nature human behaviour

Article

https://doi.org/10.1038/s41562-026-02544-y

Trial-by-trial fluctuations in decision criterion shape confidence

Received: 23 September 2025

Accepted: 9 July 2026

Published online: 24 August 2026

Check for updates

Robin Vloeberghs [1,3] ✉, Lara Navarrete Orejudo [1,3], Anne E. Urai [2] & Kobe Desender [1]

Many decisions are accompanied by confidence. One influential view couched within signal detection theory views confidence as the distance between a decision variable and a static decision criterion. However, recent work suggests that the criterion undergoes trial-by-trial fluctuations. Here, combining theory and model simulations, we predict that fluctuations in the decision criterion shape confidence. In 15 datasets, trial-by-trial estimates of the decision criterion were obtained with the hierarchical model for fluctuations in criterion. Across datasets, we confirmed our pre-registered hypothesis that confidence is shaped by a single-trial criterion state. This effect was found in 14 out of 15 datasets, indicating a robust pattern across a variety of paradigms and confidence scales. Beyond self-reports, the effect was also observed in reaction times, pupil-linked arousal and a neural signature of confidence. Our results show that variability in confidence, often treated as noise, actually reflects genuine sensitivity to the current state of the (fluctuating) decision criterion.

Every day we have to make a multitude of decisions. With each decision comes a subjective feeling of decision confidence, indicating the probability of having made the correct choice. Many computational models have helped to understand the mechanisms that give rise to people's reported decision confidence when they make choices. Signal detection theory (SDT) is a classical and influential computational framework of decision-making[1] and easily extends to modelling decision confidence.

In its simplest form, SDT describes how observers classify noisy stimuli as belonging to one of two categories. The theory assumes that observers generate an internal representation of the relevant stimulus information, typically referred to as the decision variable. To make a binary decision, this decision variable is compared against an internal decision criterion (Fig. 1; static criterion, grey line). Because the brain is inherently noisy, repeated presentations of the same stimulus do not always result in the same internal representation. Instead, a distribution of decision variables is generated (Fig. 1; two distributions, one for each stimulus). This inherent variability accounts for why participants can give different responses when presented with the same stimulus multiple times[1–3].

Going beyond binary decisions, SDT also explains how observers experience graded levels of confidence in the accuracy of their decisions. Specifically, decision confidence can be computed as the absolute distance between the decision variable and the decision criterion, with a larger distance indicating higher confidence[4–8]. This implementation is referred to as the distance-to-criterion hypothesis of confidence. Although this approach can be traced back already to the early proposals of SDT[1], in recent years, this approach has been popularized in the perceptual metacognition literature.

A key component in SDT is the decision criterion to account for biases (that is, when one response option is chosen more often than the other). For example, shifting the criterion to the right will result in overall more 'left' responses, while leaving the overall sensitivity unaffected. A large body of empirical work shows that human observers strategically shift the decision criterion in response to external feedback[9], the base rate of stimuli[10], task instructions[11], expectancy[12], previous choices[13,14] and monetary rewards[15]. To assess whether an observer uses a shifted (or biased) decision criterion, its position is calculated based on the hit and false alarm rates. Crucially, by doing so, it is implicitly assumed that the criterion is fixed and, thus, that this

[1]Brain and Cognition, KU Leuven, Leuven, Belgium. [2]Cognitive Psychology, Leiden University, Leiden, the Netherlands. [3]These authors contributed equally: Robin Vloeberghs, Lara Navarrete Orejudo. ✉e-mail: robin.vloeberghs@kuleuven.be

https://doi.org/10.1038/s41562-026-02544-y

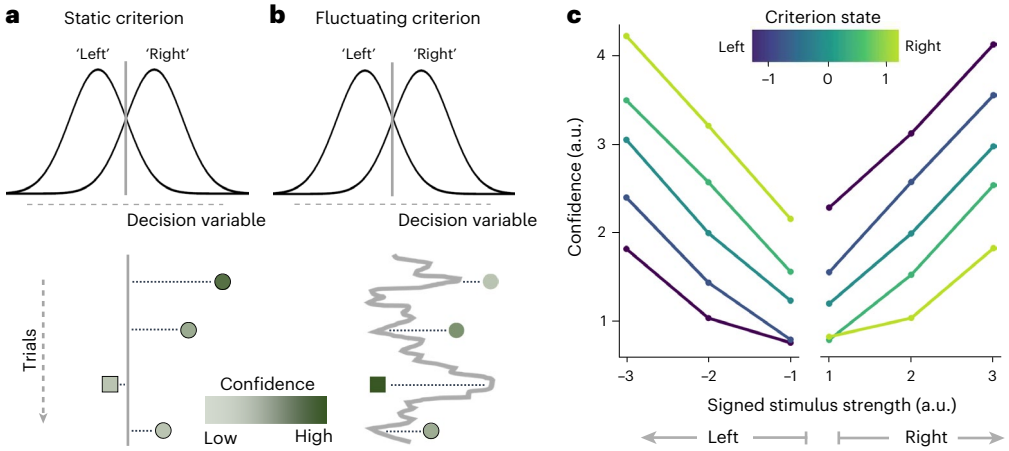

Fig. 1 | Simulating confidence as distance-to-criterion under trial-to-trial criterion fluctuations. a, According to SDT, choices are formed by comparing a decision variable to a criterion. An observer will respond 'left' (squares) or 'right' (circles), depending on whether the decision variable falls left or right of this criterion (grey line), respectively. Within SDT, confidence can be quantified as the absolute distance between the decision variable and the criterion (dotted lines), also known as the distance-to-criterion hypothesis. **b**, Whereas this criterion is typically assumed to be fixed across a series of trials (static criterion), here we investigate the consequences for decision confidence when this criterion slowly fluctuates across time. As illustrated, for identical decision variables, the predicted confidence can differ, depending on whether the criterion is static or fluctuating. **c**, Formal model simulations (6,000 trials) from an SDT framework with a fluctuating criterion (that is, as illustrated in Fig. 1b) show a clear influence of the criterion state on confidence. For the left stimuli (negative values on the *x* axis), confidence is higher when the criterion fluctuates towards the right (lighter colours). This pattern flips for the right stimuli (positive values on the *x* axis), with higher confidence when the criterion fluctuates towards the left (darker colours). a.u., arbitrary unit.

(potentially shifted) criterion is the same for all trials (Fig. 1a; static criterion). However, increasing evidence suggests that computational variables, such as the decision criterion, are not static but instead fluctuate at the level of individual trials (Fig. 1b; fluctuating criterion)[16–22].

In the SDT framework, trial-by-trial criterion fluctuations should also affect the computation of confidence, given that confidence is thought to reflect the distance between the criterion and the decision variable (Fig. 1b). More specifically, when the criterion fluctuates towards the right, on average, the distance between a decision variable favouring the 'left' stimulus and the criterion will increase, leading to higher confidence. On the contrary, when the criterion fluctuates towards the left, the distance between a decision variable favouring the 'left' stimulus and the criterion will decrease, leading to lower confidence. Indeed, exactly this pattern is observed when formally simulating data from an SDT observer with a fluctuating criterion (Fig. 1c; see Methods and 'Results' for a detailed description). These simulations also show that given the same stimulus, confidence can vary quite substantially. Thus, this suggests that variability in confidence reports, which have traditionally been treated as noise[6], might actually reflect genuine sensitivity of confidence judgments to the current state of the (fluctuating) decision criterion.

In this Article, we build on current developments in computational modelling, which allow the effective measurement of trial-to-trial fluctuations in the decision criterion. Vloeberghs et al. developed the hierarchical model of fluctuations in criterion (hMFC)[20], which uses a hierarchical Bayesian estimation approach to accurately recover criterion estimates at the single-trial level (see Methods for a detailed description). This new modelling framework now makes it possible to put the hypothesis to the test.

In summary, the current study aims to examine whether self-reported confidence ratings, reaction times and neurophysiological measures of confidence are influenced by trial-to-trial fluctuations in the decision criterion, quantified using the hMFC. We first report results from a pre-registered analysis of 15 behavioural datasets, which confirmed the influence of criterion fluctuations on confidence ratings. When these datasets were analysed individually, the effect was found in 14 out of 15 datasets. In two additional datasets, these findings were replicated using an implicit measure of confidence, namely, reaction times, and two neurophysiological markers of confidence, namely, pupil size and a neural electroencephalogram (EEG) signature of confidence. Together, these results suggest a robust effect that is present in a variety of paradigms and confidence scales and show that variability in both behavioural and neurophysiological measures of confidence is driven by fluctuations in the underlying latent decision criterion.

Results

Model simulations

To investigate whether and how confidence is shaped by criterion fluctuations, we first simulated data from an SDT observer with the criterion varying from trial to trial according to a first-order autoregressive process (AR(1)) (Fig. 1c). Using a linear regression model predicting confidence based on stimulus strength, stimulus sign, and criterion, we find that confidence increases as stimulus strength increases ($\beta = 0.840$; s.e. = 0.014; $F(1) = 3{,}313.823$; $P < 0.001$; confidence interval (CI) (95%) = (0.811, 0.868)). More interestingly, the interaction between the criterion and stimulus direction, which is central to our hypothesis, is significant ($\beta = -0.423$; s.e. = 0.035; $F(1) = 3{,}674.507$; $P < 0.001$; CI (95%) = (−0.492, −0.355)). Specifically, we find that confidence decreases for a 'right' stimulus when the criterion fluctuates to the right and increases when it fluctuates to the left. For a 'left' stimulus, the opposite pattern emerges. In addition, we observe a significant three-way interaction between the criterion, stimulus direction and stimulus strength ($\beta = -0.191$; s.e. = 0.016; $F(1) = 138.295$; $P < 0.001$; CI (95%) = (−0.223, −0.159)). Note that the significance of the three-way interaction is not a reliable signature for confidence as distance to criterion. Instead, it arises from a specific combination of stimulus configuration and the range of criterion fluctuations. Specifically, the three-way interaction will be significant when the lines in Fig. 1c are not parallel, which occurs when the dataset contains trials in which the criterion coincides with the mean of the generative SDT stimulus distribution (for example, a criterion of −1 with a stimulus of −1, or a criterion of +1 with a stimulus of +1). In Supplementary Fig. 1, we simulated an agent with a smaller range of the criterion fluctuations where such trials did not occur. For this agent, although the critical two-way interaction between the stimulus direction and criterion is significant,

https://doi.org/10.1038/s41562-026-02544-y

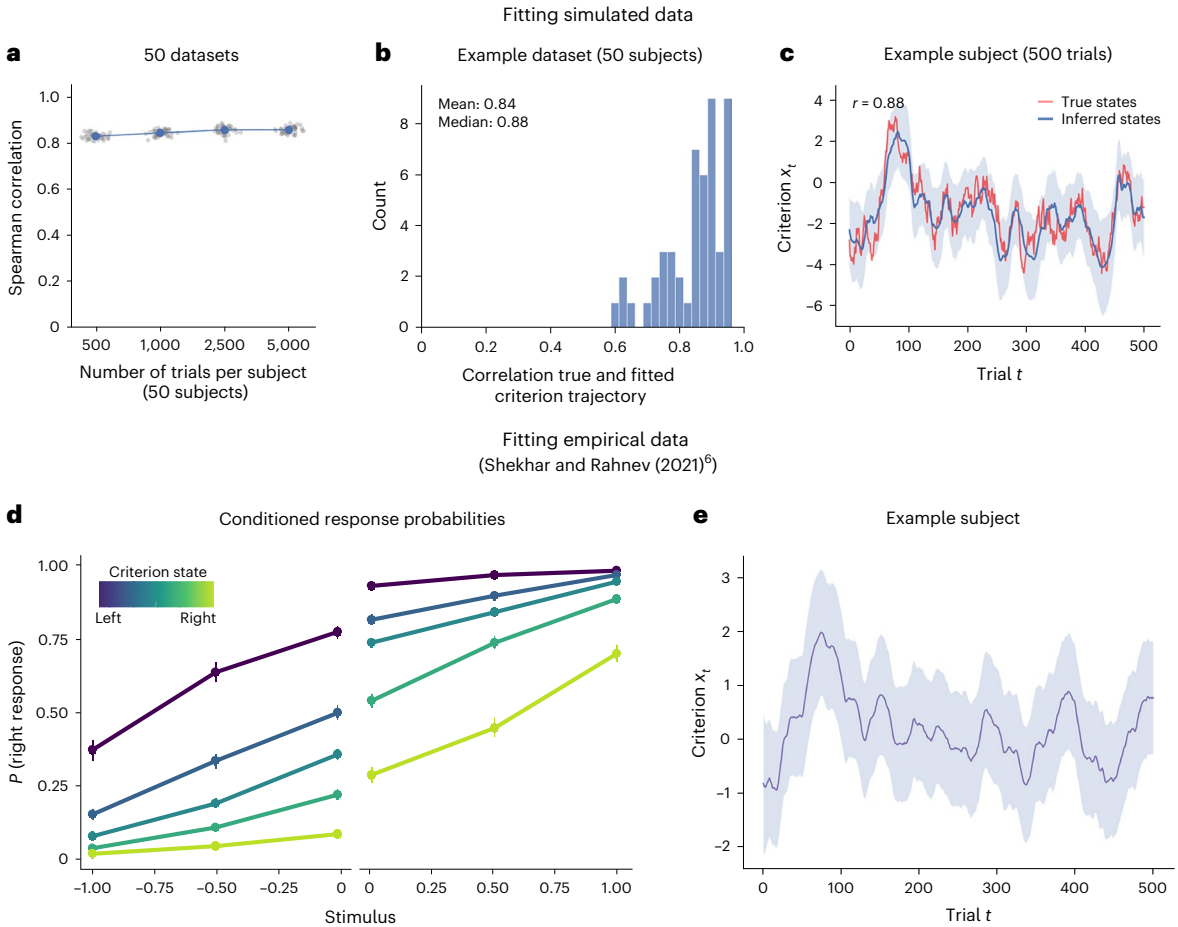

Fig. 2 | Estimating criterion fluctuations with the hMFC on simulated and empirical data. a, The recoverability of the criterion fluctuations by the hMFC was assessed in Vloeberghs et al.[20] using simulated data with a varying number of trials per subject (500, 1,000, 2,500 or 5,000 trials), with 50 subjects per dataset and 50 datasets per trial count. Each dot represents the correlation between the true and estimated criterion trajectory averaged over subjects in one dataset. The error bars are plotted but are very small. **b**, The correlation between the true and estimated criterion trajectory for each subject (with 500 trials) in a representative simulated dataset. **c**, The true and estimated criterion fluctuations for a simulated subject with 500 trials. The line for the estimated fluctuations shows the posterior mean, and the shaded area indicates the 95% credible interval of the posterior at each trial. **d**, The empirical response probabilities of the data (20 subjects with 2,800 trials each) of Shekhar and Rahnev (experiment 4)[6] conditioned on the binned trial-by-trial criterion. The more the criterion is shifted to the left, the more right responses are observed. Error bars reflect standard errors. **e**, The estimated criterion trajectory of an example subject (first 500 trials are shown). The line indicates the posterior mean of each trial, with the shaded area representing the 95% credible intervals.

the three-way interaction is not, providing evidence that this is not a key model prediction.

To assess how the results change as a function of accuracy, we performed separate analyses for correct and error trials (Supplementary Fig. 2). The pattern for correct trials mirrors that of the full dataset, whereas for error trials, the interaction between the stimulus direction and criterion reverses, with higher confidence for leftwards stimuli when the criterion fluctuates leftwards.

Empirical data

Following these simulations, we next investigated our hypothesis in empirical data, estimating criterion fluctuations using the hMFC[20]. The hMFC can accurately recover the criterion fluctuations in simulated data, even with only 500 trials per subject (Fig. 2a–c). In empirical data, the trial-by-trial criterion estimated by the hMFC captures differences in response probabilities that are consistent with criterion shifts in SDT, as illustrated by the response probabilities conditioned on criterion state shown in Fig. 2d. The more the criterion fluctuates to the right, the more left responses are observed. Conversely, a leftwards shift in the criterion leads to more rightwards responses. The model fits show stable and interpretable posterior distributions, with proper mixing of the chains of the Gibbs sampler (all $\hat{R} < 1.05$), reassuring that we sampled from the full posterior range (Supplementary Fig. 3 and Supplementary Table 1). Supplementary Table 2 reports the mean, standard deviation, and minimum and maximum values, calculated per study, for the per-subject parameters of the autoregressive model that characterizes the criterion fluctuations. Altogether, these results ensure the validity of the hMFC estimates.

Next, we turn towards the key question of the current work: is confidence shaped by trial-to-trial fluctuations in the decision criterion? In empirical data, we observed the same confidence patterns as seen in the simulations (Fig. 3a). Indeed, when fitting empirical data using a linear mixed model in which confidence is predicted based on stimulus strength, stimulus direction and criterion (model 1; see Methods), we found a highly significant interaction between stimulus direction and criterion, hereby confirming our main hypothesis ($\beta = -0.043$; s.e. = 0.002; $F(1, 456{,}507) = 781.961$; $P < 0.001$; CI (95%) = (−0.046, −0.040)). The negative slope indicates lower confidence for a right stimulus (+1) when the criterion drifts towards the right (positive values), and higher confidence when the criterion drifts towards the left (negative values). The opposite is true for left stimuli (−1). In addition, the three-way interaction was significant, similar to the simulations ($\beta = -0.032$; s.e. = 0.003; $F(1, 456{,}276) = 123.330$; $P < 0.001$; CI (95%) = (−0.038, −0.026)), with a larger interaction between stimulus

https://doi.org/10.1038/s41562-026-02544-y

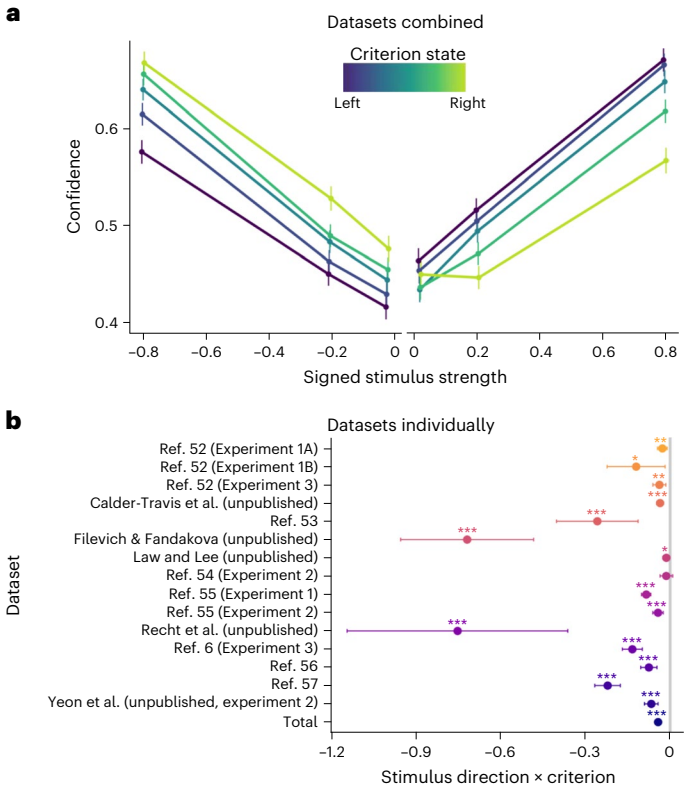

Fig. 3 | Trial-by-trial fluctuations in the decision criterion shape confidence ratings in empirical data. a, As theoretically predicted, and in line with model simulations, decision confidence was modulated by trial-by-trial fluctuations in the decision criterion. Specifically, when the stimulus direction was 'right' (positive-signed stimulus strength), confidence increased as the decision criterion fluctuated towards the left (that is, became more blue) and confidence decreased as the decision criterion fluctuated towards the right (that is, became more yellow). Data are shown across the 15 datasets. Error bars reflect standard errors. For plotting purposes only, criterion states were binned and the average within each bin is shown. **b**, Parameter estimates for the crucial interaction between stimulus direction and decision criterion and their 95% confidence intervals. The estimates are obtained from linear mixed models fitted separately to every dataset[6,52–57]. The estimate plotted at the bottom, labelled 'Total', shows the parameter estimate from the model fitted to all datasets (taking into account the nesting of subjects in experiments). The number of subjects and trials in each dataset is reported in Supplementary Table 4. The *P* values were obtained using two-sided tests. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

direction and criterion for stronger stimulus strength. Next, we found a significant effect of stimulus strength, with confidence increasing as stimulus strength increases ($\beta = 0.587$; s.e. = 0.114; $F(1, 14) = 26.715$; $P < 0.001$; CI (95%) = (0.364, 0.810)). Lastly, the main effect of stimulus direction was borderline significant ($\beta = -0.002$; s.e. = 0.001; $F(1, 456{,}483) = 3.665$; $P = 0.056$; CI (95%) = (−0.004, 0.000)). All other effects were not significant (all $P \geq 0.144$). No multicollinearity problems were detected (all variance inflation factor (VIF) ≤ 2.607). Through a sensitivity analysis, we ensured that the results were identical for different values of the fixed hyperparameters μ_σ^2 and β_σ^2 (Supplementary Table 3).

When analysing correct and error trials separately, we found a similar pattern as in the full dataset for correct trials (Supplementary Fig. 4). Consistent with the simulations, for error trials, we found that the interaction between stimulus direction and criterion flips, with higher confidence for left stimuli when the criterion fluctuates towards the left. The model selection results can also be found in Supplementary Fig. 4.

To investigate the robustness of the critical two-way interaction between stimulus direction and criterion, we reanalysed all datasets with confidence measured on 3−, 4− or 6-point rating scales using an ordinal mixed regression model (Supplementary Fig. 5). Of the 12 datasets suitable for this analysis, 10 showed a significant interaction. The remaining two datasets were not significant ($P = 0.061$ and $P = 0.066$).

To assess whether these findings reflect meaningful contributions of criterion fluctuations and not just statistical noise or model overfitting, we turned to model comparison. Specifically, we tested whether the linear mixed model reported above (model 1) outperformed a simpler model that lacked criterion fluctuations (model 2). Results provide unequivocal support that the more complex model including the criterion is preferred, even when controlling for complexity (ΔAIC (AIC model 1 − AIC model 2) = −7,482; ΔBIC (BIC model 1 − BIC model 2) = −7,372; χ^2 (10) = 7,502; $P < 0.001$). This indicates that a significant portion of variability in decision confidence can be explained by fluctuations in the decision criterion.

To further rule out the possibility that these results reflect spurious correlations that can occur between slowly drifting variables[23], we implemented the session permutation method[24]. In this approach, criterion fluctuations were randomly shuffled across participants (model 3) and compared with the model with the original fluctuations (model 1). Once again, the comparison strongly favoured model 1 (ΔAIC = −8,091; ΔBIC = −7,848). In model 3, the shuffled criterion fluctuations failed to account for any variability in confidence, as there was no interaction between stimulus direction and criterion ($\beta = -0.002$; s.e. = 0.002; $F(1, 137{,}574) = 1.626$; $P = 0.202$; CI (95%) = (−0.006, 0.001)). Also, the three-way interaction between stimulus direction, stimulus strength and criterion was absent ($\beta = 0.001$; s.e. = 0.003; $F(1, 269{,}873) = 0.182$; $P = 0.700$; CI (95%) = (−0.005, 0.008)). Only the main effect of stimulus strength remained significant ($\beta = 0.592$; s.e. = 0.116; $F(1, 14) = 26.228$; $P < 0.001$; CI (95%) = (0.364, 0.820)). For full comparability, we report the model comparison for model 2 fitted to the truncated dataset in Supplementary Analysis 1.

Previously, we found, across 15 datasets, a highly robust association between confidence and the single-trial state of the decision criterion. Next, we further unravelled to what extent this pattern was present in each of the individual datasets. To this end, we fitted the model to each dataset separately. Importantly, the key two-way interaction between stimulus direction and criterion, central to our hypothesis, was significant in 14 out of 15 datasets (Fig. 3b). The three-way interaction between stimulus direction, stimulus strength and criterion, which was also predicted by the simulations, was significant in 10 out of 15 datasets, with 7 showing a negative slope and 3 showing a positive slope. In the remaining 5 datasets, the interaction was absent (Supplementary Fig. 6). In summary, although the three-way interaction is noisier and more difficult to detect at the dataset level, the consistent presence of the crucial interaction for our main hypothesis between stimulus direction and criterion in 14 out of 15 datasets shows that the effect is robust and generalizes across diverse paradigms and confidence rating scales.

Having established that criterion fluctuations shape confidence ratings, we next explored whether this is also the case for an implicit measure of confidence (reaction times) and for neurophysiological correlates of confidence. While ratings provide valuable subjective reports, they rely on participants' introspection and may be influenced by biases or limitations in self-assessment. To complement this, we examined more objective indices of confidence that do not depend on introspective accuracy, namely, reaction times, pupil size and a neural marker of confidence (error positivity (Pe)).

First, we consider reaction times. Reaction times are regarded as an implicit behavioural measure of confidence, with longer reaction times corresponding to lower confidence[7,25]. When predicting reaction times, there was a significant effect of stimulus strength ($\beta = -0.183$; s.e. = 0.014; $F(1, 26.25) = 210.460$; $P < 0.001$; CI (95%) = (−0.211, −0.155)), with higher stimulus strength (that is, easier stimuli) resulting in lower

https://doi.org/10.1038/s41562-026-02544-y

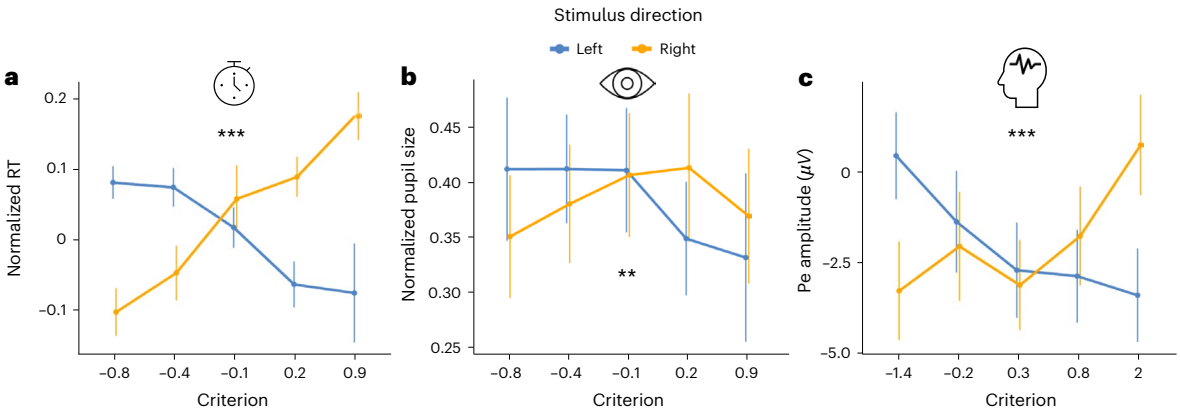

Fig. 4 | Criterion fluctuations shape implicit and neurophysiological measures of confidence. a–**c**, Reaction times (**a**), pupil size (**b**) and a neural marker of confidence (**c**) are all significantly predicted by the interaction between stimulus direction and criterion in linear mixed models (reaction times: $P < 0.001$, pupil sizes: $P = 0.003$, Pe amplitude: $P < 0.001$). Specifically, when the criterion fluctuates to the right (positive values) for a rightwards stimulus (yellow line), longer reaction times, larger pupil sizes and a higher Pe amplitude are observed, corresponding to low confidence. In contrast, when the criterion fluctuates to the left (negative values) for a rightwards stimulus, lower values for all three measures are observed, indicating high confidence. The figure shows the mean data, with error bars denoting the standard error across subjects. Reaction times and pupil size show data from Urai et al. (27 subjects with 500 trials each)[7], while Pe amplitude was taken from Boldt and Yeung[27] (16 subjects with 864 trials each). In the latter dataset, only one level of stimulus strength was included, so it was not possible to show the data as a function of stimulus strength (as in Fig. 3a). For consistency across panels, we therefore present only the two-way interaction. The *P* values were calculated using two-sided tests. $^{**}P < 0.01$; $^{***}P < 0.001$.

reaction times. Crucially, the interaction between stimulus direction and criterion was significant ($\beta = 0.366$; s.e. = 0.053; $F(1, 23.34) = 47.666$; $P < 0.001$; CI (95%) = (0.262, 0.470)), indicating longer reaction times when the criterion fluctuates to the right for rightwards stimuli (+1). Conversely, for leftwards stimuli (−1), a rightwards-shifted criterion resulted in shorter reaction times (Fig. 4c). We further found a significant three-way interaction between stimulus direction, stimulus strength and criterion ($\beta = 0.026$; s.e. = 0.007; $F(1, 1{,}996.98) = 13.171$; $P < 0.001$; CI (95%) = (0.012, 0.041)). All other effects were not significant (all $P \geq 0.156$). It should be noted that this model showed some multicollinearity, with VIF scores reaching up to 10.25. Excluding the three-way interaction reduced VIFs to ≤ 3.185, while leaving all other results unchanged.

Second, we examined a well-known neurophysiological marker of confidence: pupil size[7,26]. Phasic arousal, as quantified through pupil size, is inversely related to confidence, with smaller pupil size reflecting higher confidence[26]. When predicting pupil size, a significant main effect of stimulus strength emerged ($\beta = -0.045$; s.e. = 0.007; $F(1, 26) = 38.757$; $P < 0.001$; CI (95%) = (−0.058, −0.032)). Specifically, easier stimuli were associated with smaller pupil sizes, supporting the idea that pupil size is inversely related to confidence, as higher confidence is expected for easier stimuli. More importantly, we found that the crucial interaction between stimulus direction and criterion was significant ($\beta = 0.064$; s.e. = 0.021; $F(1, 67{,}112) = 8.992$; $P = 0.003$, CI (95%) = (0.022, 0.106)). A right-shifted criterion was associated with larger pupil sizes for rightwards stimuli compared with leftwards stimuli, and vice versa (Fig. 4a). Finally, also the three-way interaction was significant ($\beta = 0.018$; s.e. = 0.008; $F(1, 67{,}123) = 5.307$; $P = 0.021$), as well as the interaction between stimulus direction and stimulus strength ($\beta = 0.013$; s.e. = 0.005; $F(1, 67{,}126) = 7.292$; $P = 0.007$; CI (95%) = (0.003, 0.032)). No other effects were present (all $P \geq 0.294$) and multicollinearity issues were absent (all VIF ≤ 3.161).

Lastly, we investigated a neural marker of confidence, namely, the Pe measured with EEG[27]. The Pe is sensitive to graded levels of confidence, with its amplitude decreasing as confidence increases[27,28]. When predicting single-trial Pe amplitudes, we found the critical interaction between stimulus direction and criterion ($\beta = 1.912$; s.e. = 0.203; $F(1, 764.50) = 89.23$; $P < 0.001$; CI (95%) = (1.516, 2.310)), suggesting higher amplitudes for a rightwards stimulus when the criterion fluctuates to the right, and lower amplitude when it fluctuates to the left (Fig. 4b). Also, the main effect of criterion was significant ($\beta = -0.476$; s.e. = 0.150; $F(1, 13{,}086.30) = 19.786$; $P < 0.001$; CI (95%) = (−0.770, −0.182)). In this experiment, there was only one level of stimulus strength, so the main effect of stimulus strength was omitted from the model. No multicollinearity issues were present (all VIF ≤ 1.942). Note that the same key finding in the pre-registered analysis on 15 datasets (that is, interaction between stimulus direction and criterion) was also observed when analysing the confidence ratings of this study ($\beta = -1.082$; s.e. = 0.086; $F(1, 12.95) = 156.931$; $P < 0.001$; CI (95%) = (−1.252, −0.913)) with no other effects present (all $P \geq 0.564$). Note that the analysis performed on correct and error trials separately for all three confidence markers can be found under Supplementary Analysis 2.

For all these three markers of confidence, the model including criteria outperformed the model without them (all ΔAIC ≤ −78; ΔBIC ≤ −41), and it was also preferred over the model with shuffled criterion trajectories (all ΔAIC ≤ −25; ΔBIC ≤ −61). Together, these results show that behavioural and neurophysiological markers of confidence are shaped by fluctuations in the decision criterion.

Discussion

Within SDT, decision confidence is traditionally conceptualized as the absolute distance between a decision variable and a criterion, which is known as the distance-to-criterion hypothesis. Crucially, in all previous work, the criterion is assumed to be static. However, an increasing body of research suggests that computational parameters, such as the decision criterion, are dynamic and fluctuate over trials. On the basis of the distance-to-criterion hypothesis, we thus predict that fluctuations in the decision criterion influence confidence ratings. Using the recently developed hMFC[20], which allows for obtaining trial-by-trial estimates of the decision criterion, we tested this hypothesis. Across 15 datasets, we found strong and consistent evidence for this hypothesis, demonstrating a robust effect independent of the specific paradigm or confidence scale. In addition, we replicated the effect in pupil size, in a neural marker of confidence and in reaction times, suggesting that criterion fluctuations constitute a source of variability in confidence across both behavioural and neurophysiological measures.

The current work provides strong evidence in favour of the distance-to-criterion hypothesis of confidence: across 17 datasets, we found that seemingly random variability in decision confidence

actually tracks fluctuations in the decision criterion. For confidence ratings, this effect was significant in 15 out of 16 datasets (including the dataset used for the neural marker of confidence), showing that this reflects a strong and highly replicable finding. Although the interaction was not significant in one dataset, the direction of the effect was consistent with the hypothesis. From a statistical point of view, finding one null effect across 16 datasets is not surprising: given the experimental power of 0.8, it can be expected that 3 out of 16 studies show a null effect even when the effect is real. The one study where the effect was not found was not an obvious outlier in terms of design (that is, a Gabor detection task with four-choice confidence reporting) nor the estimated hMFC parameters (Supplementary Table 2). Moreover, we replicated our findings across different neurophysiological measures of confidence, indicating that the sensitivity to a fluctuating criterion reflects a fundamental aspect of how confidence is computed rather than merely a bias in how it is reported. In the current study, we used pupil size as a neurophysiological measure of confidence. While pupil size is considered a proxy of confidence[7,26], it has also been linked to arousal, with fluctuations in neuromodulatory systems such as norepinephrine and acetylcholine associated with changes in pupil diameter[29,30]. Consequently, it remains an open question to what extent some of our findings reflect fluctuations in these neurotransmitter systems.

Our findings show that a portion of variability in confidence ratings, which is typically interpreted as the result of some form of noise, can be attributed to fluctuations in the decision criterion. Models not considering the possibility of fluctuations in the criterion typically attribute variability in confidence to some form of metacognitive noise that selectively perturbs the computation of confidence[6,31–33]. However, our current findings, unravelling the sensitivity of confidence to criterion fluctuations, advocate for a well-calibrated metacognitive system, at least with respect to criterion fluctuations. Importantly, our findings do not argue against the existence of metacognitive noise per se. It is very likely that variability in confidence is driven by fluctuations in both the decision criterion, reported in the current work, and noisy confidence criteria, as put forwards by research on metacognitive noise. In future work, the respective contributions of fluctuations in criterion and metacognitive noise in explaining variability in confidence could be studied through a key distinction: fluctuations in the decision criterion affect both responses and confidence ratings, whereas metacognitive noise affects only confidence ratings. While such a model does not yet exist, an alternative approach would be to determine the relative contribution by explicitly estimating trial-to-trial fluctuations in confidence criteria (see Supplementary Fig. 7 for an explorative approach to estimate confidence criteria fluctuations using the hMFC).

One important consequence of criterion fluctuations is that they influence well-established measures of perceptual sensitivity, such as d' (ref. 1). Simulations show that the presence of criterion fluctuations leads to an underestimation of d' (Supplementary Fig. 8a–c). Interestingly, a similar pattern emerges for metacognitive sensitivity (meta-d'), which is likewise underestimated with strong criterion fluctuations. Because d' and meta-d' are affected to a comparable extent, their ratio (M-ratio, commonly used as a measure of metacognitive efficiency[34]) remains largely unchanged. Consequently, M-ratio appears relatively robust to criterion fluctuations. Remarkably, while simulations show that d' is underestimated in the presence of criterion fluctuations, simulated accuracy remains unchanged (Supplementary Fig. 8d,e). Therefore, criterion fluctuations are able to explain dissociations between confidence and accuracy. Such dissociations have been observed in many studies and are often thought to provide a hint about the underlying computations of decision confidence[35–40]. Given that confidence accuracy dissociations are so widespread in the literature, it remains beyond the scope of the current project to examine to what extent these different phenomena can be explained by criterion fluctuations and is therefore left for future work.

The pattern of increased confidence combined with reduced perceptual sensitivity has been associated with a symptom dimension characterized by compulsive behaviours and intrusive thoughts[41,42]. When accuracy is used instead of d', this relationship is no longer observed[42]. From a theoretical perspective, one explanation for these findings is that individuals scoring higher on these psychiatric symptoms show stronger criterion fluctuations (that is, increased a and/or σ), leading to apparent reductions in sensitivity and overconfidence. Going beyond single-trial criterion estimates, future work could thus investigate whether individual autoregression parameters are correlated with subclinical variability.

Another important consequence of quantifying confidence as the distance between the decision variable and a fluctuating criterion is that confidence will naturally become autocorrelated across trials. This phenomenon has been described previously in empirical data and is often referred to as the confidence leak[43]. To explain autocorrelation in confidence (that is, confidence leak), Rahnev et al.[43] proposed that participants dynamically update their confidence criterion on each trial depending on whether their confidence on the previous trial was higher or lower than the average. While fluctuations in criterion theoretically might seem an alternative explanation, we did not find strong evidence for this in our empirical data. Although simulations from our model show very strong autocorrelation of confidence (that is, a strong confidence leak), this effect disappears when adding the estimated criterion fluctuations to the model. However, when applying the same approach to empirical data from Shekhar and Rahnev[6], the effect of previous confidence on current confidence was largely unaffected by inclusion of the criterion state. From this, we can conclude that at the moment, there is no convincing evidence that criterion fluctuations are an alternative explanation to the confidence leak in empirical data.

At the same time, the interpretation of these fluctuations is not uniquely determined. In the current work, we interpret the output of the hMFC as reflecting fluctuations in the decision criterion. Alternatively, the same pattern could arise from simultaneous fluctuations in the underlying stimulus distributions, while the criterion itself remains static. For example, fluctuations in neuronal excitability could change the baseline firing rate of neurons over time, effectively shifting the internal evidence distributions[44]. From a modelling perspective, these alternatives are mathematically equivalent, and distinguishing between them would require additional neural data.

A further limitation of the hMFC is that it is formulated within the SDT framework, which does not account for reaction times. Explaining reaction time distributions typically requires sequential sampling models such as the drift diffusion model (DDM)[45]. Importantly, whereas SDT includes a single-bias parameter (the criterion), the DDM includes two potential sources of bias: the starting point and the drift bias[14]. To investigate the mapping between these parameters and the SDT criterion, we developed a DDM variant where starting point and drift bias fluctuate according to an AR(1) process (see also Vloeberghs et al.)[46]. We fitted this DDM to the same empirical datasets as the hMFC and found that criterion fluctuations were significantly predicted by drift bias, both when analysing all datasets together or individually (Supplementary Fig. 9). Criterion fluctuations also mapped onto fluctuations in the starting point when analysing all datasets together. When these datasets were analysed individually, the effect was present only in 4 out of 14 datasets. Note that one dataset was excluded from the analysis owing to a task design that was incompatible with the DDM. These results suggest that drift bias can be consistently detected at the dataset level, whereas starting point is considerably noisier and requires substantially more data to be reliably identified. Taking these together, we can conclude that criterion fluctuations measured by the hMFC toolbox reflect both fluctuations in drift bias and starting point.

An important question for future work is how these findings relate to alternative conceptualizations of the decision criterion. The current paper considers the criterion in evidence space, whereas Bayesian

models define the criterion in probability space. Future research could compare these approaches and examine whether they allow for a clearer distinction between the distance-to-criterion account and Bayesian accounts of confidence[8].

In summary, our study provides robust evidence that confidence is modulated by trial-to-trial fluctuations in the decision criterion. This finding shows that variability in confidence judgments, which is often assumed to reflect noise, reflects genuine computation of confidence as distance-to-criterion. With this work, new insights are provided into the computational underpinnings of decision confidence.

Methods

Open science statement

A first exploratory analysis was performed on the data of experiment 4 in Shekhar and Rahnev[6], which is publicly available. Afterwards, we performed a pre-registered confirmatory analysis of 14 datasets from the Confidence Database[36]. The pre-registration can be found at https://osf.io/mxuv2. For ease of reading, in the remainder, we will jointly discuss the results from the exploratory analysis together with the 14 confirmatory analyses. Finally, a pre-registered analysis regarding individual differences can be found in Supplementary Analysis 3. The analyses on the two additional datasets involving reaction times, pupil size and the neural marker of confidence are not included in the pre-registration.

Simulation study

To investigate how criterion fluctuations shape confidence, we simulated 6,000 trials from an SDT observer. On each trial, a decision variable was sampled from a normal distribution with mean $\mu = (3, -2, -1, 1, 2, 3)$ and standard deviation $\sigma = 1$. The decision criterion varied across trials according to AR(1) with autoregressive coefficient $a = 0.9995$ and noise standard deviation $\sigma = 0.05$. A binary response on each trial was obtained by comparing the decision variable with the decision criterion. Confidence was then quantified as the absolute distance between the decision variable and the decision criterion. To examine the relationship between confidence and criterion fluctuations, we used a linear regression model where confidence was predicted by a full-factorial model including as predictors the decision criterion, stimulus direction and stimulus strength. Stimulus strength was defined as the absolute value of the mean μ of the normal distribution, while stimulus direction corresponded to its sign. For visualization purposes only, the decision criterion was divided into equally sized bins.

Dataset selection

As pre-registered, we included only studies that met the following inclusion criteria: (1) participants reported trial-by-trial confidence using a confidence rating scale, (2) each participant completed a minimum of 500 trials (that is, to allow robust estimation of trial-by-trial criterion with the hMFC), (3) the main manipulation involved the stimulus strength presented to participants on the screen and (4) no trial-by-trial feedback was provided. On the basis of these criteria, a total of 14 datasets were selected (that is, in addition to Shekhar and Rahnev[6], which was used for the exploratory analysis). Note that, in contrast to the pre-registration, the datasets of Yeon et al. (experiment 1, unpublished) and Clark et al.[47] are not included in the current analysis. The former was wrongly classified and used only one level of stimulus strength, and for the latter, the fitting procedure of the hMFC did not converge owing to the low number of subjects ($n = 4$).

From the 15 studies included, 8 used Gabor patches as experimental stimuli, 4 studies used the random-dot motion task (that is, participants indicated whether they perceived left- or rightwards movement within a cloud of moving dots), 2 studies asked participants to decide whether the left or right panel contained the most dots, and 1 study used ambiguous faces, which had to be classified as either looking fearful or happy. The majority of the studies, 10 out of 15, used a discrete 4-point confidence scale, whereas the other 5 studies used either a continuous scale or a 2-, 3-, or 6-point scale. Across studies, the number of subjects varied from 10 to 54, and the number of trials ranged between 504 and 3,240. Thirteen subjects were excluded because they either performed at the chance level or provided the same confidence rating on over 90% of trials. Six additional subjects were removed owing to estimated criterion states larger than 10. Values this large saturate the sigmoid-transformed log-odds in the hMFC, thereby jeopardizing the accuracy of the model estimates. In total, after we excluded these subjects, the final datasets comprised 379 subjects with 459.091 trials in total. Detailed information about each of the studies can be found in Supplementary Table 4.

To examine whether our findings extend to implicit measures of confidence and to neurophysiological correlates of confidence, we analysed two additional datasets. For reaction time and pupil size, we used data from Urai et al.[7], which included 27 human observers performing 500 trials of a two-interval forced-choice motion coherence discrimination task. In this task, participants judged whether a test random-dot motion stimulus had higher or lower coherence than a reference stimulus, while their pupil sizes were recorded. Throughout the experiment, five stimulus strength levels were presented. Note that this dataset did not include confidence ratings. Reaction times were log-transformed and standardized. Pupil size was calculated as the mean baseline-corrected signal during the 250 ms preceding feedback. More details on pupillometry preprocessing are provided in Urai et al.[7]. To assess the modulation of criterion fluctuations in a neural marker of confidence (Pe), we reanalysed data from Boldt and Yeung[27]. In this study, 16 participants completed 864 trials in which they judged which of the two panels contained more dots and then reported their confidence on a 6-point scale. Only a single stimulus strength level was used in this experiment. Single-trial Pe amplitudes were calculated at electrode Pz in a time window of 250–350 ms post-response, following Boldt and Yeung[27]. Technical details of the EEG preprocessing are available in the original study.

Estimating the single-trial decision criterion using the hMFC

The single-trial decision criterion was obtained by fitting the hMFC separately to each of the 15 datasets[20]. This model uses a Bernoulli distribution to predict the binary response y_t at trial t:

$$y_t \sim \text{Bern}\left(f\left(\boldsymbol{w}^T \boldsymbol{u}_t + x_t\right)\right) \quad (1)$$

where the probability of a response is determined by a set of input variables u_t, weighted by $\boldsymbol{w}$, and a latent variable x_t representing the decision criterion for that trial. The sigmoid transformation $f(z)$ restricts the probability between 0 and 1. The criterion state x_t is modelled as a first-order autoregressive model with intercept b, autoregressive coefficient a and error variance σ^2:

$$x_t \sim b + a x_{t-1} + \epsilon_t, \epsilon_t \sim \text{N}\left(0, \sigma^2\right) \quad (2)$$

Adding the fluctuating value x_t to the Bernoulli log-odds is equivalent to comparing a decision variable to a fluctuating criterion. Note, however, that the estimated criterion x_t, as outputted by the hMFC, is a bias on the probability of making a certain response (for example, right response). Thus, a positive criterion state will increase the probability of a right response and therefore corresponds in SDT to a criterion shifted to the left. For ease of interpretation, in further analyses, the sign of x_t is inverted such that its interpretation of x_t is in line with the traditional SDT interpretation. Note that systematic effects on the criterion (for example, stimulus base-rate manipulations, feedback conditions or response on the previous trial) can be estimated by including this variable as covariate u_t in the hMFC. The AR(1) process for the criterion fluctuations therefore models the endogenous fluctuations on top of the systematic effects on the criterion of the included covariates. Although these criterion fluctuations are modelled as an

AR(1) process, the hMFC can recover deviations from this assumption through its use of the forward filtering–backward sampling algorithm[48], for instance, capturing the influence of omitted variables or estimating non-stationary patterns in the criterion dynamics, such as systematic trends over time (Supplementary Fig. 10).

In addition to the per-subject parameters ($\boldsymbol{w}, b, a, \sigma^2$), the hMFC estimates hierarchical (group-level) distributions, allowing the sharing of statistical strength across subjects to enhance the estimation accuracy of these individual parameters. Through this hierarchical structure, the hMFC can accurately recover the model parameters, even with as few as 500 trials per subject. Most importantly, the hMFC shows excellent recovery of the latent decision criterion trajectories (Fig. 2a–c), providing accurate trial-to-trial estimates of the decision criterion. For full details on the model and its parameter recovery, we refer the reader to Vloeberghs et al.[20].

In the current study, as input variables $\boldsymbol{u}_t$ for the hMFC, we considered the current trial stimulus, previous trial stimulus and both response and confidence on the previous trial, together with their interaction. Note that in the data of Urai et al.[7], used to investigate pupil size, the self-reported confidence rating was absent, and thus, confidence on the previous trial could not be used as a predictor in the hMFC. The model was run with four chains, each generating 5,000 posterior samples, with the first 500 samples discarded as burn-in. Convergence of the chains was assessed using the $\hat{R}$ diagnostic. Fitting the full model with all the hierarchical (group-level) parameters resulted in convergence issues of the different chains (that is, $\hat{R} > 1.05$) for the majority of the datasets. These issues were resolved by fixing μ_σ^2 and β_σ^2, the two parameters of the inverse gamma hierarchical prior for σ^2, to the estimated values for Shekhar and Rahnev[6], for which the model did properly converge ($\mu_\sigma^2 = 0.15$ and $\beta_\sigma^2 = 0.15$). For consistency, all datasets were fitted with these two hyperparameters fixed. To ensure the robustness of our main results in light of these fixed parameters, we conducted a sensitivity analysis across a range of values for μ_σ^2 and β_σ^2 (Supplementary Table 3).

Predicting confidence with the single-trial criterion state

To test the hypothesis that criterion fluctuations shape confidence ratings, we estimated three linear mixed-effects models. The dependent variable in each of these models was reported confidence, which was rescaled to range between 0 and 1 for all datasets. In the first model, confidence was predicted based on the interaction between stimulus strength, stimulus direction and the single-trial criterion estimates, obtained with the hMFC. In a random-dot motion task, stimulus strength would correspond to the motion coherence and stimulus direction would refer to left- or rightwards movement. Stimulus strength was recoded in each dataset to fall between 0 and 1, and stimulus direction was recoded as −1 and 1. The second model includes only stimulus strength and stimulus direction. The third model includes the estimated criterion trajectories but shuffled across subjects within each dataset. As such, we can use the session permutation method[24]. Statistically comparing autocorrelated time series can introduce spurious significant effects when not sufficiently controlled for. To ensure that our models picked up only on confidence variability that is truly driven by trial-to-trial criterion fluctuations, we therefore used the session permutation method by comparing model 1 (original criterion fluctuations) with model 3 (shuffled criterion fluctuations). If confidence ratings are indeed influenced by criterion fluctuations, the first model including trial-by-trial criterion estimates should be favoured over the other two models in a model selection procedure.

Model 1: confidence ~ stimulus strength × stimulus direction × criterion + (1 + stimulus strength + criterion | study/subject)

Model 2: confidence ~ stimulus strength × stimulus direction + (1 + stimulus strength | study/subject)

Model 3: confidence ~ stimulus strength × stimulus direction × shuffled criterion + (1 + stimulus strength + stimulus direction + shuffled criterion | study/subject)

First, we estimated the three models for all datasets combined. To account for the nesting of participants within studies when analysing the datasets simultaneously, we used nested random effects. For model selection, it should be noted that the comparisons between model 1 and model 2, and between model 1 and model 3, were not conducted on the same dataset. Therefore, the results of these comparisons should not be compared directly. Specifically, the session permutation method used for model 3 required all subjects within a dataset to have the same number of trials. To achieve this, within each dataset, we determined the cut-off based on the subject with the fewest trials and excluded any trials exceeding this count. As such, the estimated criterion trajectories could be shuffled across subjects. While the comparison between model 1 and model 3 was conducted on this truncated dataset, the comparison between model 1 and model 2 was performed on the full dataset. For completeness, Supplementary Analysis 1 reports the comparison between model 1 and model 2 using the truncated dataset as well. Second, in addition to analysing all datasets combined, we analysed each dataset separately using model 1, but without the random effects nested within each study.

To examine whether fluctuations in decision criterion influence reaction times and neurophysiological markers of confidence, we applied the same approach described above, replacing confidence ratings with reaction times, pupil size or Pe amplitude as the dependent variable.

For all models, the random-effects structure was obtained using an automated procedure that fits all possible random-effects structures in parallel and selects the model with the lowest AIC. All statistical analyses were performed in R (4.6.0) using RStudio (2026.05.0)[49]. The linear mixed-effects models were fitted using the lme4 package (1.1.37)[50]. The ordinal mixed regression models were fitted with the ordinal package (2022.11.16)[51].

Reporting summary

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

Data availability

All data analysed in the current study are available via GitHub at https://github.com/robinvloeberghs/Confidence-and-criterion-fluctuations.

Code availability

The code for all analyses is available via GitHub at https://github.com/robinvloeberghs/Confidence-and-criterion-fluctuations.

References

1. Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics* Vol. 1 (Wiley, 1966).
2. Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).
3. Macmillan, N. A. & Creelman, C. D. *Detection Theory: A User's Guide* (Psychology Press, 2005).
4. Hebart, M. N., Schriever, Y., Donner, T. H. & Haynes, J.-D. The relationship between perceptual decision variables and confidence in the human brain. *Cereb. Cortex* **26**, 118–130 (2016).
5. Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227–231 (2008).
6. Shekhar, M. & Rahnev, D. The nature of metacognitive inefficiency in perceptual decision making. *Psychol. Rev.* **128**, 45–70 (2021).
7. Urai, A. E., Braun, A. & Donner, T. H. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nat. Commun.* **8**, 14637 (2017).

8. Xue, K., Shekhar, M. & Rahnev, D. Challenging the Bayesian confidence hypothesis in perceptual decision-making. *Proc. Natl Acad. Sci. USA* **121**, e2410487121 (2024).
9. Haddara, N. & Rahnev, D. The impact of feedback on perceptual decision-making and metacognition: reduction in bias but no change in sensitivity. *Psychol. Sci.* **33**, 259–275 (2022).
10. Sánchez-Fuenzalida, N., Van Gaal, S., Fleming, S. M., Haaf, J. M. & Fahrenfort, J. J. Predictions and rewards affect decision-making but not subjective experience. *Proc. Natl Acad. Sci. USA* **120**, e2220749120 (2023).
11. Kloosterman, N. A. et al. Humans strategically shift decision bias by flexibly adjusting sensory evidence accumulation. *Elife* **8**, e37321 (2019).
12. Bang, J. W. & Rahnev, D. Stimulus expectation alters decision criterion but not sensory signal in perceptual decision making. *Sci. Rep.* **7**, 17072 (2017).
13. Treisman, M. & Williams, T. C. A theory of criterion setting with an application to sequential dependencies. *Psychol. Rev.* **91**, 68–111 (1984).
14. Urai, A. E., de Gee, J. W., Tsetsos, K. & Donner, T. H. Choice history biases subsequent evidence accumulation. *Elife* **8**, e46331 (2019).
15. Frithsen, A., Kantner, J., Lopez, B. A. & Miller, M. B. Cross-task and cross-manipulation stability in shifting the decision criterion. *Memory* **26**, 653–663 (2018).
16. Ashwood, Z. C. et al. Mice alternate between discrete strategies during perceptual decision-making. *Nat. Neurosci.* **25**, 201–212 (2022).
17. Cowley, B. R. et al. Slow drift of neural activity as a signature of impulsivity in macaque visual and prefrontal cortex. *Neuron* **108**, 551–567.e8 (2020).
18. Mochol, G., Kiani, R. & Moreno-Bote, R. Prefrontal cortex represents heuristics that shape choice bias and its integration into future behavior. *Curr. Biol.* **31**, 1234–1244.e6 (2021).
19. Roy, N. A., Bak, J. H., Akrami, A., Brody, C. D. & Pillow, J. W. Extracting the dynamics of behavior in sensory decision-making experiments. *Neuron* **109**, 597–610.e6 (2021).
20. Vloeberghs, R., Urai, A. E., Desender, K. & Linderman, S. W. A Bayesian hierarchical model of trial-to-trial fluctuations in decision criterion. *PLoS Comput. Biol.* **21**, e1013291 (2025).
21. Urai, A. E. Structure uncovered: understanding temporal variability in perceptual decision-making. *Trends Cogn. Sci.* **30**, 54–65 (2026).
22. Gupta, D. & Brody, C. D. Limitations of a proposed correction for slow drifts in decision criterion. *Neurons Behav. Data Anal. Theory* **1**, 1–19 (2022).
23. Meijer, G. Neurons in the mouse brain correlate with cryptocurrency price: a cautionary tale. *Peer Community J.* **1**, e29 (2021).
24. Harris, K. D. Nonsense correlations in neuroscience. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.29.402719 (2020).
25. Kiani, R., Corthell, L. & Shadlen, M. N. Choice certainty is informed by both evidence and decision time. *Neuron* **84**, 1329–1342 (2014).
26. Lempert, K. M., Chen, Y. L. & Fleming, S. M. Relating pupil dilation and metacognitive confidence during auditory decision-making. *PLoS ONE* **10**, e0126588 (2015).
27. Boldt, A. & Yeung, N. Shared neural markers of decision confidence and error detection. *J. Neurosci.* **35**, 3478–3484 (2015).
28. Desender, K., Murphy, P., Boldt, A., Verguts, T. & Yeung, N. A Postdecisional neural marker of confidence predicts information-seeking in decision-making. *J. Neurosci.* **39**, 3309–3319 (2019).
29. Aston-Jones, G. & Cohen, J. D. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.* **28**, 403–450 (2005).
30. Reimer, J. et al. Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nat. Commun.* **7**, 13289 (2016).
31. De Martino, B., Fleming, S. M., Garrett, N. & Dolan, R. J. Confidence in value-based choice. *Nat. Neurosci.* **16**, 105–110 (2013).
32. Jang, Y., Wallsten, T. S. & Huber, D. E. A stochastic detection and retrieval model for the study of metacognition. *Psychol. Rev.* **119**, 186–200 (2012).
33. Mueller, S. T. & Weidemann, C. T. Decision noise: an explanation for observed violations of signal detection theory. *Psychon. Bull. Rev.* **15**, 465–494 (2008).
34. Fleming, S. M. & Lau, H. C. How to measure metacognition *Front. Hum. Neurosci.* https://doi.org/10.3389/fnhum.2014.00443 (2014).
35. Olawole-Scott, H. & Yon, D. Expectations about precision bias metacognition and awareness. *J. Exp. Psychol. Gen.* **152**, 2177–2189 (2023).
36. Rahnev, D. A robust confidence–accuracy dissociation via criterion attraction. *Neurosci. Conscious.* **7**, 1–11 (2021).
37. Rahnev, D. Confidence-accuracy dissociations in perceptual decision making. *Vision Res.* **244**, 108831 (2026).
38. Vaghi, M. M. et al. Compulsivity reveals a novel dissociation between action and confidence. *Neuron* **96**, 348–354.e4 (2017).
39. Van Marcke, H., Le Denmat, P., Verguts, T. & Desender, K. Manipulating prior beliefs causally induces under- and overconfidence. *Psychol. Sci.* **35**, 358–375 (2024).
40. Andreassen, E. R., Frith, C. D. & Yon, D. Public communication alters private confidence. *Curr. Biol.* **36**, 171–188 (2026).
41. Benwell, C. S. Y., Mohr, G., Wallberg, J., Kouadio, A. & Ince, R. A. A. Psychiatrically relevant signatures of domain-general decision-making and metacognition in the general population. *Npj Ment. Health Res.* **1**, 10 (2022).
42. Rouault, M., Seow, T., Gillan, C. M. & Fleming, S. M. Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biol. Psychiatry* **84**, 443–451 (2018).
43. Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M. & Lau, H. Confidence leak in perceptual decision making. *Psychol. Sci.* **26**, 1664–1680 (2015).
44. Samaha, J., Iemi, L. & Postle, B. R. Prestimulus alpha-band power biases visual discrimination confidence, but not accuracy. *Conscious. Cogn.* **54**, 47–55 (2017).
45. Ratcliff, R. & McKoon, G. The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput.* **20**, 873–922 (2008).
46. Vloeberghs, R., Tuerlinckx, F., Urai, A. E. & Desender, K. Explaining temporally clustered errors with an autocorrelated drift diffusion model. Preprint at *bioRxiv* https://doi.org/10.64898/2026.03.20.713186 (2026).
47. Clark, T. K., Yi, Y., Galvan-Garza, R. C., Bermúdez Rey, M. C. & Merfeld, D. M. When uncertain, does human self-motion decision-making fully utilize complete information? *J. Neurophysiol.* **119**, 1485–1496 (2018).
48. Murphy, K. P. *Probabilistic Machine Learning: Advanced Topics* (MIT Press, 2023).
49. R Core Team *R: A Language and Environment for Statistical Computing* (R Core Team, 2023).
50. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
51. Christensen, R. H. B. Cumulative link models for ordinal regression with the R package ordinal. *J. Stat. Softw.* **35**, 1–46 (2018).
52. Adler, W. T. & Ma, W. J. Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLoS Comput. Biol.* **14**, e1006572 (2018).

53. Denison, R. N., Adler, W. T., Carrasco, M. & Ma, W. J. Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proc. Natl Acad. Sci. USA* **115**, 1109011095 (2018).
54. Maniscalco, B., McCurdy, L. Y., Odegaard, B. & Lau, H. Limited cognitive resources explain a trade-off between perceptual and metacognitive vigilance. *J. Neurosci.* **37**, 1213–1224 (2017).
55. Orchard, E. R., Dakin, S. C. & Van Boxtel, J. J. Internal noise measures in coarse and fine motion direction discrimination tasks and the correlation with autism traits. *J. Vis*. https://doi.org/10.1167/jov.22.10.19 (2022).
56. Siedlecka, M., Koculak, M. & Paulewicz, B. Confidence in action: differences between perceived accuracy of decision and motor response. *Psychon. Bull. Review* https://doi.org/10.3758/s13423-021-01913-0 (2021).
57. Wang, S. Face size biases emotion judgment through eye movement. *Sci. Rep*. https://doi.org/10.1038/s41598-017-18741-9 (2018).

Acknowledgements

We thank A. Boldt for providing us with the data of Boldt and Yeung[27].

Author contributions

Conceptualization: R.V. and K.D. Formal analysis: R.V. and L.N.O. Funding acquisition: A.E.U. and K.D. Methodology: R.V., L.N.O. and K.D. Project administration: R.V. and L.N.O. Software: R.V. and L.N.O. Supervision: A.E.U. and K.D. Visualization: R.V. and L.N.O. Writing—original draft: R.V., L.N.O. and K.D. Writing—review and editing: R.V., A.E.U. and K.D.

Funding

This research was supported by a grant from the Research Foundation Flanders, Belgium (FWO-Vlaanderen No. G0A6N24N) and the Flemish government (Methusalem project METH/24/003), both awarded to K.D. and a Veni fellowship (VI.Veni.212.184) from the Netherlands Organization for Scientific Research awarded to A.E.U.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-026-02544-y.

Correspondence and requests for materials should be addressed to Robin Vloeberghs.

Peer review information *Nature Human Behaviour* thanks Reza Ebrahimpour and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Robin Vloeberghs

Last updated by author(s): Jun 26, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No new data was collected for the current study.
Data analysis	All statistical analyses were performed in R (4.6.0) using Rstudio (2026.05.0). The linear mixed effects models were fitted using the lme4 package (1.1.37). The ordinal mixed regression models were fitted with the ordinal package (2022.11.16). An open-source model in Python (hMFC) was used to measure criterion fluctuations (https://github.com/robinvloeberghs/hMFC)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The study used 15 openly available datasets from the Confidence database (Rahnev et al., 2021; https://osf.io/s46pr/). The raw and preprocessed datasets can be found on the GitHub repo: https://github.com/robinvloeberghs/Confidence-and-criterion-fluctuations.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	The current study did not collect new data but used 15 openly available datasets from the Confidence database (https://osf.io/s46pr/). The datasets do not contain information on age and gender. More information can be found in the original studies.
Reporting on race, ethnicity, or other socially relevant groupings	The datasets do not contain information on these variables.
Population characteristics	The datasets do not contain information on these variables. More information on age can be found in the original studies.
Recruitment	This differs for each study with different methods such as online requirement or credit-based requirement for students.
Ethics oversight	This information can be found in the original studies.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The data are of quantitative nature. Responses, reaction times, and confidence levels, together with stimulus information were recorded on each trial for each participant.
Research sample	This information can be found in the original studies.
Sampling strategy	This information can be found in the original studies.
Data collection	The experiments were shown on a computer.
Timing	For the exact timing the authors of each original study should be contacted.
Data exclusions	Subjects were excluded if they violated one of the preregistered exclusion criteria.
Non-participation	This information can be found in the original studies.
Randomization	Participants were not allocated in experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.