had 21 subjects. No

subjects participated in both experiments. We excluded subjects with chance-level accuracy (lower than 55%). This led to excluding two subjects from Experiment 1 and four subjects from Experiment 2. Experimental procedures were approved by the Georgia Institute of Technology Institutional Review Board.

2.2. Experimental design

2.2.1. Experiment 1

Each trial began with subjects fixating on a small white dot (size = 0.05°) at the center of the screen for 1 s. A Gabor patch (diameter = 4°) oriented either to the left (counterclockwise) or right (clockwise), relative to a 45° tilt, was then presented for 100 milliseconds (Fig. 1). The Gabor patch was superimposed on a noisy background. Subjects judged whether the Gabor patch is tilted left (counterclockwise) or right (clockwise), relative to a 45° tilt. A red line indicating a 45° tilt was drawn on a circle for reference. Subjects then rated their confidence on a 4-point scale (1 = low confidence, 4 = high confidence). Subjects made both responses on a standard keyboard with no time constraints. The main experiment comprised 4 runs, each containing 5 blocks of 45 trials, with fixed 15-s breaks after each block and longer subject-controlled breaks after each run. In total, each subject completed 900 experimental trials.

Stimulus reliability and boundary distance were manipulated by varying the contrast and tilt offset of the Gabor patches. For the manipulation of stimulus reliability, we sampled three levels of contrast: 0.3 (low), 0.45 (medium), and 1 (high). For the manipulation of boundary distance, we first determined each individual's threshold (T) using a staircase procedure (see below), and then sampled tilt offsets of 0.5 T (low), T (medium), and 2 T (high). The mean threshold across subjects was $T = 2.26°$ (SD = 0.89°), which means that oblique-effect anisotropies (Girshick, Landy, & Simoncelli, 2011) were negligible in our design. Indeed, based on the internal-noise model from Bertana, Chetverikov, van Bergen, Ling, and Jehee (2021), our design produced only a 2% increase in average sensory precision between the smallest and largest tilt offsets (0.5 T vs. 2 T).

Within each block, three levels of tilt offset appeared in random order, and contrast levels appeared in a random cyclic pattern such that each contrast level appeared once every three trials in random order. Importantly, while each contrast level was guaranteed to appear once within every three trials, the specific order within each set of three was randomized (e.g., the sequence might be low-medium-high, then low-high-medium, then high-low-medium, etc.), preventing participants from predicting which contrast level would appear on any given trial.

Subjects underwent three training blocks and two staircase blocks before the actual experiment. The first block always had tilt offset of 10° relative to 45°. The second practice block contained 40 trials with the tilt levels being progressively more difficult (tilt offset of 5°, 3°, 2°, and 1°, with 10 trials per level). The third practice block had three tilt levels (4°, 2°, and 1°) and that randomly interleaved over 20 trials. The three levels of contrast in the training were the same as they were in actual experiments (0.3, 0.45, 1). Subjects then performed a 2-down-1-up staircase procedure and a 3-down-1-up staircase procedure to determine the individual tilt offset (T). The thresholds from both staircase blocks were then averaged to obtain the final value T used for the tilt offset manipulation. We staircased on orientation because in pilot testing we observed substantial individual differences in sensitivity to orientation, but relatively consistent sensitivity to contrast across participants. The staircase procedures were conducted at a fixed medium contrast level (0.45) to establish a baseline threshold for the boundary distance manipulation.

2.2.2. Experiment 2

Experiment 2 is identical to Experiment 1 except as noted below. Instead of a Gabor orientation discrimination task, we performed a dot motion direction discrimination task in Experiment 2. The stimulus consisted of a black circle (diameter = 3°) containing 113 moving white dots. These dots moved either to the left (counterclockwise) or right (clockwise), relative to a 45° tilt for 300 milliseconds (Fig. 1). Participants judged whether the tilt direction of the dot motion was left/right

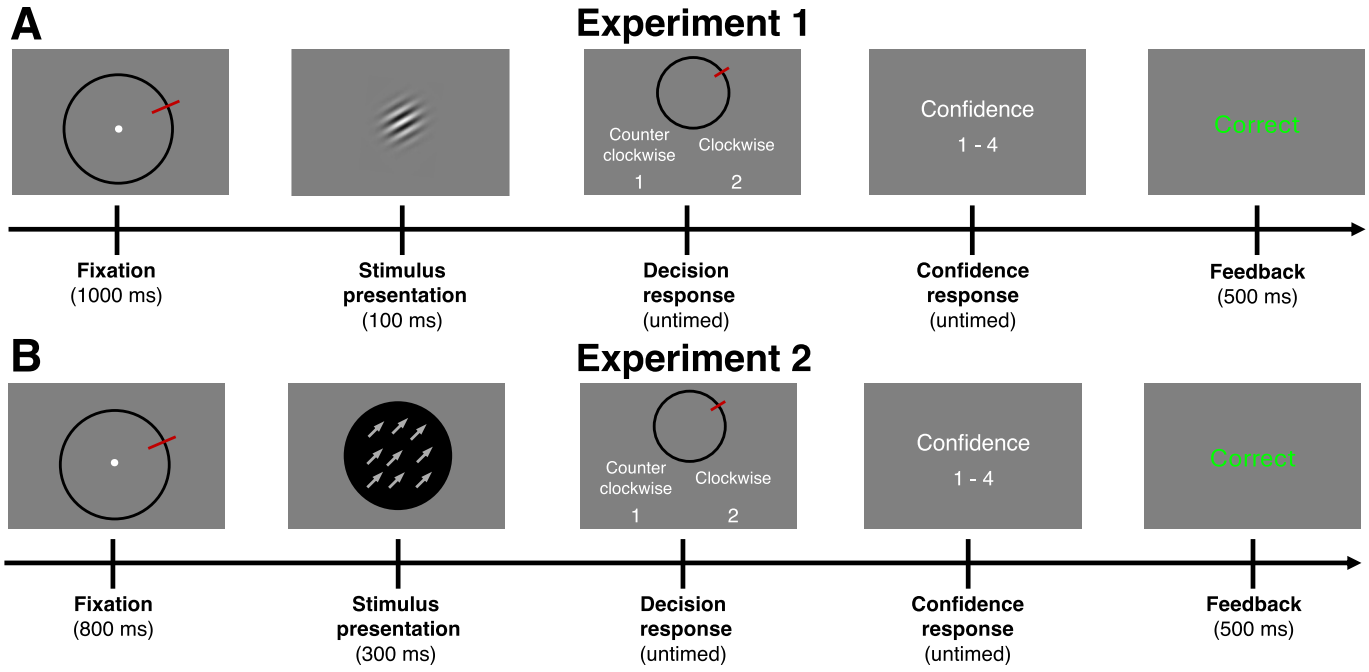

Fig. 1. Experimental design. (A) Experiment 1 design. Subjects viewed a noisy Gabor patch tilted clockwise or counterclockwise relative to 45°, indicated its tilt, and rated their confidence on a 4-point scale. We manipulated the contrast of the Gabor patches (low, medium, high) and the degree of tilt offset relative to 45° (low, medium, high). (B) Experiment 2 design. The design of Experiment 2 mirrored that of Experiment 1, except the stimulus was random dots moving clockwise or counterclockwise relative to 45°. We manipulated the coherence of the moving dots (low, medium, high) and the degree of tilt offset relative to 45° (low, medium, high).

relative to a 45° tilt and reported their confidence. We manipulated the stimulus reliability by varying the motion coherence of the dot motion (motion coherence = 0.3, 0.4, and 0.7, for low, medium, and high, respectively). We manipulated boundary distance by sampling three tilt offsets (0.5xT, T, and 2xT, for low, medium, and high, respectively, where T is the individualized threshold determined by the staircase procedure). The mean threshold across subjects was $T = 5.0°$(SD = 0.96°). We again had three training blocks. In the first training block, the tilt offset was set to 20°. In the second training block, the tilt offsets become progressively more difficult (tilt offsets of 10°, 7°, 5°, and 3°). Finally, the tilt offsets in the third training block were 6°, 3°, and 1.5°. As in Experiment 1, subjects performed two staircase procedure to determine their individual tilt offset (T) before the experimental trials. We again staircased on orientation at the medium coherence level. The staircase procedures were conducted at a fixed medium coherence level (0.4) to establish a baseline threshold for the boundary distance manipulation.

We note that dot motion displays are known to lead to direction reversals – a phenomenon where observers estimate the motion direction to be in the precise opposite direction from the true motion direction. The frequency of such estimation judgments is approximately 8–10% in motion displays similar to ours (Bae & Luck, 2022) but it is unclear whether these reversals are specific to estimation task or also affect two-choice tasks like ours. Such reversals, if present in our task, would act as a relatively small, symmetric source of additional reliability-related noise that would affect all three boundary-distance conditions roughly equally and would not undermine the conceptual distinction between stimulus reliability and boundary distance manipulations.

2.2.3. Apparatus

The experiment was counducted in a dimly lit room. Stimuli were presented on a 21.5 in. 1920 × 1080 monitor with a refresh rate of 60 Hz, using a custom software written in MATLAB (The MathWorks) using the Psychophysics Toolbox (Brainard, 1997). The viewing distance was about 60 cm.

2.2.4. Analyses

We first computed subjects' task sensitivity (*d'*) and confidence on each experimental condition. We calculated *d'* using the following formula:

$$d' = \varphi^{-1}(\textit{hit rate}) - \varphi^{-1}(\textit{false alarm rate})$$

where φ^{-1} is the inverse of the cumulative standard normal distribution that transforms hit rate and false alarm rate into *z* scores.

We first confirmed that the stimulus reliability and boundary distance manipulation affected task sensitivity and confidence. We conducted two 3 (stimulus reliability) x 3 (boundary distance) repeated-measures analyses of variance (RM-ANOVAs) with *d'* and confidence as the dependent variables. We also computed the differences between the low and high levels for both manipulations in *d'* and confidence. We then conducted paired *t*-tests to compare the effect of stimulus reliability and boundary distance.

Additionally, we examined the differences between pairs of conditions that featured different extremes. Specifically, we compared three types of trials: 1) trials with the highest level of stimulus reliability and the lowest level of boundary distance (High reliability/Low distance), 2) trials with the medium level of stimulus reliability and the medium level of boundary distance (Med reliability/Med distance), and 3) trials with the lowest level of stimulus reliability and the highest level of boundary distance (Low reliability/High distance). The comparison between these types of trials was performed using paired t-tests.

We examined the folded-X pattern where easier conditions lead to increased confidence for correct trials but decreased confidence for error trials (Hangya, Sanders, & Kepecs, 2016). We fit linear mixed-effect models separately for correct and error trials. The models were also fit separately for the stimulus reliability and boundary distance manipulations by only modeling the manipulation of interest. The sign of the slope in the resulting model (β) indicates whether confidence is increasing for easier conditions. For each subject, we obtained individual beta coefficients, which were then averaged across subjects to yield the reported average beta values. The folded-X pattern corresponds to a positive slope for correct trials and a negative slope for error trials.

This approach of fitting separate models for each manipulation type directly corresponds to our visualization strategy in Figs. 4, 5, and 7, where we plot confidence patterns separately for stimulus reliability and boundary distance manipulations. By analyzing each manipulation independently while averaging across levels of the other manipulation, we obtain beta coefficients that directly represent the slopes shown in these figures, facilitating an intuitive interpretation of whether each manipulation produces a folded-X or double-increase pattern. Nevertheless, we also conducted a unified linear mixed-effect model that simultaneously tested both manipulations and their interactions in a single analysis and obtained equivalent results (see Supplementary Results).

Finally, we performed an across-experiment comparison of confidence and *d'*. We averaged *d'* and confidence across all conditions within each experiment. We then compared the obtained average *d'* and confidence between experiments using an independent-sample *t*-tests.

2.2.5. Artificial neural network (ANN) simulations

To examine whether the human dissociations between stimulus reliability and boundary distance manipulations could arise purely from the statistical structure of the stimuli and task, we trained artificial neural networks (ANNs) on a Gabor orientation discrimination task with stimuli that closely matched our Experiment 1. To ensure the robustness of our results and enable statistical testing, we trained 30 independent ANN instances that differed only in their random initializations and used their distribution for statistical inferences (Fung, Murty, & Rahnev, 2025; Rafiei, Shekhar, & Rahnev, 2024; Shekhar, Fung, Saxena, Rafiei, & Rahnev, 2025).

Each ANN consisted of four layers: two convolutional layers followed by two fully connected layers (Green, Hu, Denison, & Rahnev, 2026). The output of the network was a single decision unit. The ANN made the perceptual decision based on the activation in this decision unit: negative activations indicated counterclockwise decisions, while positive activations indicated clockwise decisions. Confidence was defined as the absolute value of the activation in the decision unit.

All 30 instances were trained on a custom dataset of Gabor patches with about 20,000 images of Gabor patches. The Gabor images were 160 × 160 pixels and scaled to [0,1]. The tilts were randomly sampled so that their orientation deviates between 0.05° and 2° from 45°, with the three levels contrast (0.3, 0.45, 1.0) and two classes (clockwise/counterclockwise) fully counterbalanced. About 14,000 images were used during training, with the remainder serving as the validation set. All instances were trained for 2 epochs with a batch size of 32 using the binary cross-entropy loss function and the Adam optimizer (learning rate = 1e-4). All instances reached performance of above 90% in the validation dataset.

To enable model comparison with the human behavioral data, we roughly matched the average accuracy between humans and model instances by adjusting the tilt offset (search from 0.05° to 2° in 25 evenly spaced levels; Fung, Murty, et al., 2025; Rafiei et al., 2024) with a fixed contrast level of 0.45. The closest tilt values that match to human accuracy were 0.05°, 0.1313°, and 0.2125° for the three levels of tilt, respectively. For testing, we generated 2000 new Gabor patches for each combination of three tilt levels (0.05°, 0.1313°, and 0.2125°) and three contrast levels (0.3, 0.45, 1.0), for both clockwise and counterclockwise classes. This resulted in a total of 36,000 test images (2000 × 3 tilts × 3 contrasts × 2 classes). We computed accuracy and confidence for the full 3 × 3 factorial combination of contrast and tilt offset across all 30

trained ANN instances and the confidence-accuracy dissociation and folded-X pattern were analyzed following the same procedure as in the human data.

3. Results

We examined whether stimulus reliability and boundary distance manipulations produce dissociable confidence patterns (i.e., folded-X vs. double-increase pattern). We used a 3 × 3 factorial design, where each manipulation was tested at low, medium, and high levels. Subjects performed a Gabor orientation discrimination task in Experiment 1 and a dot motion direction discrimination task in Experiment 2. We first compared the results of the stimulus reliability and boundary distance manipulations on task sensitivity (*d'*) and confidence, and then examined the folded-X vs. double-increase pattern.

3.1. Replicating the dissociable effects of stimulus reliability and boundary distance on performance and confidence

3.1.1. Experiment 1: gabor orientation discrimination

We first aimed to replicate previous findings suggesting that stimulus reliability and boundary distance manipulations have dissociable effects on accuracy vs. confidence (Boldt et al., 2017; de Gardelle & Mamassian, 2015; Spence et al., 2016; Zylberberg, Roelfsema, & Sigman, 2014). We confirmed that both manipulations have significant main effects on both d' (stimulus reliability: $F(2, 108) = 82.78$, $p = 1.60 \times 10^{-22}$; boundary distance: $F(2, 108) = 252.85$, $p = 1.80 \times 10^{-41}$; Fig. 2A) and confidence (stimulus reliability; $F(2,108) = 105.07$, $p = 4.60 \times 10^{-26}$; boundary distance: $F(2,108) = 74.34$, $p = 4.98 \times 10^{-21}$; Fig. 2B). Specifically, d' increased by 0.65 from low to high stimulus reliability ($t(54) = 10.68$, $p = 6.33 \times 10^{-15}$) and by 1.23 from low to high boundary distance ($t(54) = 17.14$, $p = 1.69 \times 10^{-23}$), whereas confidence increased by 0.49 from low to high stimulus reliability ($t(54) = 11.21$, $p = 1.04 \times 10^{-15}$) and by 0.23 from low to high boundary distance ($t(54) = 9.30$, $p = 8.40 \times 10^{-13}$).

Critically, we compared the effects of stimulus reliability and boundary distance on *d'* and confidence (Fig. 2C). We quantified the effect of stimulus reliability on performance and confidence by computing the difference in *d'* and confidence between the low and high stimulus reliability. We similarly quantified the effects of boundary distance by computing the difference in *d'* and confidence between the low and high boundary distance. We found that stimulus reliability had a smaller effect on d' compared to boundary distance ($t(54) = 9.31$, $p = 7.94 \times 10^{-13}$, Cohen's $d = 1.26$). Conversely, stimulus reliability had a larger effect on confidence compared to boundary distance ($t(54) = 6.01$, $p = 1.66 \times 10^{-7}$, Cohen's $d = 0.81$). These findings replicate the previous literature by demonstrating that stimulus reliability and boundary distance differentially impact d' and confidence.

To provide a clearer illustration of the dissociation between stimulus reliability and boundary distance, we additionally examined pairs of conditions that traded off the two manipulations. Specifically, we compared trials with the highest level of stimulus reliability and the lowest level of boundary distance (High reliability/Low distance) against trials with the lowest level of stimulus reliability and the highest level of boundary distance (Low reliability/High distance). This comparison revealed a robust confidence-accuracy dissociation between these pairs of conditions (Fig. 2D). Specifically, we found that the Low reliability/High distance trials had a significantly higher d' ($t(54) = 7.98$, $p = 1.07 \times 10^{-10}$, Cohen's $d = 1.08$) but lower confidence ($t(54) = -6.06$, $p = 1.38 \times 10^{-7}$, Cohen's $d = -0.82$) than the High reliability/

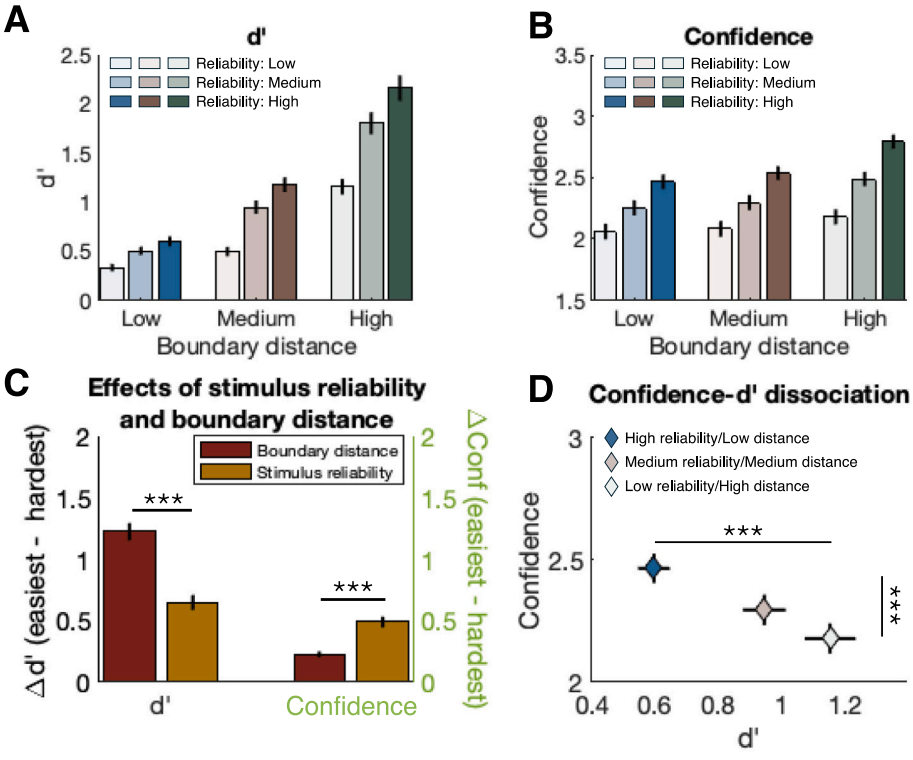

Fig. 2. Stimulus reliability and boundary distance manipulations have dissociable effects on d' and confidence in Experiment 1. (A) Task sensitivity (*d'*) for each experimental condition. Increases in both stimulus reliability and boundary distance are associated with increased *d'*. (B) Confidence for each experimental condition. Increases in both stimulus reliability and boundary distance are associated with increased confidence. (C) The strength of the effect of stimulus reliability and boundary distance on *d'* and confidence. The y-axis shows the change of *d'* and confidence for each manipulation from the hardest to the easiest condition. We found that boundary distance manipulation affects task sensitivity (*d'*) more than stimulus reliability manipulation, whereas the opposite is true for confidence. (D) Comparisons between condition that trade off the two manipulations. The Low reliability/High distance condition had the highest *d'* followed by the Medium reliability/Medium distance and High reliabiltiy/Low distance conditions. However, the ordering was opposite for confidence. Error bars show SEM. ***, $p < 0.001$.

Low distance trials. These extreme pair comparisons corroborated the previous results by confirming that, compared to boundary distance, stimulus reliability has a larger effect on confidence but a smaller impact on task sensitivity.

3.1.2. Experiment 2: motion direction discrimination

Experiment 1 demonstrated a robust confidence-accuracy dissociation between stimulus reliability and boundary distance manipulations in the context of Gabor orientation discrimination. Experiment 2 tested whether similar results would be obtained in the context of a different perceptual task – motion direction discrimination. As in Experiment 1, we first confirmed that both stimulus reliability and boundary distance manipulations affected task performance and confidence (Fig. 3A,B). Indeed, we found that both *d'* and confidence were significantly modulated by both stimulus reliability (d': $F(2,32) = 13.79$, $p = 1.05 \times 10^{-5}$; confidence: $F(2, 32) = 8.58$, $p = 5.00 \times 10^{-4}$) and boundary distance (d': $F(2, 32) = 36.16$, $p = 3.10 \times 10^{-11}$; confidence: $F(2, 32) = 8.96$, $p = 3.71 \times 10^{-4}$). Specifically, d' increased by 0.63 from low to high stimulus reliability ($t(16) = 4.18$, $p = 7.15 \times 10^{-4}$) and by 0.87 from low to high boundary distance ($t(16) = 6.68$, $p = 5.29 \times 10^{-6}$), whereas confidence increased by 0.55 from low to high stimulus reliability ($t(16) = 3.11$, $p = 6.67 \times 10^{-3}$) and by 0.15 from low to high boundary distance ($t(16) = 3.54$, $p = 2.73 \times 10^{-3}$).

Critically, we again compared the effects of stimulus reliability and boundary distance on *d'* and confidence (Fig. 3C). Similar to Experiment 1, compared to boundary distance, stimulus reliability had a smaller effect on d' ($t(16) = 2.87$, $p = 0.01$, Cohen's d = 0.70) but a larger effect on confidence ($t(16) = 2.53$, $p = 0.02$, Cohen's d = 0.61). Lastly, we again examined the differences between pairs of conditions that traded off the two manipulations. As in Experiment 1, we found that trials with the lowest level of stimulus reliability and the highest level of boundary distance (Low reliability/High distance) had a significantly higher d' ($t(54) = 7.98$, $p = 1.07 \times 10^{-10}$, Cohen's d = 1.08) but lower confidence ($t(54) = -6.06$, $p = 1.38 \times 10^{-7}$, Cohen's d = −0.82) than trials with the highest level of stimulus reliability and the lowest level of boundary distance (High reliability/Low distance). These findings confirm the conclusion from Experiment 1 that stimulus reliability and boundary distance differentially impact *d'* and confidence.

3.2. Confidence in correct vs. error trials (folded-X vs. double-increase patterns)

Having replicated previous literature by demonstrating a robust confidence-accuracy dissociation between stimulus reliability and boundary distance, we turned to our central question. Specifically, we examined the effects of the two manipulations on confidence for correct vs. error trials and tested the hypothesis that manipulating boundary distance would produce the folded-X pattern (where easier conditions lead to increased confidence for correct trials but decreased confidence for error trials), whereas manipulating stimulus reliability would produce the double-increase pattern (where easier conditions lead to increased confidence for both correct and error trials). For each subject, we fit linear regressions separately for correct and error trials to obtain individual beta coefficients, which were then averaged across subjects to yield average beta values.

The results supported this hypothesis. Indeed, we found that the boundary distance manipulation produced the folded-X pattern in both experiments (Fig. 4). Specifically, larger boundary distances (which make the task easier) led to a confidence increase for correct trials in both experiments (Experiment 1: average $\beta = 0.14$, $t(54) = 11.56$, $p = 3.1 \times 10^{-16}$, Cohen's d = 1.56; Experiment 2: average $\beta = 0.09$, $t(16) = 4.40$, $p = 4.4 \times 10^{-4}$, Cohen's d = 1.07). Critically, larger boundary

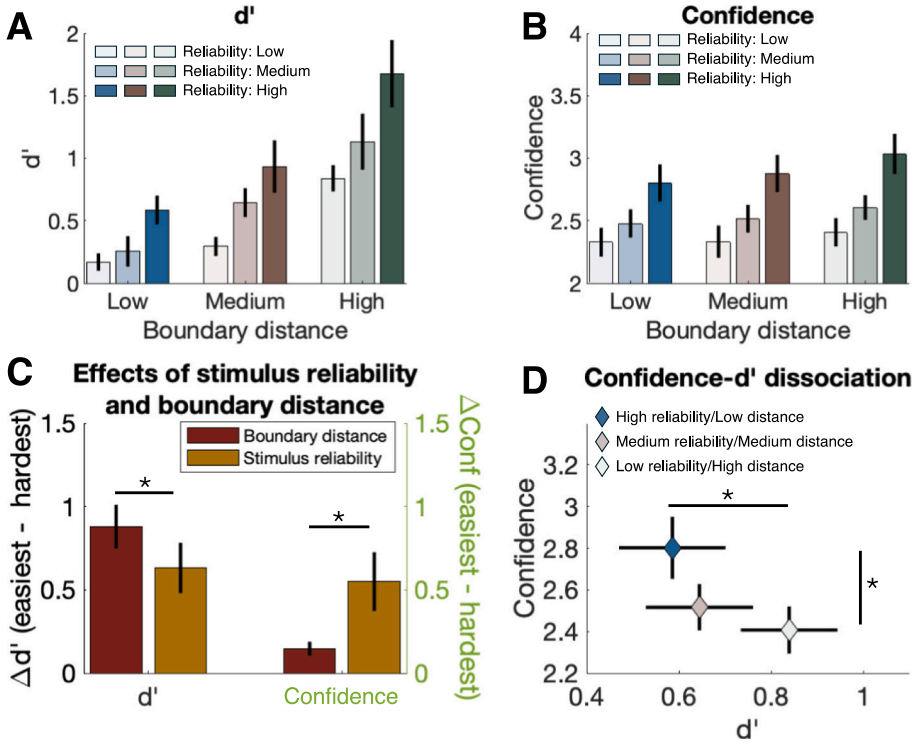

Fig. 3. Stimulus reliability and boundary distance manipulations have dissociable effects on d' and confidence in Experiment 2. (A) Task sensitivity (*d'*) for each experimental condition. Increases in both stimulus reliability and boundary distance are associated with increased *d'*. (B) Confidence for each experimental condition. Increases in both stimulus reliability and boundary distance are associated with increased confidence. (C) The strength of the effect of stimulus reliability and boundary distance on *d'* and confidence. The y-axis shows the change of *d'* and confidence for each manipulation from the hardest to the easiest condition. We found that boundary distance manipulation affects task sensitivity (*d'*) more than the stimulus reliability manipulation, whereas the opposite is true for confidence. (D) Comparisons between conditions that trade off the two manipulations. The Low reliability/High distance trials had the highest *d'* followed by the Medium reliability/Medium distance and High reliabiltiy/Low distance conditions. However, the ordering was opposite for confidence. Error bars show SEM. *, $p < 0.05$.

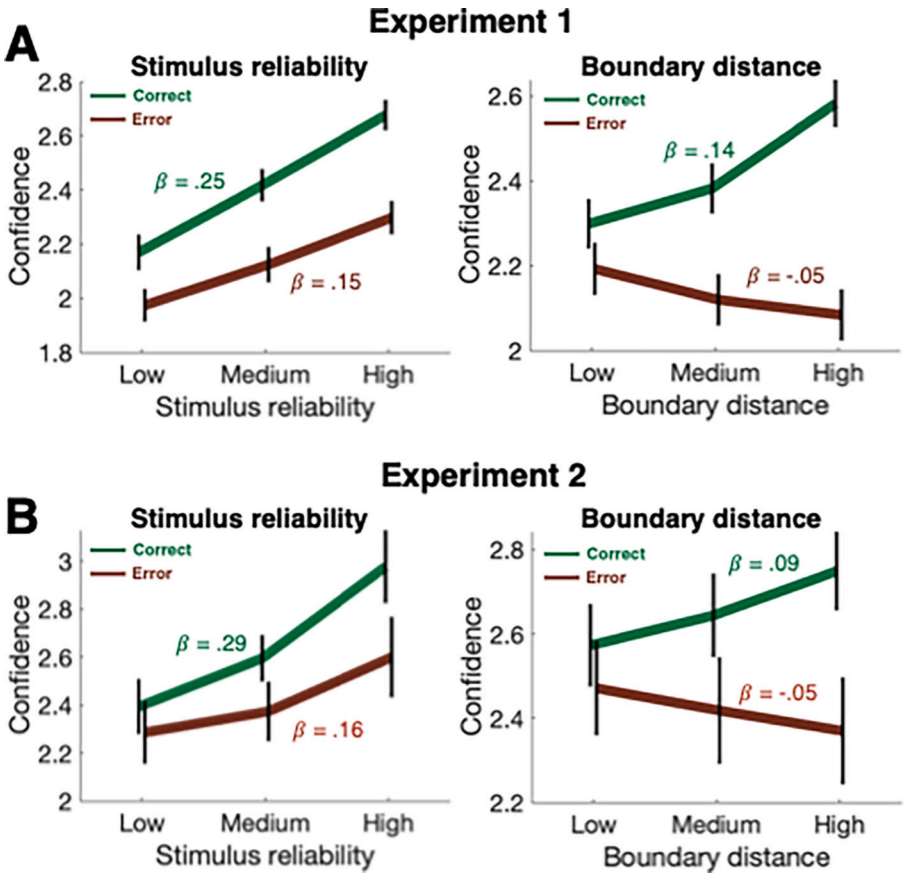

Fig. 4. Boundary distance manipulations produce the folded-X pattern, whereas stimulus reliability manipualtions produce the double-increase pattern. We found that the boundary distance manipulation produced the folded-X pattern in both experiments, such that easier conditions led to increased confidence for correct trials but decreased confidence in error trials. In contrast, the stimulus reliability manipulation produced the double-increase pattern, where easier conditions led to increased confidence for both correct and error trials. Error bars show SEM. β values show the average slope.

distances led to a confidence decrease for error trials in Experiment 1 (average $\beta = -0.05$, $t(54) = -3.50$, $p = 9.5 \times 10^{-4}$, Cohen's $d = -0.47$) and a trend towards a confidence decrease for error trials in Experiment 2 (average $\beta = -0.05$, $t(16) = -2.05$, $p = 0.057$, Cohen's $d = -0.79$).

In contrast, the stimulus reliability produced the double-increase pattern in both experiments (Fig. 4). Specifically, larger stimulus reliability (which make the task easier) led to higher confidence for correct trials in both experiments (Experiment 1: average $\beta = 0.25$, $t(54) = 10.78$, $p = 4.5 \times 10^{-15}$, Cohen's $d = 1.45$; Experiment 2: average $\beta = 0.29$, $t(16) = 3.28$, $p = 4.7 \times 10^{-3}$, Cohen's $d = 0.79$). On the other hand, larger stimulus reliability led to a confidence increase for error trials in Experiment 1 (average $\beta = 0.15$, $t(54) = 7.29$, $p = 1.4 \times 10^{-9}$, Cohen's $d = 0.98$) and a trend towards a confidence increase for error trials in Experiment 2 (average $\beta = 0.16$, $t(16) = 1.89$, $p = 0.077$, Cohen's $d = 0.46$). These results indicate that the stimulus reliability and boundary distance manipulations produce qualitatively different effects, with boundary distance leading to the folded-X pattern, whereas stimulus reliability leading to the double-increase pattern. These patterns remained unchanged when including response time as a covariate in the analysis (see Supplementary Results and Supplementary Fig. 1) or when using a unified linear mixed-effects model that simultaneously tests both manipulations and their interactions (Supplementary Results).

Given that both experiments featured a 3×3 design, the analyses above focused on each manipulation always simply ignored the other manipulation (i.e., they used all trials regardless of the level of the alternative manipulation). To further examine the robustness of our results, we additionally repeated the analyses by restricting the alternative manipulation to just one level at a time and report Bonferroni-corrected p-values. We found that these additional analyses led to a qualitatively similar pattern of results (Fig. 5). Specifically, manipulating stimulus reliability produced the double-increase pattern when boundary distance was set at low, medium, or high levels in Experiment 1, with all effects significant for both correct trials ($\beta = 0.21$, 0.25, 0.31, respectively; all p's $< 1.10 \times 10^{-11}$) and error trials ($\beta = 0.19$, 0.13, 0.16, respectively, all p's $< 1.92 \times 10^{-4}$). In Experiment 2, the pattern was similar for correct trials ($\beta = 0.26$, 0.28, 0.33, respectively; all p's < 0.011), though the effect for error trials did not reach significance for any of the three coherence levels ($\beta = 0.19$, $p = 0.102$; $\beta = 0.18$, $p = 0.144$; and $\beta = 0.10$, $p = 0.90$, respectively).

In contrast, manipulating boundary distance produced results consistent with the folded-X pattern regardless of difficulty level. Specifically, in Experiment 1, fixing stimulus reliability to low, medium, or high levels always produced significant positive slopes for correct trials ($\beta = 0.09$, 0.14, 0.19, respectively; all p's $< 1.97 \times 10^{-9}$). We observed significant negative slopes for error trials at low and medium stimulus reliability levels ($\beta = -0.041$ and -0.057; both p's < 0.019), though this effect was not significant at high stimulus reliability ($\beta = -0.065$, $p = 0.192$). In Experiment 2, correct trials showed positive slopes that were significant at low and high stimulus reliability levels ($\beta = 0.064$ and 0.13, both p's < 0.021), but not at medium stimulus reliability ($\beta = 0.070$, $p = 0.066$), while none of the error trial slopes reached significance ($\beta = -0.015$, -0.037, -0.10, respectively; all p's > 0.05). These results demonstrate that the differential effects of stimulus reliability and boundary distance on confidence for correct vs. error trials are robust across different difficulty levels of the alternative manipulation, with Experiment 1 showing stronger and more consistent effects than Experiment 2, likely due to larer sample size.

Finally, we confirmed that these patterns remain stable over time by

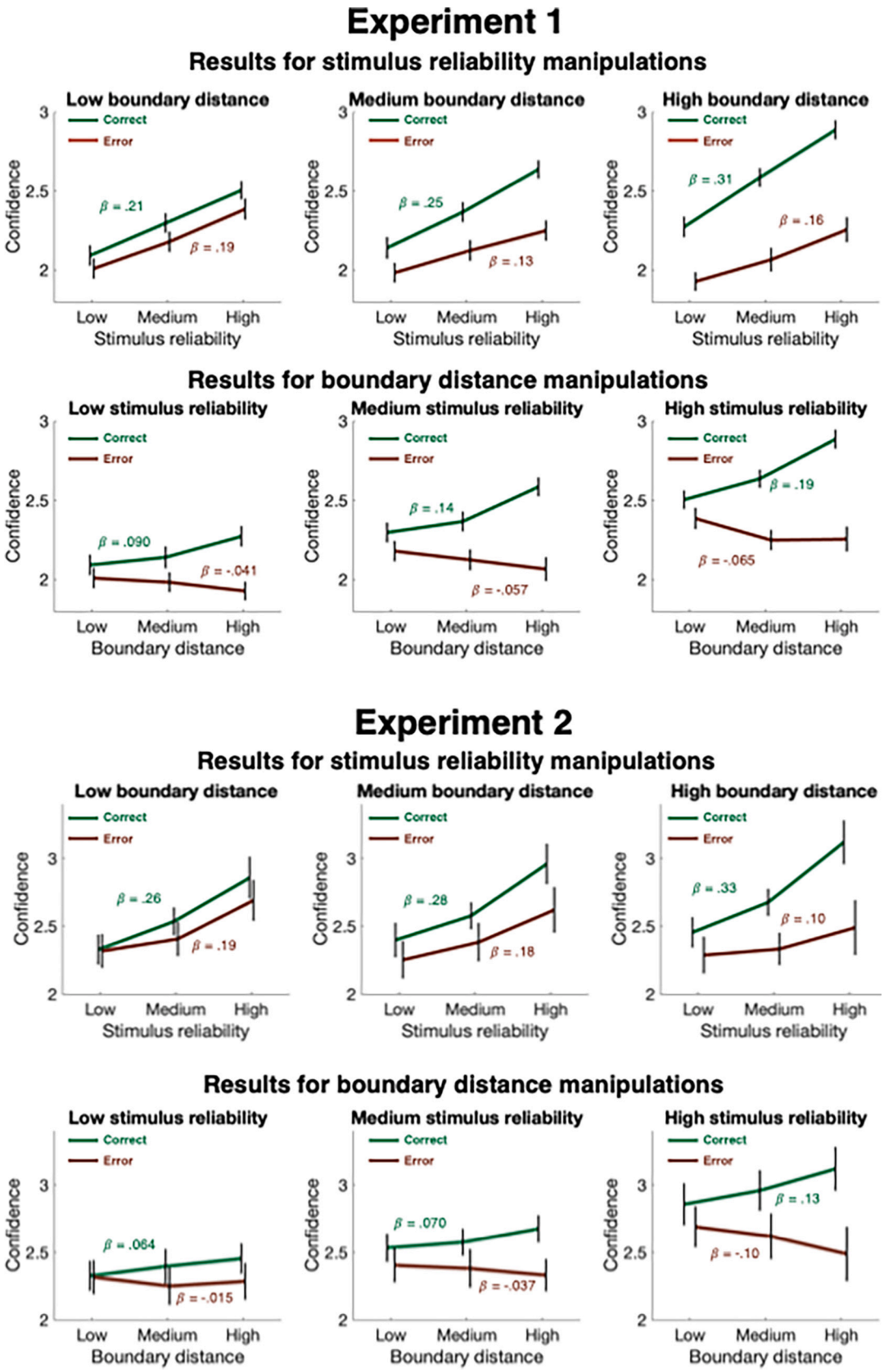

Fig. 5. Folded-X and double-increase pattern remain stable for different levels of the alternative manipulation. We examined the pattern of confidence for correct and error trials separately for each level of the alternative manipulation. For the stimulus reliability manipulation, we observed a consistent double-increase when boundary distance was held constant at low, medium, or high levels across both experiments. For the boundary distance manipulation, we observed the folded-X pattern when stimulus reliability was held constant at low, medium, or high levels across both experiments. Error bars show SEM. β values show the average slope.

conducting the analyses separately on the first and second halves of each experiment (Supplementary Figs. 2 and 3). Across all boundary distance levels, we found no significant differences in the effect of stimulus reliability on confidence between the first and second halves for either correct or error trials in both experiments (all corrected p's > 0.35). Similarly, across all stimulus reliability levels, we found no significant differences in the effect of boundary distance on confidence between the two halves (all corrected p's > 0.063). Importantly, Bayesian analyses provided moderate to strong evidence in favor of the null hypothesis (no difference between halves) for eleven out of twelve comparisons (BF_{01} range: 3.48–18.74), confirming that the absence of significant differences reflects a a general stability rather than insufficient power. These results demonstrate that the differential effects of stimulus reliability and boundary distance on the folded-X vs. double-increase patterns are not only robust across different difficulty levels but also robust to learning during the course of the experiment.

3.3. ANNs show folded-X pattern for both manipulations

The analyses above established a human dissociation in confidence patterns across the two manipulations: stimulus reliability produces a double-increase pattern, whereas boundary distance produces a folded-X pattern. To examine whether the human dissociation between stimulus reliability and boundary distance manipulations could arise purely from the statistical structure of the stimuli and task – rather than from additional metacognitive processes – we analyzed ANNs trained on Gabor stimuli as in Experiment 1. Specifically, we used a simple 4-layer architecture and trained 30 ANN instances that only differed in their random initializations.

We found that the ANNs replicated the basic performance pattern observed in humans: both higher stimulus reliability and larger boundary distance increased accuracy and confidence (Fig. 6A). This confirms that the networks were sensitive to the same task manipulations as humans. Further, the ANNs also reproduced the dissociable relationship between stimulus reliability and boundary dinstance on confidence and accuracy. Specifically, compared to boundary distance, stimulus reliability manipulations had a larger effect on confidence ($t(29) = 11.35$, $p = 3.45 \times 10^{-12}$) but a smaller effect on accuracy ($t(29) = -12.75$, $p = 2.06 \times 10^{-13}$; Fig. 6B).

Critically, however, the pattern of confidence for correct versus error trials in the ANN instances diverged from human behavior (Supplementary Fig. 4). Unlike humans, where stimulus reliability and boundary distance produced double-increase and folded-X patterns, respectively (Fig. 4), ANNs showed the folded-X pattern for both manipulations, such that easier conditions increased confidence for correct trials but decreased confidence for error trials (Fig. 6C). Indeed, larger stimulus reliability led to confidence increase for correct trials (average $\beta = 0.33$, with 30/30 ANN instances showing positive β) and confidence decrease for error trials (average $\beta = -0.037$, with 29/30 ANN instances showing negative β). A similar pattern was observed for boundary distance as well: larger boundary distance led to confidence increase for correct trials (average $\beta = 0.24$, with 30/30 ANN instances showing positive β) and confidence decrease for error trials (average $\beta = -0.10$, with ANN 30/30 instances showing negative β). This difference between humans and ANNs shows that the human double-increase and folded-X patterns cannot be explained solely by the statistical structure of the stimuli and task.

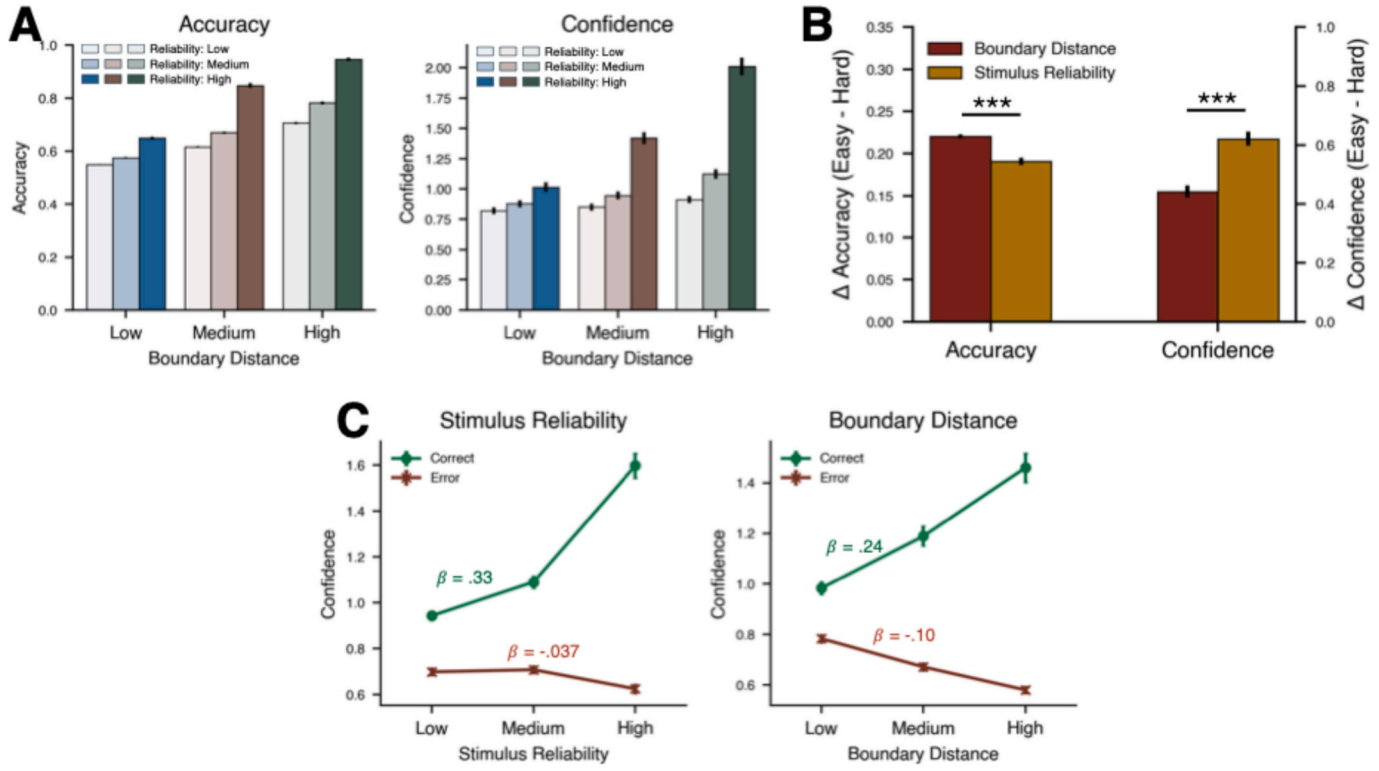

Fig. 6. ANN simulations of the Gabor orientation discrimination task. (A) Mean accuracy (left) and confidence (right) of 30 independently trained artificial neural networks (ANNs) as a function of boundary distance and stimulus reliability. Bars show the mean across ANN instances and error bars indicate SEM. (B) The strength of the effect of stimulus reliability and boundary distance on *d'* and confidence. The y-axis shows the change of *d'* and confidence for each manipulation from the hardest to the easiest condition. We found results equivalent to the human data (Fig. 2C): the boundary distance manipulation affects task sensitivity (*d'*) more than the stimulus reliability manipulation, whereas the opposite is true for confidence. ***, $p < 0.001$. (C) Mean confidence for correct (green) and error (red) responses as a function of stimulus reliability (left) and boundary distance (right). For both manipulations, the ANNs exhibit a folded-X pattern, such that easier conditions led to increased confidence for correct trials but decreased confidence in error trials. Error bars indicate SEM. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. Reaction time patterns for correct vs. error trials

Having established distinct patterns in confidence for the two manipulations, we next examined whether reaction time (RT) exhibits similar patterns. Previous studies have reported that RT often shows patterns similar to confidence (Kiani et al., 2014; Zylberberg, Fetsch, & Shadlen, 2016) and thus RT is sometimes used as a proxy for confidence (Kiani et al., 2014; Miyoshi, Rahnev, & Lau, 2025; Moran et al., 2015; Ratcliff & Starns, 2009; Volkmann., 1934). We therefore examined whether RT would follow a pattern of results similar to confidence (though, given the negative relationship between RT and confidence, we would expect the patterns to be mirror images to the confidence patterns).

We found that RT patterns showed both similarities and differences compared to confidence (Fig. 7). Stimulus reliability exhibited the double-increase pattern for confidence (Fig. 4) and could thus be expected to show a "double-decrease" pattern for RT where RT in easier conditions decreases for both correct and error trials. Experiment 1 violated this prediction: increasing stimulus reliability had no significant effect on RT for correct trials ($\beta = 0.003$, $t(54) = 0.44$, $p = 0.665$, Cohen's d = 0.06) and led to significantly increased (instead of decreased) RT for error trials ($\beta = 0.042$, $t(54) = 4.35$, $p = 6.1 \times 10^{-5}$, Cohen's d = 0.59). Experiment 2 showed a quantitative decrease for both correct and error trials but neither effect was significant (correct trials: $\beta = -0.031$, $t(16) = -0.93$, $p = 0.367$, Cohen's d = −0.22; error trials: $\beta = -0.009$, $t(16) = -0.21$, $p = 0.837$, Cohen's d = −0.05). Conversely, boundary distance exhibited the folded-X pattern for confidence (Fig. 4) and could thus be expected to also show a folded-X pattern for RT (except that easier conditions should decrease RT for correct trials and increase RT for error trials). This prediction was qualitatively confirmed, though some effects did not reach significance. Specifically, in Experiment 1, we observed a significant decrease in RT for correct trials ($\beta = -0.035$, $t(54) = -6.79$, $p = 9.0 \times 10^{-9}$, Cohen's d = −0.92) but a non-significant increase for error trials ($\beta = 0.013$, $t(54) = 1.43$, $p = 0.157$, Cohen's d = 0.19). Experiment 2 had the opposite pattern of significance, with a non-significant decrease in RT for correct trials ($\beta = -0.013$, $t(16) = -1.34$, $p = 0.199$, Cohen's d = −0.32) but a significant increase for error trials ($\beta = 0.034$, $t(16) = 2.51$, $p = 0.023$, Cohen's d = 0.61). Overall, the RT patterns followed the confidence patterns well for the boundary distance manipulation (both resulted in folded-X patterns), but diverged more strongly for the stimulus reliability manipulation. We also conducted these analyses using reciprocal RT (1/RT) to account for potential non-normality in RT distributions. These analyses showed qualitatively similar patterns to the raw RT analysis (see Supplementary Results).

3.5. Study-level confidence-accuracy dissociation

Our analyses so far compared two types of manipulations (stimulus reliability vs. boundary distance) within the context of specific tasks (Gabor orientation discrimination in Experiment 1 and motion direction discrimination in Experiment 2). Here, we further explored whether confidence-accuracy dissociations may appear when comparing across experiments that employ different stimuli and tasks. Specifically, we compared the average *d'* and average confidence between Experiments 1 and 2. We found that the task in Experiment 1 produced significantly higher average *d'* than in Experiment 2 ($t(70) = 2.37$, $p = 0.021$, Cohen's $d = 0.66$; Fig. 8). Surprisingly, however, confidence showed the opposite pattern, with Experiment 1 leading to significantly lower average confidence than Experiment 2 ($t(70) = -2.17$, $p = 0.034$, Cohen's $d = -0.60$). These results suggest that different stimuli and tasks may also lead to previously unexplored confidence-accuracy dissociations.

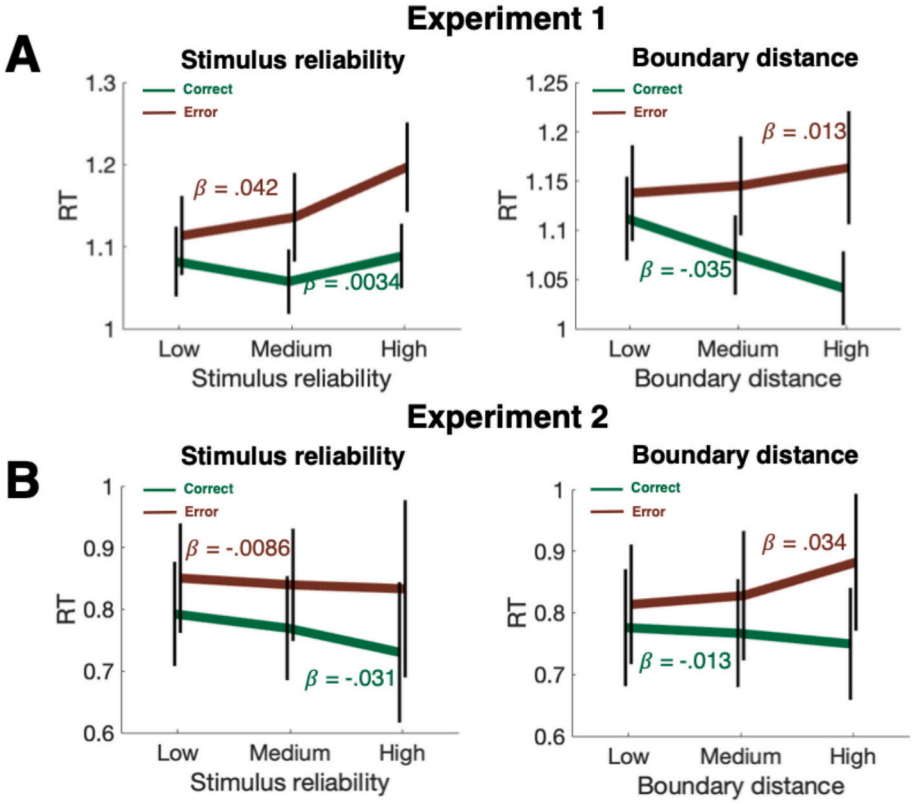

Fig. 7. Response time (RT) patterns for correct and error trials. RT exhibited both similarities and differences compared to confidence. While confidence showed a consistent double-increase pattern for stimulus reliability and a folded-X pattern for boundary distance, RT exhibited more variable patterns. For stimulus reliability (left column), RT either increased for error trials only (Experiment 1) or showed no significant change (Experiment 2). For boundary distance (right column), RT generally decreased for correct trials and increased for error trials, though not all effects were significant. Error bars show SEM. β values show the average slope.

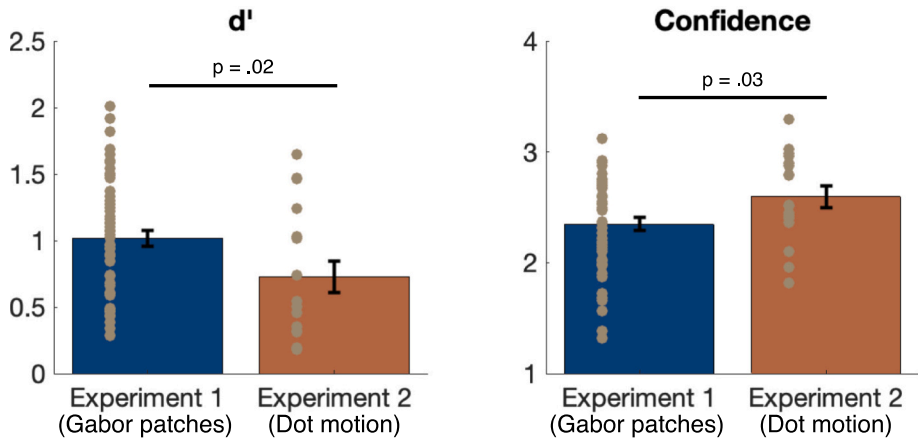

Fig. 8. Across-study confidence-accuracy dissociation. Comparison of average *d'* and confidence ratings between Experiments 1 and 2. The Gabor orientation discrimination task in Experiment 1 produced significantly higher *d'* values, but paradoxically led to significantly lower average confidence ratings compared to the motion direction discrimination task in Experiment 2. Error bars show SEM.

4. Discussion

The folded-X pattern, characterized by increased confidence for correct trials and decreased confidence for error trials under easier conditions, has long been considered a fundamental signature of decision confidence (Kepecs & Mainen, 2012). However, recent findings have identified systematic violations of this pattern, with some manipulations producing a double-increase pattern where easier conditions lead to increased confidence for both correct and error trials (Adler & Ma, 2018a; Kiani et al., 2014; Rausch et al., 2018; Van Den Berg et al., 2016). A critical question that has remained unanswered is what determines whether a particular category of stimulus manipulation will produce the folded-X versus the double-increase pattern. Here, we test the hypothesis that the type of difficulty manipulation – specifically whether it affects stimulus reliability or boundary distance – determines the resulting confidence pattern.

Our results provide strong support for this hypothesis. Across two experiments using different perceptual tasks (Gabor orientation discrimination and dot motion direction discrimination), we consistently found that boundary distance manipulations produce the folded-X pattern, while stimulus reliability manipulations result in the double-increase pattern. This dissociation was robust across different levels of the alternative manipulation and remained stable throughout the experiments. These results suggest that violations of the folded-X pattern are not random occurrences but follow systematic principles based on the nature of the difficulty manipulation. As most experimental manipulations in perceptual decision-making can be classified as affecting either stimulus reliability or boundary distance, our results provide a framework to predict a priori which confidence pattern a given manipulation will produce, which is crucial for designing experiments and interpreting neural correlates of confidence.

4.1. Mechanisms underlying the folded-X versus double-increase patterns

We consider three potential explanations for the divergent confidence patterns in correct vs. error trials produced by stimulus reliability versus boundary distance manipulations. First, the different patterns may simply reflect the nature of the task and stimuli – that is, the way these manipulations affect internal representations naturally leads to different patterns of results. Second, the folded-X pattern may emerge from post-decision evidence accumulation, which allows the system to detect errors through a "change-of-mind" mechanism. Third, the double-increase pattern may appear specifically when subjects are able to implicitly form predictions about trial-by-trial task difficulty based on salient perceptual features. Our data do not support the first two explanations but provide partial support for the third, though with important caveats suggesting that additional factors may be at play.

The first possible explanation is that the dissociation between stimulus reliability and boundary distance manipulations arises purely from the statistical structure of the stimuli and task. To test this hypothesis, we used artificial neural networks (ANNs) as "model organisms" (Cichy & Kaiser, 2019; Scholte, 2018) – simplified systems that can reveal whether observed patterns emerge from basic computational principles or require additional mechanisms. When we trained ANNs on the same Gabor orientation task and examined their confidence patterns, we found that both stimulus reliability and boundary distance produced folded-X pattern rather than the double-increase pattern observed for reliability manipulation in human. In other words, a standard feedforward classifier that has access to the same sensory input and is optimized solely for accuracy does not spontaneously produce higher confidence for errors in easier, high-reliability conditions. This divergence between human and ANN behavior suggests that the dissociation between stimulus reliability and boundary distance manipulations is unlikely to arise from the statistical properties of the task alone and instead reflects additional metacognitive mechanisms not present in standard neural networks.

The second possible explanation is that the folded-X pattern emerges from post-decision evidence accumulation, where evidence continues to accumulate after the initial decision, allowing the system to detect errors through a "change-of-mind" mechanism (Desender, Donner, & Verguts, 2021; Fetsch, Kiani, Newsome, & Shadlen, 2014). However, several recent studies suggest that confidence in perceptual tasks is not typically based on post-decisional information (Chen & Rahnev, 2023; Green & Rahnev, 2025; Vivar-Lazo & Fetsch, 2025; Xue, Zheng, Rafiei, & Rahnev, 2023). Further, our results directly demonstrate that such post-decision evidence accumulation cannot be the main reason for the folded-X pattern. Specifically, we observed systematically different confidence signatures – double-increase patterns under stimulus reliability manipulations and folded-X patterns under boundary distance manipulations – within the same set of trials. If the presence or absence of post-decisional evidence accumulation is the sole driver of the folded-X vs. the double-increase pattern, then analyzing the same set of trials in two different ways – as done here – should not lead to two different patterns. Thus, while paradigms that encourage post-decisional evidence accumulation (e.g., continuing to present new sensory evidence after the decision) may indeed result in stronger folded-X patterns, our results cannot be explained by post-decisional evidence accumulation.

The third possible explanation is that confidence patterns depend on whether subjects are able to implicitly form predictions about the trial-by-trial task difficulty (Hellmann et al., 2023; Rausch et al., 2018;

Rausch, Hellmann, & Zehetleitner, 2021; Rausch & Zehetleitner, 2019). According to the Weighted Evidence and Visibility (WEV) model, confidence integrates decision evidence and stimulus visibility (Hellmann et al., 2023; Rausch, Zehetleitner, Steinhauser, & Maier, 2020, 2021). Under this framework, the decision evidence and stimulus visibility signals are differentially affected by stimulus reliability and boundary distance manipulations. Specifically, salient stimulus features like contrast or coherence allow observers to predict task difficulty, essentially tagging trials as having "high-quality" or "low-quality" evidence. High reliability trials are tagged as high quality, leading to elevated confidence even for the errors, thus producing the double-increase pattern. In contrast, when all stimuli appear equally clear – as with boundary distance manipulations – confidence relies primarily on decision-specific evidence strength. In this case, easy trials lead to very strong evidence on correct trials but very weak evidence on error trials (even weaker, on average, than the evidence on error trials for difficult stimuli; Sanders et al., 2016). To further examine this hypothesis, we reanalyzed data from our previous work where we systematically varied subjects' ability to predict trial difficulty (Xue, Shekhar, & Rahnev, 2024). While explicit difficulty information eliminated the folded-X pattern, the results did not show a clear flip to the double-increase pattern (Supplementary Fig. 5). These findings suggest that while visibility-based accounts capture important aspects of the underlying mechanism, factors intrinsic to the type of manipulation may also play a critical role in determining confidence patterns.

4.2. RT as a proxy for confidence

We examined whether RT exhibits patterns similar to confidence, as previous studies have suggested that RT can serve as a proxy for confidence (Kiani et al., 2014; Miyoshi et al., 2025; Moran et al., 2015; Ratcliff & Starns, 2009; Volkmann., 1934). We found that RT patterns showed both similarities and differences compared to confidence. While RT generally mirrored the folded-X pattern for boundary distance manipulations, it showed more variable patterns for stimulus reliability manipulations across experiments. Most notably, for stimulus reliability in Experiment 1, RT increased only for error trials rather than showing the expected double-decrease pattern that would mirror the confidence double-increase pattern. These findings highlight both the utility and limitations of using RT as a proxy for confidence, demonstrating that while RT can sometimes provide insights into confidence-related processes, it does not always reliably reflect the same patterns observed in explicit confidence judgments.

4.3. Relationship to previous research on comparing stimulus reliability and boundary distance

Our finding that reliability manipulations produce a double-increase pattern, while boundary distance manipulations produce a folded-X pattern is consistent with the majority of the literature. Indeed, studies using reliability manipulations have generally reported double-increase patterns (Adler & Ma, 2018b; Kiani et al., 2014; Rausch et al., 2018; Van Den Berg et al., 2016), whereas studies using boundary distance manipulations have typically reported folded-X patterns (Kepecs et al., 2008; Rausch & Zehetleitner, 2019; Sanders et al., 2016). One exception to this general pattern is the study by Shekhar and Rahnev (2021) where a contrast manipulation (an example of a stimulus reliability manipulation) resulted in a folded-X pattern, as reported by Hellmann, Zehetleitner, and Rausch (2024). Critically, the Shekhar & Rahnev study used a unique design where Gabor patches were embedded in pixel noise in such a way that the overall stimulus contrast (the difference between the brightest and darkest pixels) was equated across conditions. This manipulation may have made it difficult for observers to quickly assess stimulus quality based on global perceptual features, as the overall image contrast remained constant despite changes in signal-to-noise ratio. In our framework, this represents a reliability manipulation for which stimulus quality information was not readily accessible to the metacognitive system – consistent with the WEV-based account that violations of the folded-X pattern depend on whether observers can use perceptual features to implicitly assess task difficulty. Future work systematically examining how the perceptual accessibility of difficulty cues affects confidence patterns would help further refine predictions about when reliability manipulations may produce folded-X versus double-increase patterns.

Beyond establishing these distinct confidence patterns for correct versus error trials, our results on the differential effects of stimulus reliability and boundary distance on accuracy vs. confidence (Figs. 2 and 3) help clarify an important debate in the literature. Specifically, our findings that, compared to boundary distance, stimulus reliability manipulations impact confidence more than accuracy are in line with Spence et al. (2016) and Boldt et al. (2017). However, our results diverge from those in two previous reports (de Gardelle & Mamassian, 2015; Zylberberg et al., 2014). De Gardelle & Mamassian (2015) did not find robust group differences in the effects of stimulus reliability and boundary distance on confidence, but instead found individual differences, such that each subject consistently showed a higher confidence in either high stimulus reliability or high boundary distance across two experiment sessions. We suspect that the differences in our results may be attributed to differences in study design. For example, their sample size (15 participants) is relatively small for detecting robust group differences. Further, de Gardelle & Mamassian used a more complex display consisting of two stimuli on each trial and a more complex confidence procedure where subjects judged which of every consecutive two trials they were more confident in. It is possible that the added complexity in their design coupled with a relatively small effect size reduced the power for detecting differences between the manipulations to emerge. Unlike de Gardelle & Mamassian who did not find group differences between stimulus reliability and boundary distance manipulations, Zylberberg et al. (2014) found results that on the surface appear opposite to ours. Specifically, Zylberberg et al. reported that lower stimulus reliability led to higher levels of confidence (though they did not directly compare stimulus reliability and boundary distance). However, Zylberberg et al. used a very different design where the boundary distance was very close to zero, thus making the task extremely difficult. In such cases, adding noise to the images (which usually reduced reliability) can lead to spurious signals that increase confidence. Specifically, as has been argued before (Sanders et al., 2016), overconfidence for noisy stimuli may occur because observers mistake noise as useful signals.

4.4. Possible other manipulation categories

We have argued that manipulations of task difficulty in perceptual decision making can generally be categorized as manipulations of stimulus reliability or boundary distance. The critical distinction between these two manipulations is that decreasing stimulus reliability increases the uncertainty of the stimulus itself, while decreasing boundary distance leaves stimulus uncertainty unaffected. One way to make this distinction concrete is to consider the effects of these manipulations on the 2-choice task of interest vs. a separate continuous identification task. In this context, a stimulus reliability manipulation is any manipulation that affects the difficulty of both the 2-choice and continuous identification tasks. Conversely, a boundary distance manipulation is any manipulation that affects the difficulty of the 2-choice task, while leaving the difficulty of the continuous identification task unchanged (because decisions in this task are not made by comparing to a boundary). Indeed, manipulations of contrast, noise, or motion coherence affect both 2-choice and continuous identification tasks. In contrast, manipulations of tilt offset or distance of a criterion in color space affect the difficulty of 2-choice tasks but do not change how well subjects can identify the absolute tilt or absolute color of the stimulus.

Based on the definition above, it appears that virtually any difficulty manipulation in purely perceptual tasks should fall under one of these categories. That being said, sometimes researchers use a combination of manipulations. For example, several papers have explored the effects of energy manipulations where both the signal and the noise of a stimulus have increased (Gao, Xue, Odegaard, & Rahnev, 2025; Koizumi, Maniscalco, & Lau, 2015; Samaha & Denison, 2022). Energy manipulations simultaneously involve two distinct types of changes: increasing signal (which increases reliability) while simultaneously increasing noise (which decreases stimulus reliability). Because these two reliability changes push confidence in opposite directions, the net effect on confidence patterns becomes unpredictable and depends on the relative weighting of these opposing forces. This dual nature makes energy manipulations fundamentally different from the pure stimulus reliability manipulations examined in our study, where only one dimension varies at a time, and makes them difficult to classify cleanly as either pure stimulus reliability or pure boundary distance manipulations. In other words, the distinction between stimulus reliability vs. boundary distance manipulations is meant to apply only to changes in a single dimension at a time, whereas complex manipulations that have two or more dimensions would not necessarily fall under either category.

5. Conclusion

In conclusion, our study demonstrates that the type of difficulty manipulation determines whether confidence will exhibit the folded-X or double-increase pattern. Specifically, we found that boundary distance manipulations consistently produce the folded-X pattern, while stimulus reliability manipulations result in the double-increase pattern. Our results provide researchers with a predictive framework for designing experiments and interpreting confidence-related neural signals. Additionally, we demonstrate that while RT patterns largely mirror confidence patterns, they show notable exceptions that highlight the limitations of using RT as a proxy for confidence. These findings advance our understanding of how different experimental manipulations shape the relationship between confidence and accuracy in perceptual decision-making.

CRediT authorship contribution statement

Kai Xue: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Herrick Fung:** Writing – review & editing, Visualization, Investigation. **Dobromir Rahnev:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Acknowledgments

This work was supported by the National Institute of Health (award: R01MH119189) and the Office of Naval Research (award: N00014-20-1-2622).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2026.106490.

Data availability

All data and codes are available at https://osf.io/nu4b7/.

References

Adler, W. T., & Ma, W. J. (2018a). Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLoS Computational Biology, 14*(11), Article e1006572. https://doi.org/10.1371/journal.pcbi.1006572

Adler, W. T., & Ma, W. J. (2018b). Limitations of proposed signatures of Bayesian confidence. *Neural Computation, 30*(12), 3327–3354. https://doi.org/10.1162/neco_a_01141

Bae, G.-Y., & Luck, S. J. (2022). Perception of opposite-direction motion in random dot kinematograms. *Visual Cognition, 30*(4), 289–303. https://doi.org/10.1080/13506285.2022.2052216

Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences, 115*(23), 6082–6087. https://doi.org/10.1073/pnas.1800795115

Bertana, A., Chetverikov, A., van Bergen, R. S., Ling, S., & Jehee, J. F. M. (2021). Dual strategies in human confidence judgments. *Journal of Vision, 21*(5), 21. https://doi.org/10.1167/jov.21.5.21

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

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Cannon, M. W. (1979). Contrast sensation: A linear function of stimulus contrast. *Vision Research, 19*(9), 1045–1052. https://doi.org/10.1016/0042-6989(79)90230-X

Chen, S., & Rahnev, D. (2023). Confidence response times: Challenging postdecisional models of confidence. *Journal of Vision, 23*(7), 11. https://doi.org/10.1167/jov.23.7.11

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences, 23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science, 29*(5), 761–778. https://doi.org/10.1177/0956797617744771

Desender, K., Donner, T. H., & Verguts, T. (2021). Dynamic expressions of confidence within an evidence accumulation framework. *Cognition, 207*, Article 104522. https://doi.org/10.1016/j.cognition.2020.104522

Fetsch, C. R., Kiani, R., Newsome, W. T., & Shadlen, M. N. (2014). Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron, 84*(1), 239. https://doi.org/10.1016/j.neuron.2014.09.020

Fung, H., Murty, N. A. R., & Rahnev, D. (2025). Human-like individual differences emerge from random weight initializations in neural networks (p. 2025.10.25.684448). *bioRxiv*. https://doi.org/10.1101/2025.10.25.684448

Fung, H., Shekhar, M., Xue, K., Rausch, M., & Rahnev, D. (2025). Similarities and differences in the effects of different stimulus manipulations on accuracy and confidence. *Consciousness and Cognition, 136*, Article 103942. https://doi.org/10.1016/j.concog.2025.103942

Gao, Y., Xue, K., Odegaard, B., & Rahnev, D. (2025). Automatic multisensory integration follows subjective confidence rather than objective performance. *Communications Psychology, 3*(1), 38. https://doi.org/10.1038/s44271-025-00221-w

de Gardelle, V., & Mamassian, P. (2015). Weighting mean and variability during confidence judgments. *PLoS One, 10*(3), Article e0120870. https://doi.org/10.1371/journal.pone.0120870

de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences, 108*(32), 13341–13346. https://doi.org/10.1073/pnas.1104517108

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience, 14*(7), 926–932. https://doi.org/10.1038/nn.2831

Green, M., & Rahnev, D. (2025). Visual confidence is an online process. *Journal of Vision, 25*(9), 2402. https://doi.org/10.1167/jov.25.9.2402

Green, M. L., Hu, M., Denison, R. N., & Rahnev, D. (2026). Using artificial neural networks to relate external sensory features to internal decisional evidence. *Open Mind, 10*, 29–46. https://doi.org/10.1162/OPMI.a.317

Hangya, B., Sanders, J. I., & Kepecs, A. (2016). A mathematical framework for statistical decision confidence. *Neural Computation, 28*(9), 1840–1858. https://doi.org/10.1162/NECO_a_00864

Hellmann, S., Zehetleitner, M., & Rausch, M. (2023). Simultaneous modeling of choice, confidence, and response time in visual perception. *Psychological Review, 130*(6), 1521–1543 (2023-53313-001) https://doi.org/10.1037/rev0000411.

Hellmann, S., Zehetleitner, M., & Rausch, M. (2024). Confidence is influenced by evidence accumulation time in dynamical decision models. *Computational Brain & Behavior, 7*(3), 287–313. https://doi.org/10.1007/s42113-024-00205-9

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 367*(1594), 1322–1337. https://doi.org/10.1098/rstb.2012.0037

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455*(7210), 227–231. https://doi.org/10.1038/nature07200

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron, 84*(6), 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015

Koizumi, A., Maniscalco, B., & Lau, H. (2015). Does perceptual confidence facilitate cognitive control? *Attention, Perception, & Psychophysics, 77*(4), 1295–1306. https://doi.org/10.3758/s13414-015-0843-3

Lesica, N. A., Jin, J., Weng, C., Yeh, C.-I., Butts, D. A., Stanley, G. B., & Alonso, J.-M. (2007). Adaptation to stimulus contrast and correlations during natural visual

stimulation. *Neuron, 55*(3), 479–491. https://doi.org/10.1016/j.neuron.2007.07.013
Locke, S. M., Landy, M. S., & Mamassian, P. (2022). Suprathreshold perceptual decisions constrain models of confidence. *PLoS Computational Biology, 18*(7), Article e1010318. https://doi.org/10.1371/journal.pcbi.1010318
Magliero, A., Bashore, T. R., Coles, M. G. H., & Donchin, E. (1984). On the dependence of P300 latency on stimulus evaluation processes. *Psychophysiology, 21*(2), 171–186. https://doi.org/10.1111/j.1469-8986.1984.tb00201.x
Miyoshi, K., Rahnev, D., & Lau, H. (2025). *Response Time as Decision Confidence: Insights from Type-2 ROC Analysis*. https://doi.org/10.31234/osf.io/6gyjf_v1
Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology, 78*, 99–147. https://doi.org/10.1016/j.cogpsych.2015.01.002
Nauhaus, I., Busse, L., Carandini, M., & Ringach, D. L. (2009). Stimulus contrast modulates functional connectivity in visual cortex. *Nature Neuroscience, 12*(1), 70–76. https://doi.org/10.1038/nn.2232
Rafiei, F., Shekhar, M., & Rahnev, D. (2024). The neural network RTNet exhibits the signatures of human perceptual decision-making. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-024-01914-8
Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116*(1), 59–83. https://doi.org/10.1037/a0014086
Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception & Psychophysics, 80*(1), 134–154. https://doi.org/10.3758/s13414-017-1431-5
Rausch, M., Hellmann, S., & Zehetleitner, M. (2021). Modelling visibility judgments using models of decision confidence. *Attention, Perception, & Psychophysics, 83*(8), 3311–3336. https://doi.org/10.3758/s13414-021-02284-3
Rausch, M., & Zehetleitner, M. (2019). The folded X-pattern is not necessarily a statistical signature of decision confidence. *PLoS Computational Biology, 15*(10), Article e1007456. https://doi.org/10.1371/journal.pcbi.1007456
Rausch, M., Zehetleitner, M., Steinhauser, M., & Maier, M. E. (2020). Cognitive modelling reveals distinct electrophysiological markers of decision confidence and error monitoring. *NeuroImage, 218*, Article 116963. https://doi.org/10.1016/j.neuroimage.2020.116963
Samaha, J., & Denison, R. (2022). The positive evidence bias in perceptual confidence is unlikely post-decisional. *Neuroscience of Consciousness, 2022*(1), Article niac010. https://doi.org/10.1093/nc/niac010
Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a statistical computation in the human sense of confidence. *Neuron, 90*(3), 499–506. https://doi.org/10.1016/j.neuron.2016.03.025
Scholte, H. S. (2018). Fantastic DNimals and where to find them. *NeuroImage, 180*, 112–113. https://doi.org/10.1016/j.neuroimage.2017.12.077
Shekhar, M., Fung, H., Saxena, K., Rafiei, F., & Rahnev, D. (2025). Using artificial neural networks to reveal the human confidence computation. *PLoS Computational Biology, 21*(12), Article e1013827. https://doi.org/10.1371/journal.pcbi.1013827
Shekhar, M., & Rahnev, D. (2021). The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review, 128*(1), 45–70. https://doi.org/10.1037/rev0000249
Shekhar, M., & Rahnev, D. (2024). How do humans give confidence? A comprehensive comparison of process models of perceptual metacognition. *Journal of Experimental Psychology: General, 153*(3), 656–688. https://doi.org/10.1037/xge0001524
Spence, M. L., Dux, P. E., & Arnold, D. H. (2016). Computations underlying confidence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance, 42*(5), 671–682. https://doi.org/10.1037/xhp0000179
Van Den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife, 5*, Article e12192. https://doi.org/10.7554/eLife.12192
Vivar-Lazo, M., & Fetsch, C. R. (2025). Neural basis of concurrent deliberation toward a choice and confidence judgment. *Nature Neuroscience*. https://doi.org/10.1038/s41593-025-02116-9
Volkmann.. (1934). The relation of time of judgment to certainty of judgment. *Psychological Bulletin, 31*, 672–673.
Williams, D., Phillips, G., & Sekuler, R. (1986). Hysteresis in the perception of motion direction as evidence for neural cooperativity. *Nature, 324*(6094), 253–255. https://doi.org/10.1038/324253a0
Xue, K., Shekhar, M., & Rahnev, D. (2024). Challenging the Bayesian confidence hypothesis in perceptual decision-making. *Proceedings of the National Academy of Sciences, 121*(48), Article e2410487121. https://doi.org/10.1073/pnas.2410487121
Xue, K., Zheng, Y., Rafiei, F., & Rahnev, D. (2023). The timing of confidence computations in human prefrontal cortex. *Cortex, 168*, 167–175. https://doi.org/10.1016/j.cortex.2023.08.009
Yuille, A. L., & Grzywacz, N. M. (1988). A computational theory for the perception of coherent visual motion. *Nature, 333*(6168), 71–74. https://doi.org/10.1038/333071a0
Zylberberg, A., Fetsch, C. R., & Shadlen, M. N. (2016). The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *eLife, 5*, Article e17688. https://doi.org/10.7554/eLife.17688
Zylberberg, A., Roelfsema, P. R., & Sigman, M. (2014). Variance misperception explains illusions of confidence in simple perceptual decisions. *Consciousness and Cognition, 27*, 246–253. https://doi.org/10.1016/j.concog.2014.05.012