eLife

Reviewed Preprint
v1 • June 23, 2026
Not revised

For correspondence:
nicocomay@gmail.com

* Co-last authors

Funding: See page 28

Reviewing editor: Nathan Faivre, Centre National de la Recherche Scientifique, France

Confidence phenotypes: a unified computational account of value and decision certainty in reinforcement learning

Nicolás A Comay[a], **Guillermo Solovey**[b,c,d,*], **Pablo Barttfeld**[a,*]

[a]Cognitive Science Group. Instituto de Investigaciones Psicológicas (IIPsi, CONICET-UNC), Facultad de Psicología, Universidad Nacional de Córdoba, Córdoba, Argentina • [b]Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, UBA-CONICET, Buenos Aires, Argentina • [c]Laboratorio de Neurociencia, Universidad Torcuato Di Tella, Buenos Aires, Argentina • [d]Escuela de Negocios, Universidad Torcuato Di Tella, Buenos Aires, Argentina

eLife Assessment

This **valuable** study advances our understanding of confidence in reinforcement learning by considering value confidence and decision confidence within a common Bayesian computational framework. The evidence is **solid**, supported by converging analyses across multiple datasets, though the direct interaction between the two forms of confidence and the model identifiability requires further clarification. The work will be of primary interest to researchers in reinforcement learning, decision-making, and metacognition.

https://doi.org/10.7554/eLife.111820.1.sa3

Abstract

Confidence, the “feeling of knowing” that accompanies every cognitive process, plays a critical role in human reinforcement learning; yet its computational bases in learning scenarios have only recently begun to be studied. Prior work has distinguished between value confidence (certainty in value estimates) and decision confidence (certainty that a choice is correct), but how these two forms of confidence are computed and interact has not been directly tested. Here we combine two experiments and previously published datasets to test competing computational hypotheses. We find that value confidence is best explained by a Bayesian computation reflecting the precision of value estimates, and that it adaptively guides behaviour by reducing exploration and promoting exploitation as certainty increases. In contrast, decision confidence departs from Bayesian predictions, especially on errors. A hybrid model integrating the Bayesian probability of being correct with the overall value confidence better accounts for decision confidence. Moreover, individual differences in the relative weighting of these two information sources explain variation in confidence reports and predict both task performance and metacognitive accuracy: subjects whose confidence judgments more closely track Bayesian computations perform better. Together, these results provide a unified computational mechanism through which distinct forms of confidence shape learning and choices in uncertain environments.

Introduction

Humans routinely face uncertainty, about the state of the world, the outcomes of their actions, and the intentions of others. To navigate this uncertainty, the brain constructs internal beliefs and attaches to them a sense of confidence: a graded estimate of how reliable those beliefs are (Meyniel, Sigman, et al., 2015). Confidence, in this broad sense, permeates virtually every domain of cognition. For instance, it shapes how we learn from feedback (Meyniel, Schlunegger, et

al., 2015), monitor and adjust decisions (Yeung & Summerfield, 2012), seek information (Desender et al., 2018), and coordinate with others (Bang et al., 2017). Its pervasive influence has led to the view that confidence is a central component of behavioural control (Schulz et al., 2023), and a key construct in transdiagnostic models of mental health (Hoven et al., 2019).

Despite its ubiquity, confidence is typically treated as a summary of how certain the brain is about its own states. This assumption has been highly productive, primarily through the use of perceptual paradigms and the study of decision confidence—the subjective probability of having made a correct choice (Figure 1; Fleming, 2024; Pouget et al., 2016). In this framework, confidence can bias subsequent choices (Lisi et al., 2020) or determine when to stop accumulating evidence (Balsdon et al., 2020; Balsdon & Philiastides, 2024), yet it is usually conceived as a more reflective quantity—a metacognitive evaluation of one's accuracy—rather than as a variable exerting a direct or instrumental influence on the decision process. In contrast, research in human reinforcement learning (RL) has centered on value confidence—the confidence or certainty in the values or expected outcomes associated with available options (Figure 1)—which plays a pivotal role in adaptive behavior by, for instance, governing the rate of belief updating, modulating the exploration-exploitation trade-off, and determining how new information is weighted against prior expectations (Behrens et al., 2007; Boldt et al., 2019; Nassar et al., 2010; Payzan-LeNestour & Bossaerts, 2011).

The contrast between the approaches used in perceptual and reinforcement-learning paradigms exposes a fundamental tension in how these two traditions conceptualize confidence: while perceptual models define it as a reflective readout of certainty, learning models reveal it as a generative signal that drives behaviour. How these roles coexist within a single computational architecture remains unknown—especially given that confidence itself is highly idiosyncratic (Navajas et al., 2017). Indeed, while individuals tend to report confidence consistently across time (Ais et al., 2016), their mappings between internal uncertainty and reported confidence vary widely across the population, suggesting that distinct computational strategies may underlie how uncertainty is evaluated and used to guide behaviour. Only a handful of studies have begun to address these questions in human RL (Boldt et al., 2019; Salem-Garcia et al., 2023; Ting et al., 2023), leaving open how the brain integrates confidence across the representational hierarchy—from beliefs about the world to confidence in the choices derived from them—and how it shapes decisions under uncertainty across individuals.

Here, we develop a unified framework that distinguishes multiple computational forms of confidence within reinforcement learning. Using new and previously published datasets, we show that value confidence is captured by the precision of Bayesian value estimates, whereas decision confidence reflects a hybrid computation combining probability of being correct with a more global estimate of value certainty. The relative weighting of these components characterises individual computational profiles—'confidence phenotypes'—that predict performance, exploration strategies, and metacognitive accuracy. Together, these findings suggest that confidence is best understood not as a single scalar quantity, but as a set of related computations that jointly regulate learning and choice under uncertainty.

Results

Value confidence emerges from Bayesian inference

We began by modeling value confidence, i.e.: the confidence on the estimated values of the options. In the value confidence task from Boldt et al. (2019), on each trial participants (N=21) observed a reward randomly sampled from one of two arm-bandits. Then, they reported their belief on their mean value of that option and their confidence on this estimation. We tested several models to fit these confidence ratings: one model set was based on extensions of the Rescorla-Wagner (RW) algorithms, and the others were based on Bayesian inference (see Methods for details). Figure 2a depicts model fitting results of the best two models to this dataset (see Supplementary Figure 1 for the predictions of all tested models). We found that value confidence ratings were best explained by a Bayesian model (Model frequency (p_{model}) = 0.60,

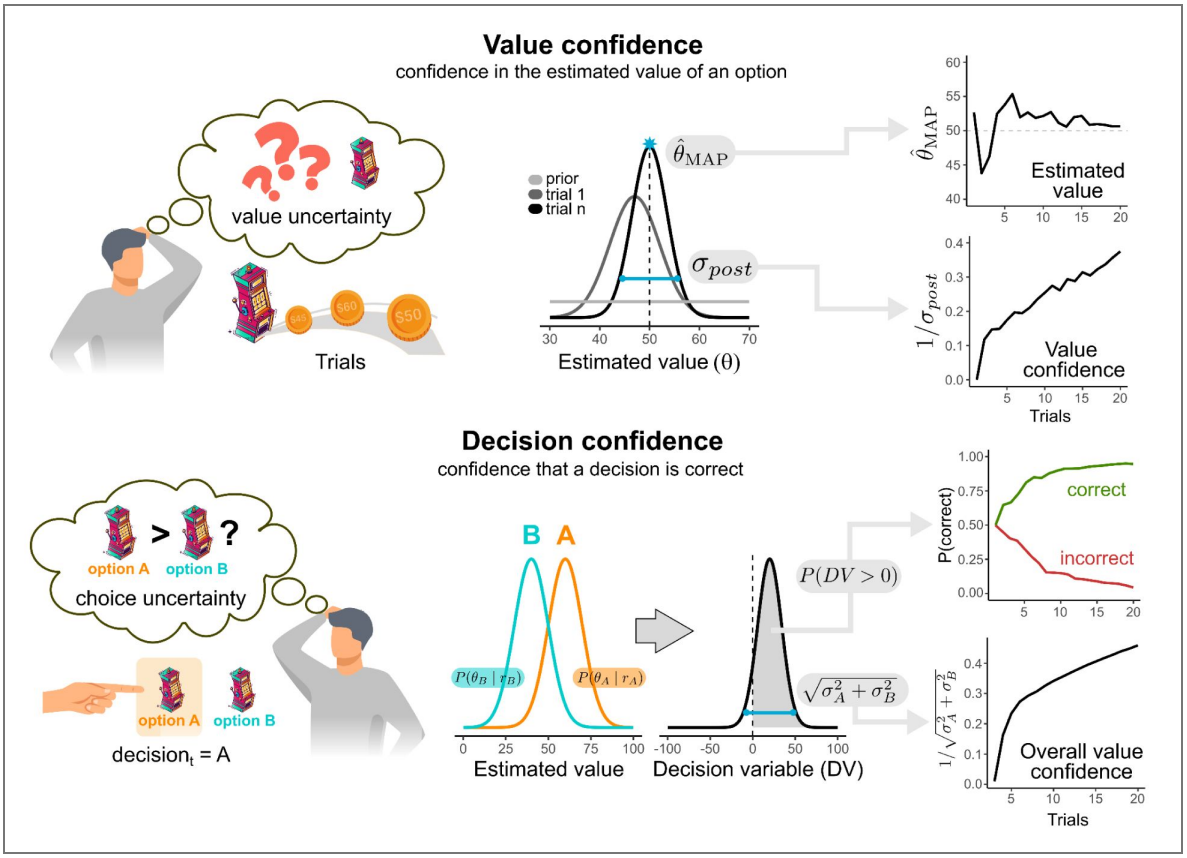


Figure 1. Different levels of confidence in reinforcement learning contexts.

Value confidence (top) represents the certainty about latent states of the environment, such as the value of one option in multi-armed bandit tasks. Under a Bayesian framework, value confidence can be straightforwardly modelled as the inverse of the standard deviation of a posterior distribution over the option's estimated value (represented by 0 in the figure). By observing subsequent rewards, not only the estimation of the option's value gets more precise but the posterior shrinks, which is reflected in an increase in value confidence. Decision confidence (bottom) reflects the certainty of having made a correct choice. Under a Bayesian formulation, it can be understood as the probability of being correct, illustrated here by the area under the decision variable (the distribution obtained by subtracting the unchosen option's posterior from the chosen option's posterior) that it is greater than zero. As trials progress and certainty about the environment increases, decision confidence better distinguishes between correct and incorrect responses. This increase in correct trials and decrease in incorrect trials is known as the "folded-x pattern" of confidence (Hangya et al., 2016; Sanders et al., 2016).

exceedance probability ($p_{exc.}$) > .99, protected exceedance probability ($p_{p.exc.}$) > .99; Figure 2c), in which value confidence reflected the inverse of the standard deviation of the posterior distribution over the estimated value of the option at play.

To test whether the Bayesian account of value confidence generalizes beyond the specific conditions of the Boldt et al. (2019) task, we next applied our models to data published by to Quandt et al. (2022) (N=62 for their Experiment 1; N=60 for their Experiment 2). In their task, participants saw a sequence of a hundred rewards from one option in rapid presentation and then, as in Boldt et al. (2019), they had to report their estimated mean value of the option and their confidence on this estimate. A key manipulation was present in this dataset, namely that the variance of the reward distributions were different across alternatives, thus generating different levels of certainty across options which significantly affected value confidence judgments. This key manipulation allowed us to even better differentiate between our models, as some of the candidate models do not take the variance of the reward distributions into account, thus predicting a constant level of confidence regardless of the variance of the rewards (indeed, the inclusion of this dataset improved model recovery results, see Supplementary Figure 2). Specifically, we tested three models on Quandt et al. (2022) data: the mentioned Bayesian model, a RW model where value confidence reflected the square root of the number of rewards seen of the particular option at play (as it was the best non-Bayesian model in Boldt et al. dataset; also note that, in this dataset, this model make the same predictions as the other models that were a function of the number of rewards seen) and a RW model with a separate learning algorithm that tracked the variance of the rewards (as this one was the only RW model that had information about the variance of the rewards). Note that we did not include any of the models which involved a surprise term (nor the model where value confidence explicitly reflected the surprise of the reward seen, see Methods) as these models could not account for the data in the Boldt et al. (2019) dataset. We again found that the Bayesian model was the best fitting model (Exp. 1: p_{model} = .94; $p_{exc.}$ > .99; $p_{p.exc.}$ > .99; Exp. 2: p_{model} = .96; $p_{exc.}$ > .99; $p_{p.exc.}$ > .99; Figure 2b and Figure 2d), as it was able to capture the decrease in value confidence levels as the standard deviation of the options reward distributions increased (Figure 2b).

Value confidence modulates the exploration-exploitation trade-off

A central functional role proposed for confidence is to regulate the balance between exploration and exploitation during learning (Boldt et al., 2019). Indeed, it is reasonable for an agent learning from the environment to increase exploitation as it becomes more certain about the values of the options at play. To test whether value confidence serves this adaptive control function, we extended the Bayesian model to include a parameter, b_1, that modulates decision noise as a function of value confidence (see Methods section and Figure 3a). A positive b_1 indicates that decisions become more deterministic—that is, more exploitative—as value confidence increases. For testing this idea, we leveraged on the decision data from Boldt et al. (2019) experiment 2 study (N=30), and in two new studies conducted in our laboratory (N=29 and N=30; the latter a pre-registered replication). In all cases, participants performed a classic two-armed bandit task in which they had to choose between two options and rate their confidence on having selected the best one. After that, they saw the reward obtained.

To assess whether value confidence modulates the exploration-exploitation balance, we compared the Bayesian model including the the b_1 parameter against the same model not including the parameter i.e., a model with constant decision noise. We found that the model including b_1 better explained the decision data in all datasets (Boldt et al. dataset: p_{model} = .96; $p_{exc.}$ > .99; $p_{p.exc.}$ > .99; Study 1: p_{model} = .82; $p_{exc.}$ > .99; $P_{p.exc.}$ = .98; Study 2: p_{model} = .85; $p_{exc.}$ > .99; $p_{p.exc.}$ > .99). Note that the main divergence between models emerged toward the end of each block, in line with the increase of exploitation as value confidence accumulates (Figure 3b, first and second rows). We also found that the b_1 parameter was significantly greater than zero in the three datasets employed (Boldt et al. dataset: t_{29} = 8.76; p < .001, d = 1.60; Study 1: t_{28} = 3.95; p < .001, d = 0.73; Study 2: t_{29} = 5.58; p < .001, d = 1.02; Figure 3b, third row), further validating the proposed mechanism.

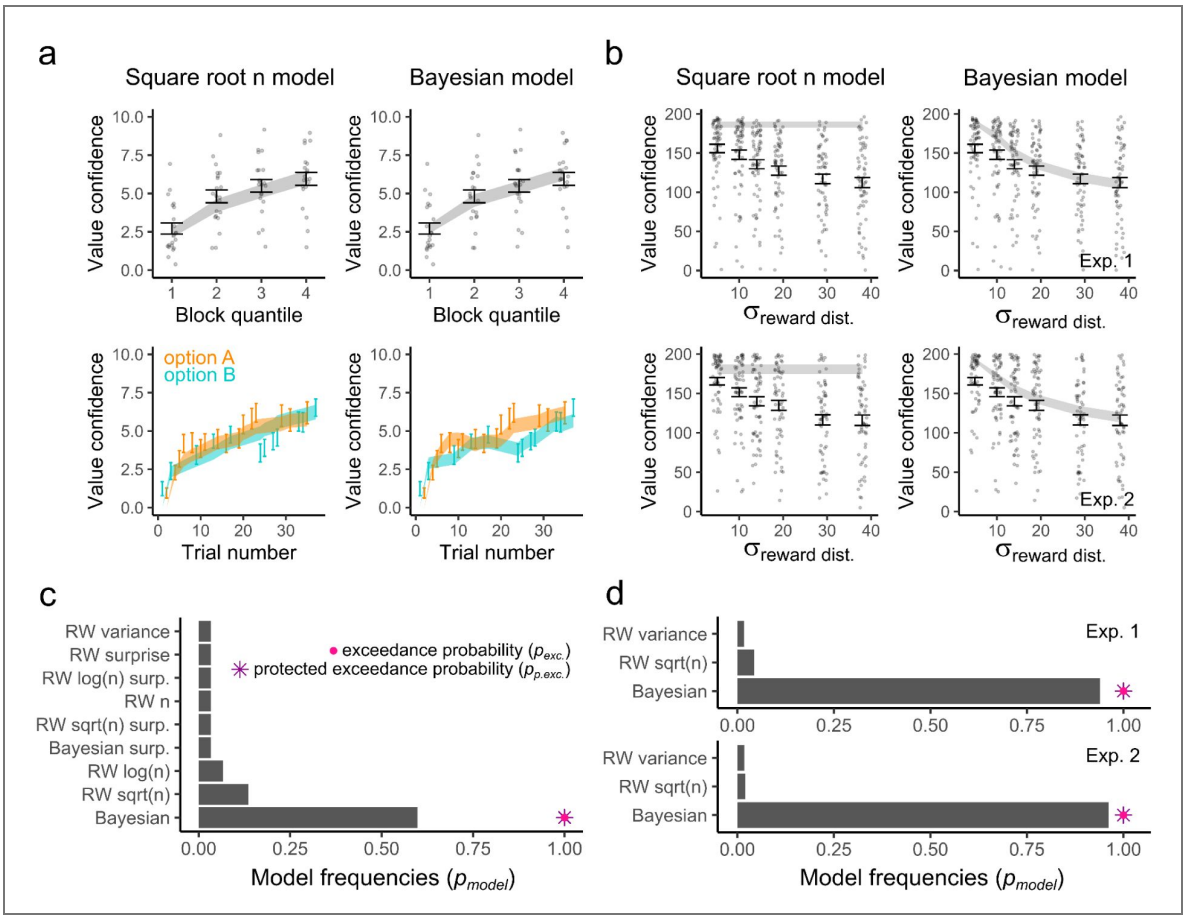


Figure 2. Value confidence is captured by a Bayesian model.

(a) Model fitting results for the two best models in the Boldt et al. (2019) data: the Bayesian model, where value confidence reflects the inverse of the standard deviation of the posterior distribution over an option's inferred value, and the Rescorla-Wagner model, where value confidence equals to the square root of the number of rewards observed of a specific option. First row: models' fits to all the data. As blocks had different lengths, we divided blocks in four quantiles with respect to the trials to be able to pool all blocks together (x axis). Second row: example of models' fits in one block only. (b) Model fitting results to the Quandt et al. (2022) data. First row: models' fits to Exp. 1 data. Second row: models' fits to Exp. 2 data. Models that are a function of the number of rewards experienced (such as the RW sqrt(n) model) cannot account for the negative effect that the spread of the reward distribution has on value confidence. (c) Model comparison results. The Bayesian model was the best fitting model in Boldt et al. (2019) data. (d) Model comparison results. The Bayesian model was the best fitting model in both experiments in Quandt et al. (2022) data. In panels (a) and (b) error bars represent the standard error of the mean (SEM) of the behavioral data, shaded regions represent the SEM of models' predictions and dots represent individual averages.

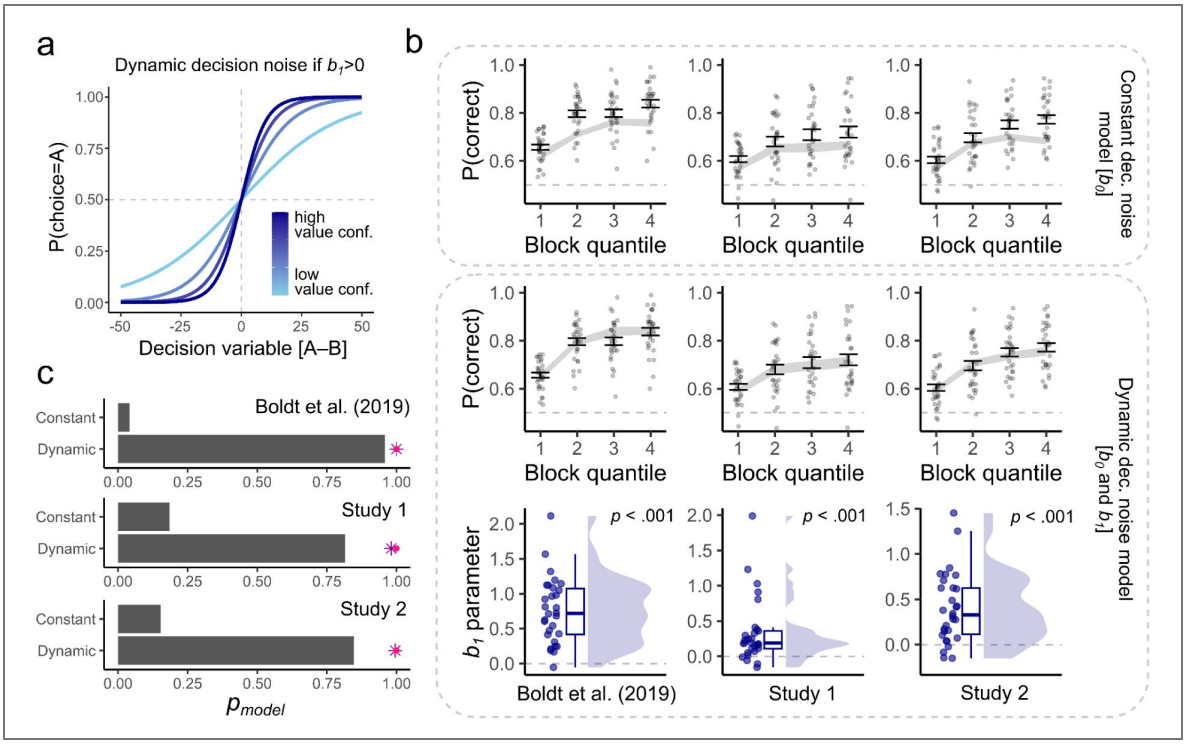


Figure 3. Value confidence modulates the exploration-exploitation trade-off.

(a) Illustration of an agent's behaviour if the b_1 parameter is positive. In such a case, as value confidence increases decision noise decreases because the slope of the softmax increases. This results in more exploitative decisions as value confidence increases throughout the trials. (b) Model fitting results of the models without (first row) and with (second row) dynamic decision noise (i.e.: without and with the b_1 parameter). Dashed lines in these two rows represent chance level performance (i.e.: a proportion of correct choices equal to .5). The third row depicts the distribution of the b_1 parameter across the three datasets analysed. Regarding the columns, the first column represents Boldt et al. data, the second column the Study 1 data and the third column the Study 2 data. (c) Model comparison results. The model that includes the modulation of decision noise by value confidence was the best fitting model across the three datasets. The conventions in this figure are equal to the ones in Figure 2.

Decision confidence deviates from the probability of being correct

Having established that value confidence and choice behavior are well described by Bayesian inference, we next evaluated whether the same Bayesian model could also explain *decision confidence*—the subjective certainty that a chosen option is correct. According to the Bayesian confidence hypothesis (Meyniel, Sigman, et al., 2015; Sanders et al., 2016), confidence should reflect the probability of being correct. Interestingly, while this model was able to capture the pattern of confidence in correct trials (r_{354} = 0.87; p < .001), it failed to account for confidence in incorrect trials (r_{354} = 0.74; p < .001; Figure 4b and Figure 4c, top row). Indeed, the Bayesian model predicts that confidence should progressively increase across trials for correct responses but decrease for the incorrect ones, a signature known as the "folded-x pattern" of Bayesian confidence (Hangya et al., 2016; Sanders et al., 2016). Inspired by Navajas et al. (2017), in which a similar deviation of Bayesian confidence in incorrect trials was found but in categorical decisions, we constructed an Bayesian-hybrid model, where confidence reflects a weighted combination of the probability of being correct (i.e.: the Bayesian confidence computation) and the overall certainty of value estimates (i.e.: the overall value confidence). This model was able to capture decision confidence data both in correct (r_{354} = 0.93; p < .001) and incorrect trials (r_{354} = 0.90; p < .001; Figure 4b and Figure 4c, bottom row), and provided the best fit in two out of the three evaluated datasets (Boldt et al.: p_{model} = 0.5; $p_{exc.}$ = 0.5; $p_{p.exc.}$ = 0.5; Study 1: p_{model} = 0.9; $p_{exc.}$ > 0.99; $p_{p.exc.}$ > 0.99; Study 2: p_{model} = 0.81; $p_{exc.}$ > 0.99; $p_{p.exc.}$ > 0.99; Figure 4d). Across the three studies, in 65 out of 89 participants the hybrid model was preferred. We also evaluated a family of models that ignore option uncertainty and rely solely on estimated option values (as in Desender & Verguts, 2024; Salem-Garcia et al., 2023). A confidence model incorporating chosen option's value emerged as a reasonable alternative, but it failed to capture behaviour in bandit tasks where both the mean and variance of reward distributions are manipulated (see Methods & Supplementary material).

Individual signatures of decision confidence reveal distinct behavioural profiles

Having established a population-level deviation from Bayesian confidence, we next examined whether individuals differ systematically in how they combine the probability of being correct and overall value confidence when reporting confidence. Following Navajas et al. (2017) we ran ordered linear regressions for each participant predicting decision confidence using the mentioned latent variables from the Bayesian-hybrid model. This yielded two regression coefficients—one for p(correct), $\beta_{p(correct)}$, and one for overall value confidence, $\beta_{value\ conf.}$—quantifying each individual's reliance on these sources of information (Fig. 5a).

Importantly, participants assigned significantly greater weight to the probability of being correct than to overall value confidence (paired t-test among beta values: t_{88} = 9.37, p < .001, d = 0.99), indicating that Bayesian estimates of being correct constituted the primary driver of confidence reports, whereas value confidence contributed comparatively less. This dominance is particularly important given that the overall value confidence does not necessarily track the discriminability between the options (i.e.: the difficulty of a decision). Consequently, relying more heavily on this signal could, in principle, give rise to confidence patterns that are misaligned with decision difficulty. By contrast, simulations showed that when probability of being correct is weighted more strongly—as in the empirical data—the model yields well-calibrated confidence, with higher confidence for easier decisions even when overall uncertainty remains high (see Supplementary Material).

We then asked whether these computational weights were related to task and metacognitive performances. First, we ran a linear regression predicting task performance using both $\beta_{p(correct)}$ and $\beta_{value\ conf.}$ as well as their interaction. We found that higher $\beta_{p(correct)}$ values were predictive of higher performance (β = 0.053; p < .001; Figure 5b, left), but $\beta_{value\ conf.}$ values were not (β = −0.034; p = .233; Figure 5b, right). No interaction was found between the two predictors (β = −0.009; p = .538). Second, using a linear regression again, we evaluated whether these beta values

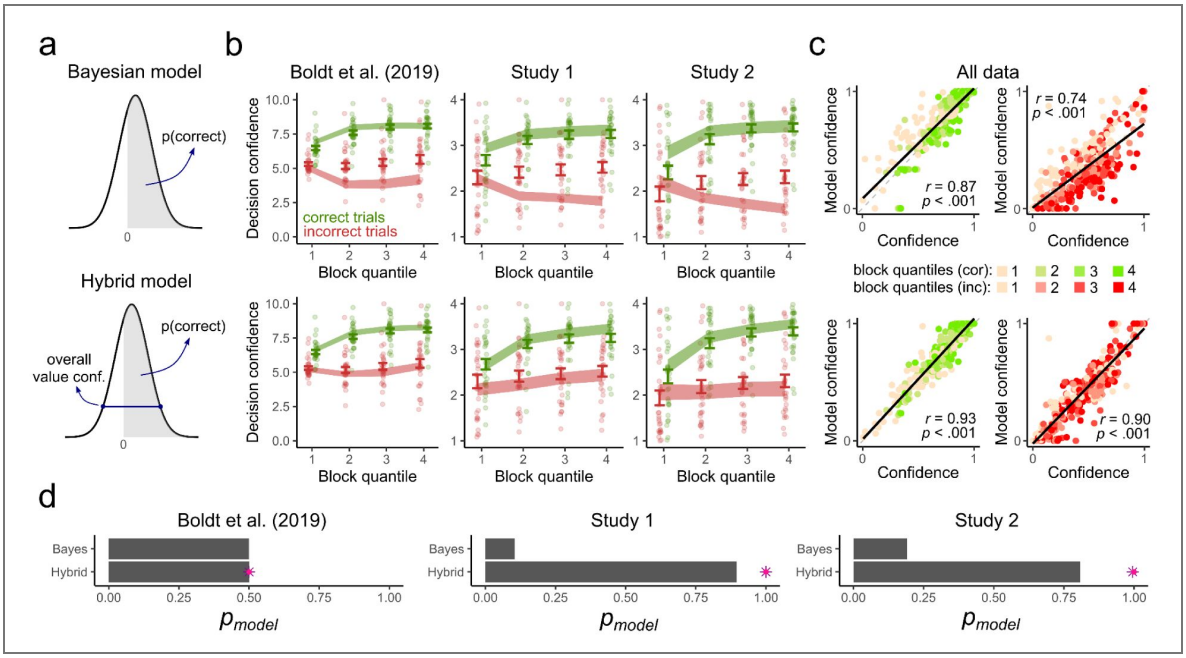


Figure 4. Decision confidence deviates from the probability of being correct.

(a) Schematic representation of the models. The subtraction between the two options' expected values posterior distributions (i.e.: chosen posterior - unchosen posterior) creates a decision variable. In such a case, the probability of being correct is the area of this distribution that is greater than zero. This probability of being correct represents the Bayesian confidence (top panel). The Bayesian-hybrid model (lower panel) combines this probability with the overall certainty about value estimates (i.e.: the overall value confidence). (b) Model fitting results. The Bayesian model can capture correct trials but its prediction deviates from the data specially in incorrect trials (top-row). The Bayesian-hybrid model, on the other hand, can account for both correct and incorrect trials (bottom-row). (c) Correlations between models' predictions and the data. Both models' predicted confidence and empirical confidence data were rescaled to a range from 0 to 1 before computing the correlations (as confidence scales differed between datasets). The same pattern is found: the Bayesian model predictions deviate from the data specially in incorrect trials, where the correlation is considerably diminished. (d) Model comparison results. The hybrid model was the best fitting model in two out of three datasets. Considering all the data, the hybrid model was the preferred model for 65 out of 89 participants.

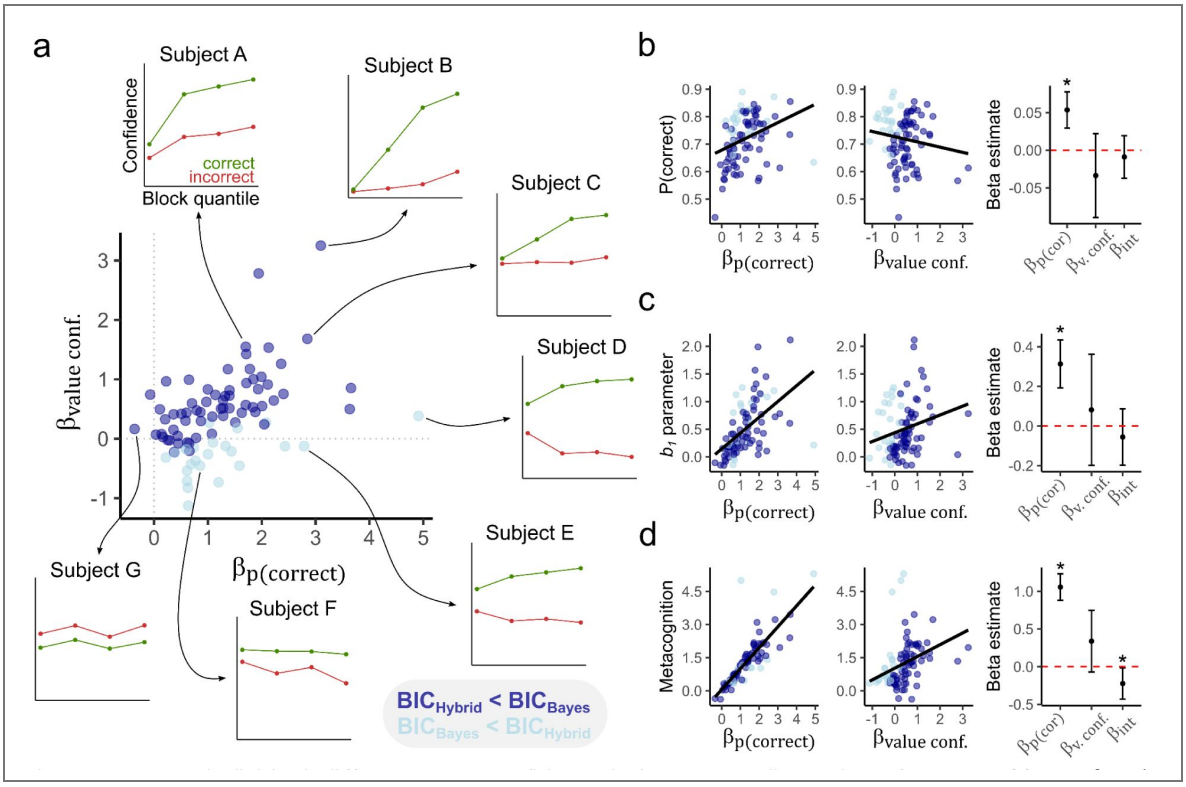


Figure 5. Interindividual differences on confidence judgments predict task performance.

(a) We found substantial individual variation in the way decision confidence was reported. By running a regression using the latent variables from the model, that is the probability of being correct and the overall value confidence, we were able to quantify this variation. Here we plot the beta values associated with the mentioned latent variables: $\beta_{p(correct)}$ and $\beta_{value\ conf.}$. Participants that had both betas with positive values showed a pattern where confidence tended to increase in both correct and incorrect decisions (Participant A and Participant B), or not decrease in incorrect decisions (Participant C). Participants with near zero beta $\beta_{value\ conf.}$ values had a more Bayesian style of confidence reporting (illustrated by Participant D and Participant E). Interestingly, several participants had negative $\beta_{value\ conf.}$ values. At first value, this is surprising as an increase in certainty should intuitively lead to greater confidence. However, the Bayesian model was the best model for virtually all of these participants (lightblue dots; note for instance the pattern of confidence judgments for Participant E), pointing out that the overall value confidence variable here should not be contributing and therefore its negative value suggests overfitting of the regression model. Note, however, that for some of these participants, like Participant F, a negative value appears to be justified as indeed her confidence decreased throughout the trials, that is, as certainty increases. Finally, some participants had negative $\beta_{p(correct)}$ values, which consequently led to a poor metacognitive ability as incorrect decisions were associated with higher confidence (Participant G). (b) Increasing values of the $\beta_{p(correct)}$ predicted higher task performance. (c) As with task performance, the $\beta_{p(correct)}$ values predicted the values of the b_1 parameter (the parameter that controlled the exploration-exploitation trade-off with increasing value confidence levels). (d) As in the b_1 parameter case, only $\beta_{p(correct)}$ values were associated with the metacognitive ability of the participants, that is, their ability to distinguish their correct and incorrect decisions with their confidence judgments. Note, however, that in this case a negative interaction between the two betas were found, meaning that increasing values of $\beta_{value\ conf.}$ negatively affected the ability of $\beta_{p(correct)}$ to predict metacognition.

were associated with the b_1 parameter values (i.e., the parameter that controlled the modulation by value confidence of the exploration-exploitation trade-off). We found that higher $\beta_{p(correct)}$ values were associated with higher b_1 parameter values ($\beta = 0.313$; $p < .001$; Figure 5c, left), while $\beta_{value\ conf.}$ values were not ($\beta = 0.082$; $p = .564$; Figure 5c, right), and no interaction was found between the two predictors ($\beta = -0.056$; $p = .442$). Finally, for evaluating metacognition—the ability to distinguish between correct and incorrect decisions—we run logistic regressions on each participant predicting accuracy from confidence judgments. We found that, as expected, larger $\beta_{p(correct)}$ values were associated with higher metacognition ($\beta = 1.056$; $p < .001$; Figure 5d, left), but no relationship was found between metacognition and $\beta_{value\ conf.}$ ($\beta = 0.338$; $p = .104$; Figure 5d, right). This time, however, a negative interaction was found, suggesting that at higher values of $\beta_{value\ conf.}$ the positive effect of $\beta_{p(correct)}$ diminished ($\beta = -0.224$; $p = .035$). Together, these results indicate that variability in confidence computation reflects distinct behavioral phenotypes rather than mere idiosyncrasies in reporting, suggesting that Bayesian estimates of correctness provide the dominant normative signal for confidence, while reliance on value confidence may reflect an additional, more idiosyncratic component that biases confidence reports away from this primary source. Indeed, participants who relied more heavily on Bayesian computations of confidence tended to make more accurate decisions (Figure 5b), used their value confidence more effectively to guide exploitative choices (Figure 5c), and showed greater accuracy in evaluating the correctness of their decisions (Figure 5d). In contrast, greater reliance on value confidence computation proved suboptimal: it impaired metacognitive accuracy while exerting no detectable influence on performance exploration-exploitation strategies.

Discussion

Overall, our results reveal how value and decision confidence arise and interact in human reinforcement learning. By quantifying inter-individual differences in these two forms of confidence computations we uncovered different behavioral patterns, where participants whose confidence computations more closely matched the Bayesian model showed higher task performance and metacognitive accuracy—in line with normative principles of decision-making.

Value confidence—the certainty in the estimated values of available options—was best captured by Bayesian computations, reflecting the precision of a posterior distribution over those values. This aligns with several findings in the literature where Bayesian models capture human learning remarkably well (Bounmy et al., 2023; Gershman, 2018; Kang et al., 2024; Meyniel, 2020; Meyniel, Schlunegger, et al., 2015). Our work further strengthens this notion by validating the predictions of these models not only against choice behavior but also against participants' explicit value confidence ratings, providing a more direct test of the model's internal variables. It should be noted, however, that human reinforcement learning is not always strictly normative. Perhaps the most prominent deviations from normative behaviour in human reinforcement learning are the positivity and the confirmation biases, weighting new evidence more heavily when it is associated with a positive valence or consistent with prior beliefs (Chambon et al., 2020; Lefebvre et al., 2017; Palminteri et al., 2017). These effects are usually modelled using separate learning rates for positive or confirmatory evidence, allowing for a greater update of positive and confirmatory information (Palminteri & Lebreton, 2022). Given that we focused on the computations underlying the uncertainty around the estimated values—rather than the estimated values per se—it remains to be tested whether including separate learning rates would further improve model fits. Interestingly, recent work has proposed that these apparent biases can themselves emerge under Bayesian principles (Godara, 2025)—which may align with their evolutionary value (Hoxha et al., 2025)—therefore being a possibility that Bayesian updating can suffice to explain these phenomena.

Extending the Bayesian framework to decision data, we found that value confidence acts as a key variable modulating the exploration-exploitation trade-off. Indeed, as certainty increased over trials, decision noise systematically decreased, resulting in more exploitative decisions (see Desender & Verguts, 2024, for a similar result but using decision confidence rather than value confidence as a modulatory signal). This aligns closely with Thompson sampling algorithms where

an agent explores more if uncertainty is higher (Gershman, 2018). It should be noted, however, that human behavior sometimes departs from this principle. In directed exploration cases, for instance, humans tend to deliberately select more uncertain options (Abir et al., 2024; Gershman, 2018). For other choice scenarios where direct and random exploration are likely to be at play, hybrid algorithms that combine directed and random exploration better explain human behavior (Gershman, 2018; E. Schulz & Gershman, 2019).

Although our model does not aim to capture all possible effects of (un)certainty on learning and decision-making, it highlights how Bayesian frameworks—by naturally representing uncertainty through precision metrics—can be extended to encompass a wide range of certainty-driven influences on behavior. For instance, Bayesian models have been used to explain how balancing the approach and avoidance of uncertainty can reduce the cognitive costs of exploration in a resource-rational manner (Abir et al., 2024), and how uncertainty modulates the use of simple decision heuristics that imperfectly exploit immediate rewards (Paunov et al., 2024). It is also worth mentioning that, outside human RL and Bayesian modelling, work in value-based decisions has shown that, by fitting certainty-informed drift-diffusion models to human choices, certainty about options' values modulates not only choices but response times (Lee & Usher, 2023), and that confidence provides a benefit signal for allocating cognitive resources during decision formation (Bénon et al., 2024; Lee & Daunizeau, 2021). Including such models that account for decision-level dynamics as a decision rule in RL contexts could provide even more insight into the mechanisms by which confidence guides decisions in uncertain environments.

Decision confidence, on the other hand, deviated from the Bayesian computations that accounted for value confidence ratings and decisions. This is inline with previous work showing deviations of normative principles in decision confidence in the perceptual domain (Comay et al., 2023; Li & Ma, 2020; Lisi et al., 2020; Miyoshi & Sakamoto, 2025; Xue et al., 2024) as well as in reinforcement learning tasks (Salem-Garcia et al., 2023; Ting et al., 2023). Specifically, confidence departed from Bayesian predictions specially on incorrect trials, violating the well-known "folded-x pattern" of confidence: with increasing evidence, confidence should increase for correct trials but decrease for incorrect ones (Hangya et al., 2016; Sanders et al., 2016)—although it should be noted that this pattern is not always a prediction of normative models of confidence (Adler & Ma, 2018; Rausch & Zehetleitner, 2019). Decision confidence was better captured by a weighted combination of two sources: the (Bayesian) probability of being correct and the certainty of the value estimates (which reflects the overall level of value confidence). Following the work of Navajas et al. (2017) in categorical decisions, we showed that the relative contributions of these two sources of information characterized the inter-individual patterns of decision confidence, reinforcing the notion that confidence is constructed in an idiosyncratic manner.

An important clarification concerns the computational role of overall value confidence in the model. This quantity does not uniquely determine option discriminability and therefore does not map directly onto decision difficulty. For example, a choice may be easy when the posterior value distributions of two options are well separated, even if both value estimates remain highly uncertain. Conversely, a choice may be maximally difficult when the two options have the same value, even if those values are known with high certainty. If confidence relied heavily on overall value confidence alone, this dissociation could produce counterintuitive patterns. For instance, higher confidence in the objectively harder case. However, both the empirical data and our simulations show that the probability of being correct is the dominant determinant of confidence, ensuring that confidence remains appropriately calibrated to decision difficulty (Supplementary Figure 6). In this framework, overall value confidence is better understood as a secondary biasing signal that modulates confidence without overriding the primary Bayesian estimate. This component is nevertheless important: it captures idiosyncratic individual differences and helps explain deviations from normative confidence, particularly on incorrect trials, where reliance on global uncertainty can produce systematic departures from the expected relationship between evidence and confidence.

Furthermore, the idea that the certainty about the options' values has an influence in decision confidence has a parallelism in influential models of confidence in perceptual decisions where stimulus reliability plays a key role in confidence computations (Boldt et al., 2017; Hellmann et al., 2023; Rausch et al., 2018; Shekhar & Rahnev, 2024) pointing to a possible cross-domains mechanism (although see Brus et al., 2021; Quandt et al., 2022). Extending this view, recent work has proposed that confidence itself reflects a noisy estimate of decision reliability constrained by a higher-order "meta-uncertainty" about the precision of the decision variable (Boundy-Singer et al., 2022). Incorporating such second-order uncertainty into learning frameworks could provide a fruitful direction for future research, offering a computational account of how confidence variability arises not only from task-related uncertainty but also from uncertainty about one's own inferential precision.

Importantly, these idiosyncratic patterns were not merely descriptive variations in confidence computation, but reflected distinct behavioral profiles. Participants whose confidence computations more closely followed the Bayesian ideal—weighting the probability of being correct more heavily—achieved higher overall accuracy, relied more strongly on value confidence to regulate their exploration-exploitation balance, and exhibited superior metacognitive insight into their own decisions. This suggests that the extent to which individuals approximate Bayesian confidence principles is not just a matter of internal calibration, but a signature of more adaptive learning and decision-making strategies. In this line, given that alterations in confidence are predictable of symptoms across multiple psychiatric dimensions (Hoven et al., 2019, 2023), these results may point to a continuum between optimal, Bayesian-like confidence computations and the kinds of suboptimality that characterize clinical populations.

Beyond idiosyncratic biases, recent work in human RL has shown that confidence judgments are also shaped by systematic biases (Salem-Garcia et al., 2023). Humans tend to overestimate their accuracy—the well-known overconfidence bias (Baranski & Petrusic, 1994)—and to report higher confidence when seeking gains than when avoiding losses, a phenomenon termed the valence-induced confidence bias (Lebreton et al., 2018). These effects can be formally accounted for by an overweighting of the learned value of the chosen option in the computation of confidence, suggesting that confidence judgments partially inherit biases originating in the learning process. Moreover, individual differences in the learning parameters underlying these value biases—such as confirmatory updating and outcome-context dependency—predict the magnitude of metacognitive biases, establishing a computational bridge between biased learning and biased confidence (Salem-Garcia et al., 2023). While the Bayesian model proposed here does not explicitly capture such value- and valence-related biases (although see Godara, 2025), extended Bayesian models—incorporating for instance reward-context dependencies—could provide a fruitful avenue for future research, offering a unified computational account of how learned values, value certainty, and decision confidence jointly give rise to both adaptive and biased behavior.

Finally, given that this recent research has modelled decision confidence using estimated option values but, unlike our approach, without incorporating uncertainty in those estimates (Desender & Verguts, 2024; Salem-Garcia et al., 2023), we therefore tested this kind of models on our datasets, as they offer a plausible alternative account of confidence behaviour. Consistent with these studies, a model driven by the value of the chosen option—akin to the "positive evidence bias" widely reported in perceptual decision-making (Maniscalco et al., 2016; Peters et al., 2017; Zylberberg et al., 2012)—provided a reasonable fit (see Supplementary material for model fitting results). However, it deviated from human behaviour mainly on incorrect trials, similarly to the Bayesian model (although over-estimating rather than under-estimating confidence in those trials). In addition, this model struggled to explain confidence patterns in a bandit task where both the mean and the variance of the reward distributions were manipulated (see Supplementary Material). Together, these findings support the idea that uncertainty in value estimates plays a central role in shaping confidence in learning settings. More broadly, we

believe that future research could exploit systematic manipulations of reward distributions to further dissociate the respective contributions of estimated uncertainty and chosen-option value to decision confidence.

By linking Bayesian principles of learning with metacognitive evaluations of choice, our results bridge two traditionally separate literatures—on reinforcement learning and decision confidence—into a unified account of behavior under uncertainty. Beyond their theoretical relevance, these insights may inform how confidence guides learning and exploration across domains, and how its miscalibration contributes to suboptimal decision-making strategies.

Methods

All data & scripts to reproduce the reported results, as well as the Supplementary material, are available online at: *osf.io/8tex5*.

Datasets description

Here we describe the nature of the datasets employed in our study. For datasets that have been previously published, we refer the reader to the original publications for a more detailed description of the tasks employed. For the newly collected datasets, participants were university students who provided informed consent prior to participation. These new experimental protocols were approved by the Ethics Committee of the Instituto de Investigaciones Psicológicas (UNC-CONICET).

Value confidence

For the value confidence analysis, we employed datasets from two previous studies. First, we used the data from the experiment 1 of Boldt et al. (2019) study. Participants (N=21) performed a two-armed bandit task in which, on each trial, they observed a reward randomly sampled from one of the options. After seeing the reward, participants reported their belief on the mean value of this alternative and their confidence on this estimate on a continuous square scale. Rewards were drawn from Beta distributions with the following parameters:

{α = 1; β = 3}; {α = 2; β = 3}; {α = 3; β = 3}; {α = 3; β = 2}; {α = 3; β = 1}.

Participants performed 600 trials divided in 15 blocks of different lengths, ranging from 20 trials to 60 trials. Twenty five % of the trials were “decision trials” where participants had to choose one of the two options and rate their confidence on having chosen the best one.

The second dataset comprised the two experiments carried by Quandt et al. (2022). Both experiments followed identical procedures, with 62 participants in the first one and 60 in the second. Participants observed a hundred rewards from one alternative and then rated their belief on the expected value of the option and their confidence in this rating. They performed five trials with each alternative, for a total of six alternatives. The rewards associated each alternative were drawn from Normal distributions with the following parameters:

{μ = 80, σ = 15}; {μ = 100, σ = 20}; {μ = 120, σ = 30}; {μ = 130, σ = 40}; {μ = 150, σ = 10}; {μ = 160, σ = 5}.

Decision confidence

For the decision confidence analysis we used four datasets. First, we used the data from experiment 2 of Boldt et al. (2019) study. Participants (N=30) performed a two-armed bandit task in which they had to choose, on each trial, one of the options with the aim to maximize their rewards. After choosing, they had to report on a continuous scale their confidence in having selected the best alternative. The alternatives’ rewards distributions were sampled from the same Beta distributions of the value confidence experiment. Participants performed 600 trials, divided in 15 blocks of different lengths. Block length ranged from 20 trials to 60 trials. In 25% of the trials participants did not make a choice but instead reported their estimated mean value of the options (“rating trials”). These trials were excluded from the analysis.

The second dataset corresponded to a study conducted by us in our laboratory (" *Study 1* "). Participants (N=29) performed 20 blocks of 30 trials of a two armed bandit task. After choosing, they had to report their confidence in having selected the best option on a 1 (not sure at all) to 4 (completely sure) scale. Rewards were sampled from the same Beta distributions of the Boldt et al. (2019) study. The third dataset ("*Study 2*") was a pre-registered replication of this protocol (N=30). Pre-registration is available at: *https://doi.org/10.17605/OSF.IO/Z5TCA* .

We employed a fourth dataset ("*Study 3*") to differentiate between models with and without including certainty in estimated values (results are reported in the Supplementary material, see also Computational models section for the description of the models). This dataset was similar as the previous ones, with participants (N=15) performing 20 blocks of 30 trials of a two armed bandit task and reporting confidence on a 1 to 4 scale after each choice. However, the reward distributions were manipulated differently in this task. Inspired by Hertz et al. (2018) design, rewards from each option were sampled from Gaussian distributions, one with higher mean than the other, thus constituting one "correct" and one "incorrect" option. The variances of each one could independently be high (H = 25^2 = 625) or low (L = 10^2 = 100). This resulted in a design with four experimental conditions (thus participants completing 5 blocks with each condition, randomly ordered in the experimental session): H-H, H-L, L-H, L-L, where the first letter indicates the variance of the correct (higher expected value) and the second indicates the variance of the incorrect (lower expected value) option. The mean of the correct option was sampled uniformly from the range {40; 60} and the mean of the incorrect option was the mean of the correct option minus 30. The difference of value between options, therefore, was always the same.

Computational models

Values and value confidence

Two main classes of models were tested to account for the pattern of the value confidence data. One class consisted of models based on Rescorla-Wagner (RW) algorithms (R. A. Rescorla & A. R. Wagner, 1972), and the other was based on Bayesian inference. For the RW models, expected values for the options were updated as follows:

$$V_t = V_{t-1} + \delta\,(r_t - V_{t-1})$$

Were V represents the value of the option, t indexes the trial number, δ is the learning rate parameter and r is the reward obtained.

For the Bayesian models, posterior probability distributions of the options' expected values were constructed using Bayesian inference. As mentioned, in the Boldt et al. datasets, rewards were drawn from Beta distributions. Therefore, a Bayesian observer computes, on each trial, a posterior distribution over the mean of a Beta distribution, represented by θ, by integrating the likelihood of the rewards given θ and the prior probability of θ, which we modelled as uniform. Formally:

$$P\,(\theta_t|r_{1:t}) \sim P\,(r_{1:t}|\theta)\,P\,(\theta)$$

The maximum-a-posteriori (MAP) value of this posterior is therefore the predicted expected value of the option on trial t, i.e., V_t. These Bayesian models were implemented using Stan (*Stan*, 2025; MCMC diagnostics are reported in the Supplementary material). A re-parametrization was used to infer the mean of a Beta distribution. Specifically, we estimated the parameters that govern the shape of a Beta distribution, α and β, and then obtained the mean of the Beta as:

$$\theta = \frac{\alpha}{\alpha + \beta}.$$

Given that in Quandt et al. (2022) datasets rewards came from Normal distributions, we inferred the options' expected values (θ) by leveraging on the Normal-Normal conjugate family (Johnson, 2022). The posterior of θ, therefore, was therefore computed as follows:

$$P(\theta|r_{1:100}) \sim N(\bar{p}\frac{\sigma^2}{n\pi^2+\sigma^2} + \overline{r_{1:100}}\frac{n\pi^2}{n\pi^2+\sigma^2}, \frac{\pi^2\sigma^2}{n\pi^2+\sigma^2})$$

Where $\bar{p}$ represents the mean of the prior (which we set at zero), π^2 represents prior variance (which we set at 100, thus constructing an uninformative prior), $\overline{r_{1:100}}$ is the mean of the rewards seen and σ^2 is the variance of the rewards seen (i.e.: the likelihood variance). The subscript 1:100 represents that subjects faced a hundred rewards for each option.

Finally, for the dataset of Study 3, reported in the Supplementary material, Bayesian inference was implemented using a Kalman-filter following Gershman (2018) to recursively infer posteriors' mean (θ) and variances (σ^2) for each option. Formally:

$$\theta_{t+1} = \theta_t + \delta_t(r_t - \theta_t)$$
$$\sigma^2_{t+1} = \sigma^2_t - \delta_t\sigma^2_t$$

where the learning rate δ_t is given by:

$$\delta_t = \frac{\sigma^2_t}{\sigma^2_t + \tau^2}$$

where τ^2 was set to the specific option's reward distribution variance. For all options initial means (θ) were set at zero and initial σ^2 at 100 (uninformative prior).

Given these two modeling frameworks for value updating (i.e.: the Bayesian inference framework and the RW framework), we compared several models for explaining value confidence. The Bayesian framework naturally offers a measure for the certainty associated to the inferred values of the options, which is the inverse of the standard deviation of the posterior distribution over the expected value of the option at play. Thus, under the *Bayesian model*, value confidence can be expressed as:

$$value\ conf = \frac{1}{\sigma_{P(\theta_t|r_{1:t})}}$$

Here σ represents the standard deviation of the posterior distribution over θ, the estimated value of the specific option at play.

In contrast, as RW algorithms do not have an explicit measure of certainty in the values associated with each alternative, we extended this approach with several possible mechanisms for computing value confidence. The first of these extensions, the *RW surprise model*, defines value confidence as the inverse of the absolute surprise of the outcome. Formally:

$$value\ conf = \frac{1}{|r_t - V_{t-1}|}$$

The second RW model comprised a separated RW rule to track the variance of the outcomes (*RW tracking variance model*, Hertz et al., 2018):

$$\tau_t = \tau_{t-1} + \gamma\left[(r_t - V_{t-1})^2 - \tau_{t-1}\right]$$

Here τ refers to the computed variance of the rewards observed, and γ is the variance learning rate. Value confidence under this model reflected the inverse of the tracked variance, i.e.:

$$value\ conf = \frac{1}{\tau_t}$$

For the third RW model, value confidence reflected the number of rewards observed of the specific option sampled (*RW n model*). Let *n* refers to that number; value confidence is represented as:

$$value\ conf = n$$

Next, based on the *RW n* model, we constructed two additional variants: In this models value confidence reflects the logarithm of the number of rewards experienced with the specific option at play (*RW log(n) model*) or the square root of the same number (*RW sqrt(n) model*).

Finally, we tested a weighted combination of the Bayesian model, the RW sqrt(n) model and the RW log(n) model with the surprise generated by the outcome. The *Bayesian-surprise model* can be expressed as follows:

$$value\ conf = w\frac{1}{\sigma_{P(\theta_t|r_{1:t})}} - (1-w)\left|(r_t - \theta_{t-1})\right|$$

The *RW sqrt(n)-surprise model* can be then stated as:

$$value\ conf = w\sqrt{n} - (1-w)\left|(r_t - V_{t-1})\right|$$

And the *RW log(n)-surprise model* is the same as above but with the logarithm of *n* instead of the square root. Note that the weights (w) were independent parameters for each model.

All the RW models were tested in the Boldt et al. (2019) dataset. In the Quandt et al. (2022) dataset we only tested the *RW sqrt(n) model* and the *RW tracking variance model* given that 1) the *RW sqrt(n) model* was the best non-Bayesian model in Boldt et al. (2019) data; 2) the *RW tracking variance model* was the only non-bayesian model including information of the reward distribution variance for value confidence, the key manipulation in Quandt et al. (2022) dataset. The *Bayesian model* was included in all datasets.

Decisions and decision confidence

The decision confidence experiments involved classic two-armed bandit tasks where participants chose between two alternatives (which we will represent as *A* and *B*) and then had to rate their confidence on having chosen the best option (see Datasets description section). Given that we found that value confidence was best explained by a Bayesian model, we used Bayesian inference to update the values of the options on each trial as described for the value confidence data. Bayesian inference for options' posterior means and variances on each trial was implemented in Stan for the Boldt et al. (2019), Study 1 and Study 2 datasets and with a Kalman-filter for Study 3 dataset, as described in the Values and value confidence section above.

For fitting decisions, we used a classic softmax equation:

$$P\left(d_t = A\right) = \frac{1}{1 + \exp\left(-DV_t^{\lambda}\right)}$$

Here *d* represents the decision, *DV* represents the decision variable (which is $V_{t-1}^{A} - V_{t-1}^{B}$, note that *V* here is the maximum-a-posteriori value of each option), and A is the slope of the sigmoid that controls the noise in the decision process as an inverse temperature parameter. We tested two variations for fitting decisions. In one of them λ was treated as a constant. In the other, λ dynamically varied across trials, being modulated by value confidence in the following way:

$$\lambda_t = b_0 + b_1\frac{1}{\sqrt{\sigma_{At-1}^2 + \sigma_{Bt-1}^2}}$$

Where σ represent the standard deviation of the posteriors over the options' expected values. The term multiplying b_1 can be interpreted as a global value confidence, as it reflects the inverse of the combined uncertainty associated with the value estimates of both options. As long as b_1 is positive, value confidence will modulate the exploration-exploitation trade-off by reducing decision noise with increasing values of value confidence. Conversely, if b_1 is indistinguishable from zero, decision noise is constant across trials and participants do not take into account the options' uncertainty for making decisions.

For decision confidence we tested several models. In the first one, confidence followed the Bayesian confidence hypothesis (Meyniel et al., 2015; Sanders et al., 2016), i.e.: confidence reflects the probability of being correct (*Bayesian model).* As the posteriors over options' expected values have Gaussian shape, this probability can be computed as follows:

$$confidence = P\left(correct_t\right) = \Phi\left(\frac{V_{t-1}^{chosen} - V_{t-1}^{unchosen}}{\sqrt{\sigma_{chosen,t-1}^2 + \sigma_{unchosen,t-1}^2}}\right)$$

Where Φ is the standard cumulative Gaussian function. In the second model, the *Bayesian-hybrid model*, decision confidence reflected a weighted sum of the probability of being correct and the overall certainty of the estimated values (i.e.: the overall value confidence; Navajas et al., 2017). Formally:

$$confidence = \omega\Phi\left(\frac{V_{t-1}^{chosen} - V_{t-1}^{unchosen}}{\sqrt{\sigma_{chosen,t-1}^2 + \sigma_{unchosen,t-1}^2}}\right) + \left(1-\omega\right)\frac{1}{\sqrt{\sigma_{chosen,t-1}^2 + \sigma_{unchosen,t-1}^2}}$$

In the main manuscript we report the results of these two primary models. However, because prior work on confidence in bandit tasks has relied on models that do not incorporate uncertainty in value estimates (Desender & Verguts, 2024; Salem-Garcia et al., 2023), we sought to assess whether including the certainty of the decision variable in the Bayesian-hybrid model was indeed warranted. To do so, we compared its fit to that of models that rely solely on the estimated option values. Specifically, we tested two extra models. For one of them, confidence reflected the absolute difference between the estimated values of the options (*Mean difference model)*, i.e.:

$$confidence = \left|V_{t-1}^{chosen} - V_{t-1}^{unchosen}\right|$$

For the other extra model, and given that the value of the chosen option have been identified as important in explaining decision confidence both in bandit tasks and in perceptual decision making (i.e.: the "positive evidence bias", Zylberberg et al., 2012), confidence reflected a weighted sum of the absolute difference between the estimated values of the options and the value of the chosen option (*Mean difference + PEB model).* Formally:

$$confidence = \omega\left|V_{t-1}^{chosen} - V_{t-1}^{unchosen}\right| + \left(1-\omega\right)V_{t-1}^{chosen}$$

Model fitting results of these two extra models are reported in the Supplementary material.

Model fitting and model comparison procedure

We fitted all models by maximizing the log-likelihood of the parameters given each subject data using the *optim* function in R with the simulated annealing method. Ten optimization routines with different starting points were run per subject. The probability of a specific decision was given by the softmax equations detailed above. For computing the probability of confidence ratings (both value and decision confidence) in the case of Boldt et al. (2019) datasets and our datasets,

we used a set of confidence criteria, $\{c_1,..., c_{k-1}\}$, where k is the number of confidence ratings, that divided a Gaussian distribution with mean at the confidence predicted by the model and a standard deviation that was fitted to each subject (akin as observational noise, Nunez et al., 2024). The probability of a specific level of confidence is then the area under this Gaussian that is restricted by corresponding confidence criteria. Boldt et al. (2019) confidence data was discretized from 0 to 10 to be able to apply this procedure. For Quandt et al. (2022) data we applied a linear transformation for mapping models' confidence to participants' confidence. We computed the likelihood of the parameters by determining the proportion of transformed confidence predictions from the model that matched the reported confidence divided by the total number of predictions (we used this approach to reduce the number of parameters as there were only 30 trials per subject).

We compared all models using Bayesian model selection at the group level (Stephan et al., 2009). Specifically, we computed Bayesian Information Criterion weights (Wagenmakers & Farrell, 2004) for each subject and for each model as a proxy for model evidence (i.e., the belief that model m generated data from subject i) and then used the *bmsR* R package to compute exceedance and protected exceedance probabilities (the probabilities that participants were more likely to use a certain model to generate behavior rather than any other alternative model) which were used as the metric for model comparison. We report in the Supplementary material parameter and model recovery results.

Statistical analyses

We also carried out several model free statistical tests. We used one-sample t-tests against zero (two-sided) for testing whether the bl parameter was greater than zero in the three datasets employed (Figure 3b). For each subject we run ordinal regressions predicting decision confidence on each trial with the two model-derived variables: the probability of being correct and the overall value confidence. Each of these two terms were therefore associated with two beta values, $\beta_{p(correct)}$ for the first term and $\beta_{value\ conf.}$ for the second term, which allowed us to characterize individual differences in confidence reporting (Figure 5a). We then used these beta values (as well as their interaction) to predict task performance (the proportion of correct choices), the b_1 parameter and the metacognitive ability of each participant using the following linear regression models (Figure 5b, 5c, 5d):

$$p\,(correct) \sim \beta_{p(correct)} + \beta_{value\ conf.} + \beta_{p(correct)} * \beta_{value\ conf.}$$
$$b_1 \sim \beta_{p(correct)} + \beta_{value\ conf.} + \beta_{p(correct)} * \beta_{value\ conf.}$$
$$metacognition \sim \beta_{p(correct)} + \beta_{value\ conf.} + \beta_{p(correct)} * \beta_{value\ conf.}$$

Metacognition was measured using a logistic regression per subject predicting a correct (1) or incorrect (0) response using the confidence level reported by the participant on each trial:

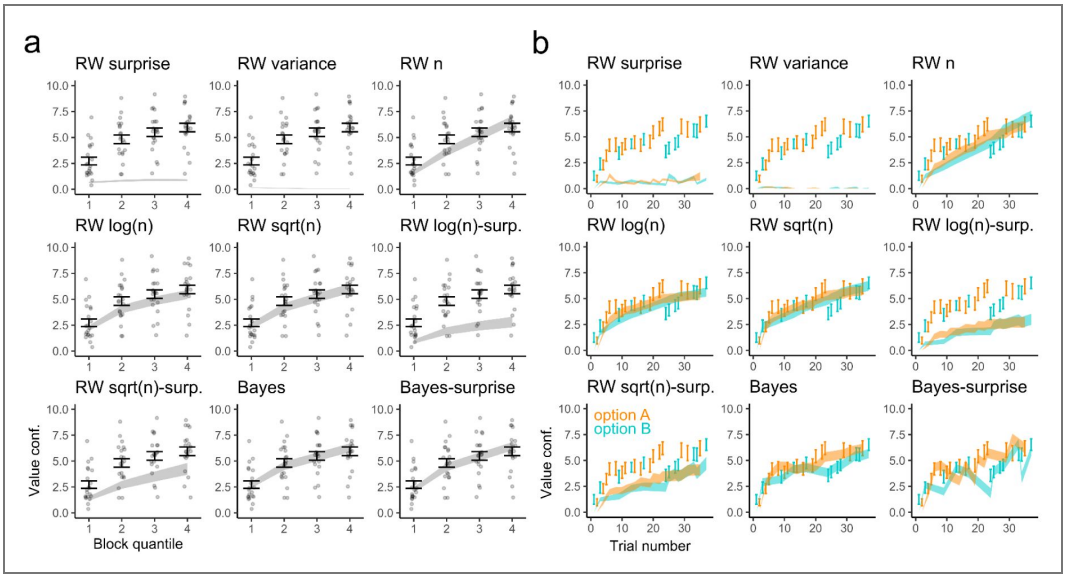
$$P\,(response = correct) = \frac{e^x}{1 + e^x}$$

Where $x = \beta_0 + \beta_1 * confidence$.

Supplementary material

Value confidence predictions for all models

Here we illustrate the predictions in the Boldt et al. (2019) dataset of all tested models for value confidence. Supplementary Figure 1a shows models' predictions for all the trials, while Supplementary Figure 1b shows models' predictions for one specific block as an example.



Supplementary Figure 1. Model fits for value confidence data, all models.

(a) Models' predictions considering all the trials in Boldt et al. (2019) dataset. (b) Models' predictions considering only one block, for illustrative purposes. Error bars represent the standard error of the mean (SEM) of the behavioural data and shaded regions represent the SEM of the models' predictions.

MCMC diagnostics

For the datasets of Boldt et al. (2019) and our 2 main studies rewards came from Beta distributions. As part of the Bayesian modelling, we used Stan to numerically compute trial-by-trial posterior estimates of the options' values. Two chains were run with 8000 iterations each, using a burn-in of 4000 iterations and a thinning factor of 1.

To assess the reliability of this procedure, we examined standard MCMC diagnostics—R-hat, effective sample sizes ratios (*Neff/N*), divergent transitions, trace plots and checks of the posterior samples—using two simulation scenarios. In the first, we performed inference after observing only a single reward, illustrating sampler behaviour under minimal information (i.e., at the start of a block). This procedure was repeated 1000 times and we extracted the mean and range of the diagnostic metrics. The second simulation repeated the same procedure but after observing twenty rewards, providing an example of sampler performance with more informative data.

We report the results in the Supplementary Table 1 below. Across both scenarios, all diagnostics indicated proper convergence of the Stan sampler, supporting the validity of the numerical posterior estimates used in our modelling.

Supplementary Table 1. MCMC diagnostics summary.

Simulation	R-hat	Neff/N	Divergent transitions
One reward seen.	1 {1;1}	0.83 {0.6;1.02}	0 {0;0}
Twenty rewards seen.	1 {1;1}	1.01 {0.71;1.21}	0 {0;0}

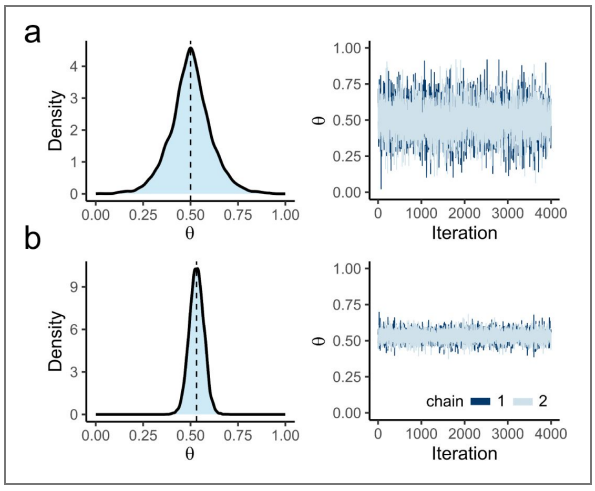
Note: values represent the mean of the obtained diagnostics in the 1000 simulations of each scenario, and values in curly braces represent the full range {min;max} of the obtained values.

We also include representative posterior samples from these two scenarios, which showed an approximately Gaussian shape, as well as trace plots confirming stable and well-mixed chains in Supplementary Figure 2 below.

Model and parameter recovery results. Value confidence

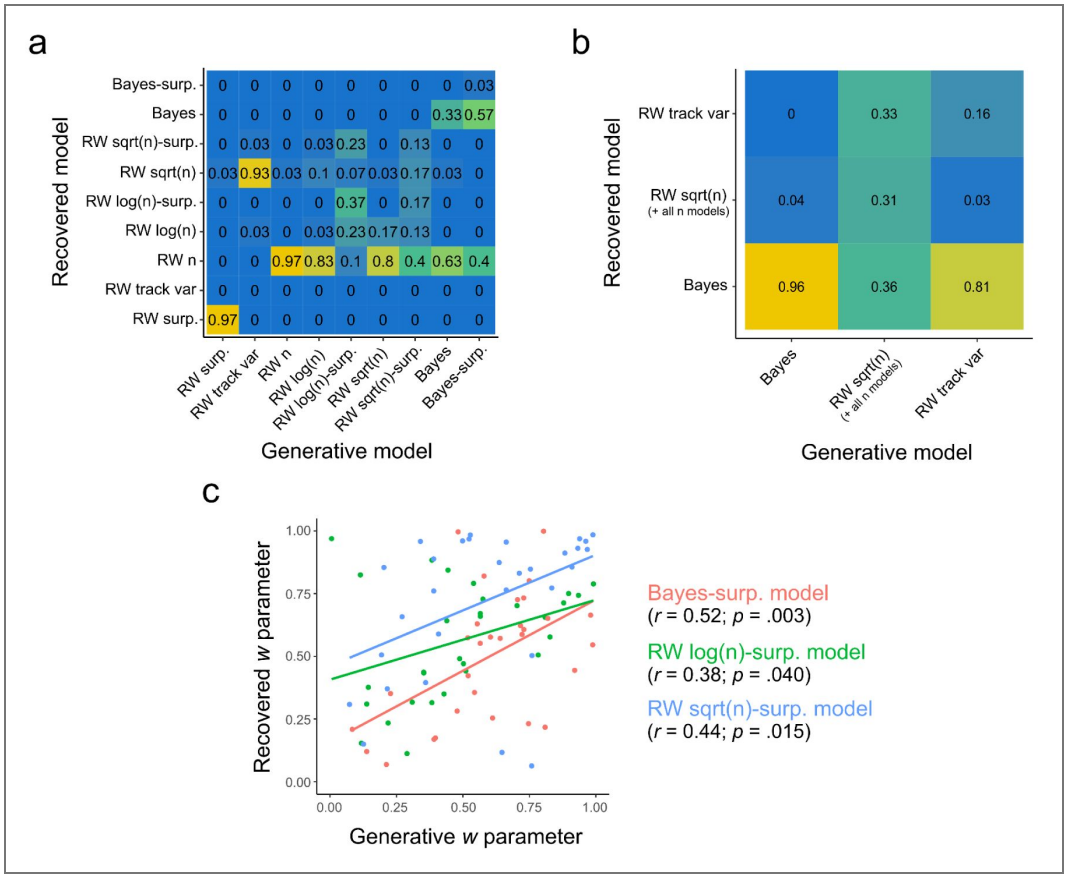
To evaluate our ability to differentiate between the models tested for value confidence data, we conducted model recovery analysis. We simulated 30 subjects with each model both with the Boldt et al. (2019) experiment design and then fitted all the models to the simulated data with the same procedure described in the model fitting section of the main manuscript. We then checked the proportion of times that the best fitting model (according to the Bayesian Information Criterion) was equal to the generative model among all the simulated subjects with that specific generative model. Supplementary Figure 3a below depicts the results obtained.

We also carried out model recovery with the Quandt et al. (2022) experiment design. For this second model recovery analysis, we only tested the RW tracking variance model, the Bayesian model, and the RW n model due to two reasons: first, models that included a surprise term were worse at fitting the Boldt et al. (2019) data (see Supplementary Figure 1); second, in the Quandt et al (2022) design all models that are a function of the number of trials experienced (such as the RW sqrt(n) model or the RW log(n) model) make the same predictions as the RW n model. Due to computation time, having less models also allowed us to simulate more subjects per model: 100. Once data which each model was simulated, we then followed the same procedure as described above for the Boldt et al. (2019) experiment design. We found that including the Quandt et al (2022) dataset improves model recovery results (Supplementary Figure 3b), as we could better distinguish between models that take (the RW track variance and the Bayesian model) and do not take (the RW n family of models) the reward distribution variance into account. Note, however, that the RW tracking variance model was confused with the Bayesian model in 81% of the cases.



Supplementary Figure 2. Illustration of posterior distributions and trace plots.

(a) Scenario 1: only one reward (r = 0.5) was observed. (b) Scenario 2: twenty rewards (mean r = 0.54) were observed. Theta represents the inference over the mean value of a Beta distribution, i.e.: $\frac{\alpha}{\alpha+\beta}$. The vertical dashed lines at the posteriors represent the mean of the posterior distribution.



Supplementary Figure 3. Model and parameter recovery results for value confidence data.

(a) Model recovery results for the Boldt et al. (2019) experiment design. We found that the RW n model is mistakenly considered to be the model that generated the data for several generative models, with the most prominent examples being the RW log(n) model and the RW sqrt(n) model. (b) Model recovery results for Quandt et al. (2022) experiment design. While this result is uninformative to distinguish between the models that are a function of the number of rewards seen, it helps to differentiate that group of models and the models that take the variance of the reward distribution into account. (c) Parameter recovery results for the parameter that weighted the contribution of the surprise term in the three models that included it.

Furthermore, with the same simulated datasets we also checked parameter recoverability (Supplementary Figure 3c). We especially focused the analysis on the parameters related to value confidence, that is, the weights applied to the surprise term in the Bayesian-suprise model, the RW sqrt(n)-surprise model and the RW log(n)-surprise model. We found significant correlations between the generative and the recovered weights for the surprise term in the three models (Bayesian-surprise model: r = 0.52, p = .003; RW log(n)-surprise model: r = 0.38, p = .040; RW sqrt(n)-surprise model: r = 0.44, p = .015; Overall correlation: r = 0.41, p < .001).

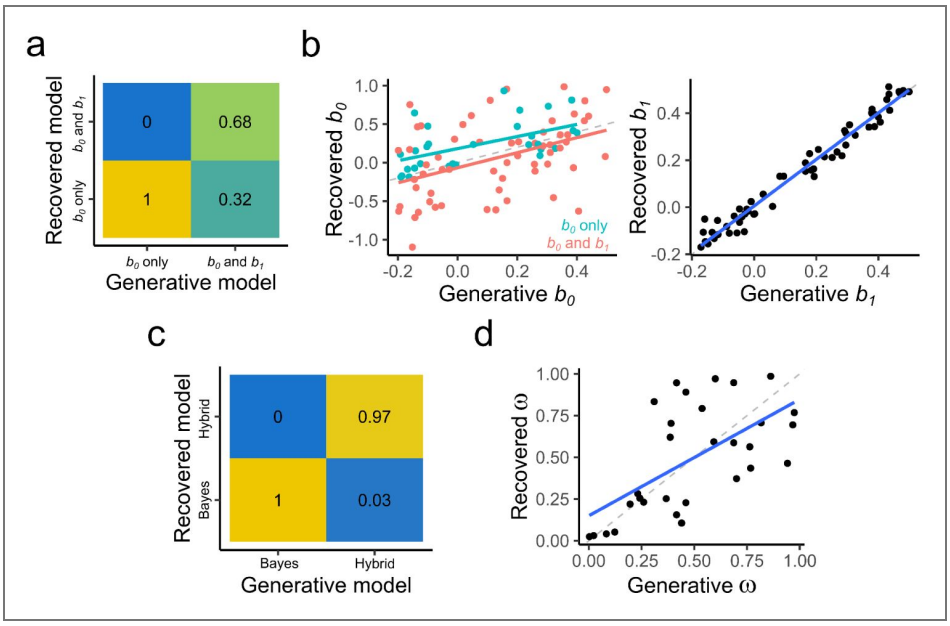
Model and parameter recovery results. Decision confidence.

Using the same procedure described above (i.e.: simulating data with all models, then fitting the simulated data with all models and finally checking the proportion of times that the recovered model equals the generative model) we also carried out model and parameter recovery analyses in the models tested for decision confidence data. First, we checked the recoverability of the model including the modulation of decision noise by value confidence via the b_1 parameter (dynamic decision noise model) in contrast with the model without the decision noise modulation (constant decision noise model) (Supplementary Figure 4a). We used this simulated data to also check for parameter recoverability of the b_0 and b_1 parameters that modulated the softmax we implemented as a decision policy. We found strong correlations between the generative and the recovered parameters for both parameters (b_0 : r = 0.55, p = .002 [constant decision noise model]; b_0: r = 0.43, p < .001 [dynamic decision noise model]; b_1 : r = 0.99, p < .001; Supplementary Figure 4b). This is important as we used these parameters (especially the b_1 parameter) to characterize individual differences in the way decisions were made, that we then also correlated with individual differences in confidence computations (main manuscript Figure 5c).

Moving to confidence data, we checked model recoverability on the decision confidence models, that is, the Bayesian and the Bayesian-hybrid model. The recoverability of the two models were practically perfect (Supplementary Figure 4c). Finally, we checked whether we can recover the weight associated with the overall value confidence in the Hybrid model. We found a significant correlation between the generative and the recovered model for this particular parameter (r = 0.62; p < .001; Supplementary Figure 4d).

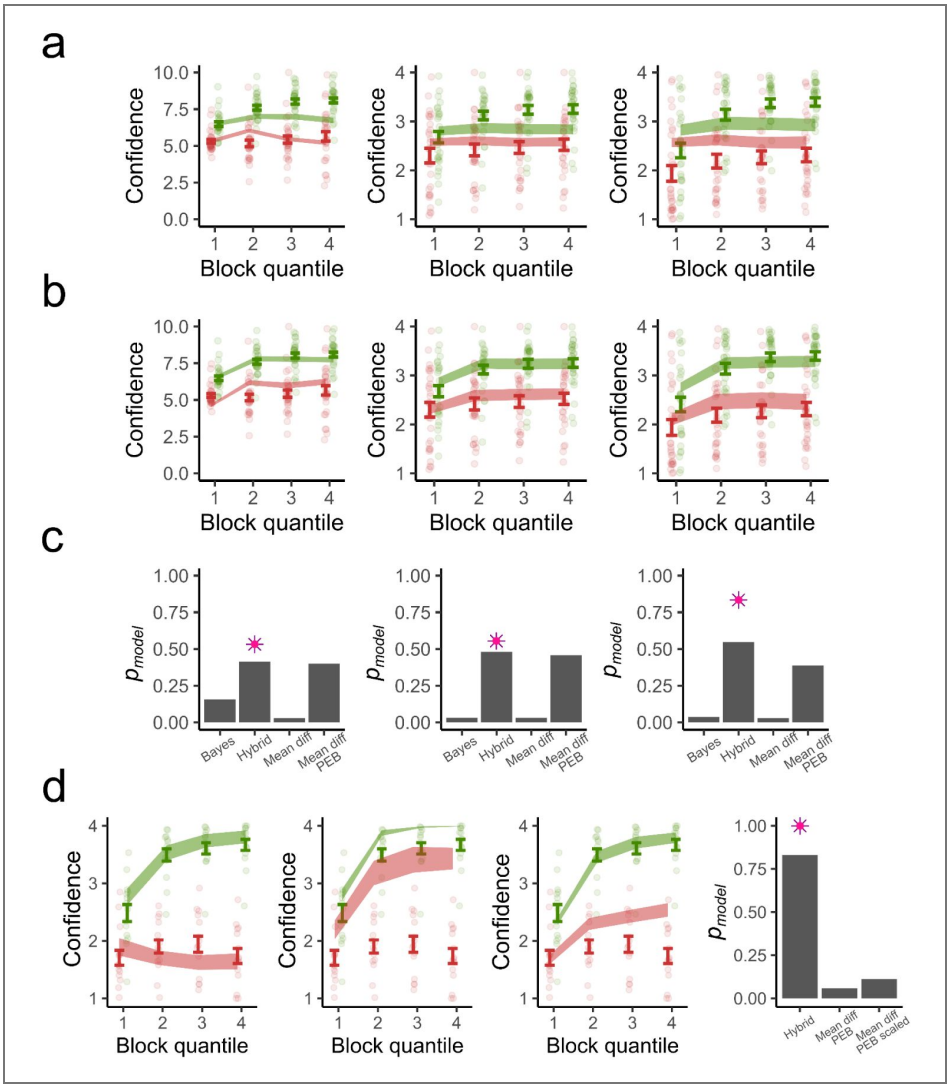
Evaluating confidence models that ignore uncertainty in estimated values

The Bayesian and Bayesian-hybrid models incorporate posterior uncertainty into their confidence computations. However, previous work has shown that confidence can also be captured using only estimated option values, without explicitly modelling their uncertainty (Salem-García et al., 2023; Desender & Verguts, 2024). For this reason, we tested two additional models: one in which confidence reflected the absolute difference between estimated option values (Mean Difference model), and another in which confidence was a weighted combination of this difference and the value of the chosen option (Mean Difference + PEB model). Further details are provided in the Methods section of the main manuscript. We first fitted these models to the three datasets reported in the main manuscript. Supplementary Figure 5a illustrates the predictions of the Mean Difference model and Supplementary Figure 5b illustrates the predictions of the Mean Difference + PEB model. Although the Bayesian-hybrid model remained slightly superior to the other models (Boldt et al.: p_{model} = 0.41; $p_{exc.}$ = 0.53; $p_{pexc.}$ = 0.53; Study 1: p_{model} = 0.48; $p_{exc.}$ = 0.55; $p_{p.exc.}$ = 0.55; Study 2: p_{model} = 0.55; $p_{exc.}$ = 0.84; $p_{p.exc.}$ = 0.84; Supplementary Figure 5c) and enabled meaningful characterisation of inter-individual profiles, the Mean Difference + PEB model emerged as a competitive alternative across these datasets.



Supplementary Figure 4. Model and parameter recovery results for decision confidence data.

(a) Model recovery results regarding the models with and without the b_1 parameter. (b) Parameter recovery results for the b_0 and b_1 parameters. (c) Model recovery results for the decision confidence models. (d) Parameter recovery results for the ω parameter that weighted the sources of information for decision confidence in the Hybrid model: the probability of being correct and the overall value confidence.



Supplementary Figure 5. Evaluating the predictions of models that do not incorporate posterior uncertainty.

(a) "Mean difference" model predictions to the three main datasets reported in the main manuscript (first figure: Boldt et al. data; second figure: study 1 data; third figure: study 2 data). (b) "Mean difference + PEB" model predictions. Figures are ordered in the same way as in (a). (c) Model comparison results. Figures are ordered in the same way as (a) and (b). (d) Model predictions to the dataset manipulating the mean and variance of the reward distributions of the options plus model comparison results. First figure: Hybrid model; second figure: Mean difference + PEB model; third figure Mean difference + PEB model including a scale parameter for the estimated values of the options; fourth figure: model comparison results. The conventions of the whole figure are the same as in the main manuscript, namely: green and red represent correct and incorrect responses respectively, error bars and shaded lines represent empirical data and model fits respectively, dots are individuals' average values, and pink dots and purple asterisks in model comparison plots are exceedance and protected exceedance probabilities respectively.

Therefore, we next sought to arbitrate between this model and the Bayesian-hybrid model, and to further assess whether incorporating value-estimate uncertainty is necessary for modelling confidence. To this end, we fitted both models to an additional dataset in which we manipulated both the means and variances of the reward distributions (details are given in Methods). This manipulation was intended to differentiate the computations of the two models, as only the Bayesian-hybrid model explicitly integrates uncertainty.

The Mean Difference + PEB model failed to generalise to this dataset with more complex mean/variance manipulations (Supplementary Figure 5d). One possible explanation is that changes in the reward distributions across blocks modified the scale of option values in ways that the model—which relies heavily on those values—struggles to accommodate. We therefore implemented a variant in which estimated values were rescaled via an additional parameter to reduce across-block variability, which indeed improved model fit (Supplementary Figure 5d). Nonetheless, the Bayesian-hybrid model was clearly preferred in this dataset (p_{model} = 0.83; $p_{exc.}$ > 0.99; $p_{p.exc.}$ > 0.99). This finding reinforces the main conclusion of the manuscript: incorporating uncertainty in estimated values appears essential for explaining confidence in bandit tasks. Finally, we again compared the constant and dynamic decision-noise formulations and found that the dynamic model was superior (p_{model} = 0.64; $p_{exc.}$ = 0.89; $p_{p.exc.}$ = 0.62) with b_1 values, on average, greater than zero (t_{14} = 3.36; p = .005, d = 0.87).

Implications of dissociating overall uncertainty from decision difficulty in the hybrid model

In the hybrid model, confidence is computed as a weighted combination of the probability of being correct and overall value confidence, the latter defined as the inverse of the total uncertainty (i.e., the sum of the variances associated with the value representations of the options). While this formulation captures an intuitive notion of global uncertainty, it does not necessarily track the discriminability of the options, and thus does not map directly onto decision difficulty.

As a consequence, these two quantities can, in principle, diverge. For instance, overall uncertainty can be high even when the difference between options is large (i.e., an easy decision), and conversely low when options are similar (i.e., a difficult decision). This dissociation raises a potential concern: if confidence were to rely heavily on overall uncertainty, the model could, in principle, give rise to counterintuitive and poorly calibrated confidence patterns that are misaligned with decision difficulty.

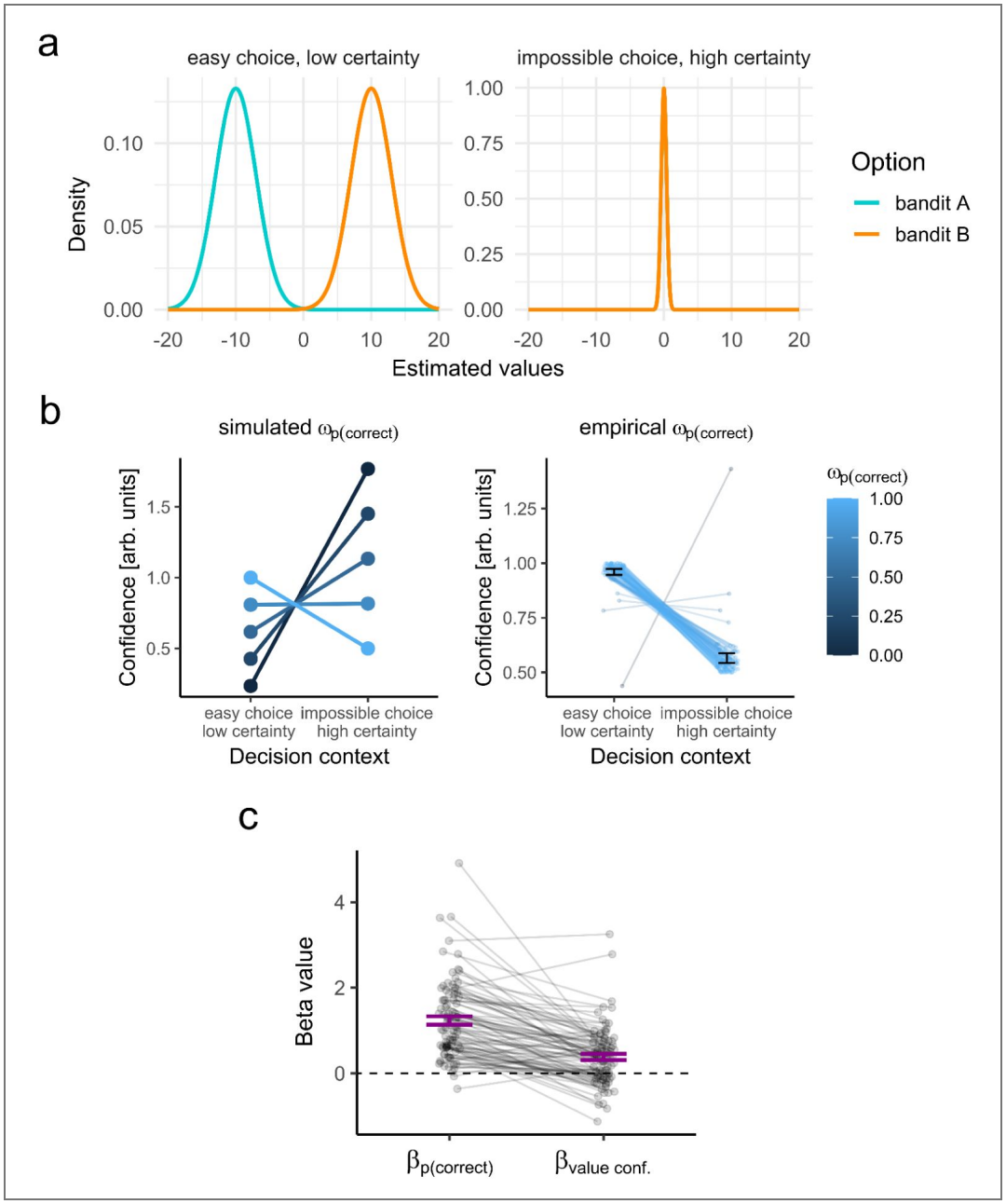
To evaluate whether this issue arises in practice, we simulated model predictions in two extreme scenarios that explicitly dissociate these quantities (see Supplementary Figure 6a below): (i) an easy decision under low certainty, and (ii) an "impossible" decision (i.e.: the two options having the same expected value) under high certainty. For the first case, we constructed posterior estimates of two options (A and B) as normal distributions with the following parameters: {μ_A = −10, σ_A = 3}; {μ_B = 10, σ_B = 3}. For the second case, normal distributions had the following parameters: {μ_A = 0, σ_A = 0.4}; {μ_B = 0, σ_B = 0.4}. Confidence in these two scenarios was computed as follows:

$$conf. = \omega\Phi\left(\frac{\mu_B - \mu_A}{\sqrt{\sigma_B^2 + \sigma_A^2}}\right) + (1-\omega)\frac{1}{\sqrt{\sigma_B^2 + \sigma_A^2}}$$

Crucially, we varied the relative weight assigned to the probability of being correct and to overall value confidence, ω = {0, 0.25, 0. 5, 0. 75, 1}, and additionally examined the confidence predictions using the weights observed empirically in our data (M=0.95, SD=0.08, range=[0.26;0.99]).

Simulations (Supplementary Figure 6b below) show that only when overall value confidence is given a large weight do counterintuitive patterns emerge. However, when the probability of being correct receives a dominant weight (as consistently observed in our participants; Supplementary Figure 6b and 6c) the model produces well-calibrated confidence estimates, with higher

confidence assigned to easier decisions despite elevated uncertainty. This supports the interpretation that overall value confidence acts as a secondary, biasing component in confidence computation, modulating but not overriding the primary Bayesian estimate of correctness.



Supplementary Figure 6. Dissociating overall uncertainty from decision difficulty in the hybrid model. (a) The two proposed scenarios of decisions: an easy choice (estimated expected values are very different) with high uncertainty (expected values are not precisely estimated) and an impossible choice (estimated expected values are equal) with low uncertainty (expected values are precisely estimated). Note that in the right panel the cyan distribution cannot be seen as it is behind the orange one. (b) Model predictions varying the w value. (c) Model predictions using the empirical ω values found by fitting the model to the data. (d) A comparison of the beta values obtained from the regression shows that the contribution of the probability of being correct is significantly higher than the contribution of overall value confidence to confidence judgments. We opted for comparing the regression beta values instead of the ω values as we normalised the predictors for the regression, thus beta values are directly comparable between them.

Additional information

Funding

Funder	Grant reference number	Author
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)	2018-03614	Pablo Barttfeld
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)	2021-0083	Pablo Barttfeld
Universidad de Buenos Aires (UBA)	UBACyT 20020170100330BA	Guillermo Solovey
Center for Open Science, Templeton World Charity Foundation and the Association for the Scientific Studies of Consciousness	Funding Consciousness Research with Registered Reports	Pablo Barttfeld

Author ORCID iDs

Nicolás A Comay: https://orcid.org/0000-0002-1693-9253
Pablo Barttfeld: https://orcid.org/0000-0001-8530-6938

References

Abir Y., Shadlen M. N., Shohamy D. (2024) Human Exploration Strategically Balances Approaching and Avoiding Uncertainty. *eLife* **13**:RP94231 https://doi.org/10.7554/eLife.94231 | PubMed

Adler W. T., Ma W. J. (2018) Limitations of Proposed Signatures of Bayesian Confidence. *Neural Computation* **30**:3327-3354 https://doi.org/10.1162/neco_a_01141 | PubMed

Ais J., Zylberberg A., Barttfeld P., Sigman M. (2016) Individual consistency in the accuracy and distribution of confidence judgments. *Cognition* **146**:377-386 https://doi.org/10.1016/j.cognition.2015.10.006 | PubMed

Balsdon T., Philiastides M. G. (2024) Confidence control for efficient behaviour in dynamic environments. *Nature Communications* **15**:9089 https://doi.org/10.1038/s41467-024-53312-3 | PubMed

Balsdon T., Wyart V., Mamassian P. (2020) Confidence controls perceptual evidence accumulation. *Nature Communications* **11**:1753 https://doi.org/10.1038/s41467-020-15561-w | PubMed

Bang D., Aitchison L., Moran R., Herce Castanon S., Rafiee B., Mahmoodi A., Lau J. Y F., Latham P. E., Bahrami B., Summerfield C. (2017) Confidence matching in group decision-making. *Nature Human Behaviour* **1**:0117 https://doi.org/10.1038/s41562-017-0117

Baranski J. V., Petrusic W. M. (1994) The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics* **55**:412-428 https://doi.org/10.3758/BF03205299 | PubMed

Behrens T. E. J., Woolrich M. W., Walton M. E., Rushworth M. F. S. (2007) Learning the value of information in an uncertain world. *Nature Neuroscience* **10**:1214-1221 https://doi.org/10.1038/nn1954 | PubMed

Bénon J., Lee D., Hopper W., Verdeil M., Pessiglione M., Vinckier F., Bouret S., Rouault M., Lebouc R., Pezzulo G., *et al.* (2024) The online metacognitive control of decisions. *Communications Psychology* **2**:23 https://doi.org/10.1038/s44271-024-00071-y | PubMed

Boldt A., Blundell C., De Martino B. (2019) Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of Consciousness* **2019**:niz004 https://doi.org/10.1093/nc/niz004 | PubMed

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**:1520-1531 https://doi.org/10.1037/xhp0000404 | PubMed

Boundy-Singer Z. M., Ziemba C. M., Goris R. L. T. (2022) Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour* **7**:142-154 https://doi.org/10.1038/s41562-022-01464-x | PubMed

Bounmy T., Eger E., Meyniel F. (2023) A characterization of the neural representation of confidence during probabilistic learning. *NeuroImage* **268**:119849 https://doi.org/10.1016/j.neuroimage.2022.119849 | PubMed

Brus J., Aebersold H., Grueschow M., Polania R. (2021) Sources of confidence in value-based choice. *Nature Communications* **12**:7337 https://doi.org/10.1038/s41467-021-27618-5 | PubMed

Chambon V., Théro H., Vidal M., Vandendriessche H., Haggard, P, & Palminteri S. (2020) Information about action outcomes differentially affects learning from self-determined versus imposed choices. *Nature Human Behaviour* **4**:1067-1079 https://doi.org/10.1038/s41562-020-0919-5 | PubMed

Comay N. A., Della Bella G., Lamberti, P, Sigman M., Solovey G., Barttfeld P (2023) The presence of irrelevant alternatives paradoxically increases confidence in perceptual decisions. *Cognition* **234**:105377 https://doi.org/10.1016/j.cognition.2023.105377 | PubMed

Desender K., Boldt A., Yeung N. (2018) Subjective Confidence Predicts Information Seeking in Decision Making. *Psychological Science* **29**:761-778 https://doi.org/10.1177/0956797617744771 | PubMed

Desender K., Verguts T (2024) Decision Confidence and Outcome Variability Optimally Regulate Separate Aspects of Hyperparameter Setting. *BioRxiv* https://doi.org/10.1101/2024.10.03.616475

Fleming S. M. (2024) Metacognition and Confidence: A Review and Synthesis. *Annual Review of Psychology* **75**:241-268 https://doi.org/10.1146/annurev-psych-022423-032425 | PubMed

Gershman S. J. (2018) Deconstructing the human algorithms for exploration. *Cognition* **173**:34-42 https://doi.org/10.1016/j.cognition.2017.12.014 | PubMed

Godara P. (2025) Apparent learning biases emerge from optimal inference: Insights from master equation analysis. *Proceedings of the National Academy of Sciences* **122**:e2502761122 https://doi.org/10.1073/pnas.2502761122 | PubMed

Hangya B., Sanders J. I., Kepecs A. (2016) A Mathematical Framework for Statistical Decision Confidence. *Neural Computation* **28**:1840-1858 https://doi.org/10.1162/NECO_a_00864 | PubMed

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

Hertz U., Bahrami B., Keramati M. (2018) Stochastic satisficing account of confidence in uncertain value-based decisions. *PLOS One* **13**:e0195399 https://doi.org/10.1371/journal.pone.0195399 | PubMed

Hoven M., Lebreton M., Engelmann J. B., Denys D., Luigjes J., Van Holst R. J. (2019) Abnormalities of confidence in psychiatry: An overview and future perspectives. *Translational Psychiatry* **9**:268 https://doi.org/10.1038/s41398-019-0602-7 | PubMed

Hoven M., Luigjes J., Denys D., Rouault M., Van Holst R. J. (2023) How do confidence and self-beliefs relate in psychopathology: A transdiagnostic approach. *Nature Mental Health* **1**:337-345 https://doi.org/10.1038/s44220-023-00062-8

Hoxha I., Sperber L., Palminteri S. (2025) Evolving choice hysteresis in reinforcement learning: Comparing the adaptive value of positivity bias and gradual perseveration. *Proceedings of the National Academy of Sciences* **122**:e2422144122 https://doi.org/10.1073/pnas.2422144122 | PubMed

Johnson A. A., Ott M. Q., Dogucu M. (2022) *Bayes Rules! An Introduction to Applied Bayesian Modeling* (1st) Chapman and Hall: CRC.

Kang P, Tobler P N., Dayan P (2024) Bayesian reinforcement learning: A basic overview. *Neurobiology of Learning and Memory* **211**:107924 https://doi.org/10.1016/j.nlm.2024.107924 | PubMed

Lebreton M., Langdon S., Slieker M. J., Nooitgedacht J. S., Goudriaan A. E., Denys D., Van Holst R. J., Luigjes J. (2018) Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. *Science Advances* **4**:eaaq0668 https://doi.org/10.1126/sciadv.aaq0668 |

PubMed

Lee D. G., Daunizeau J. (2021) Trading mental effort for confidence in the metacognitive control of value-based decision-making. *eLife* **10**:e63282 https://doi.org/10.7554/eLife.63282 | PubMed

Lee D. G., Usher M. (2023) Value certainty in drift-diffusion models of preferential choice. *Psychological Review* **130**:790-806 https://doi.org/10.1037/rev0000329 | PubMed

Lefebvre G., Lebreton M., Meyniel F., Bourgeois-Gironde S., Palminteri S. (2017) Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour* **1**:0067 https://doi.org/10.1038/s41562-017-0067

Li H.-H., Ma W. J. (2020) Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. *Nature Communications* **11**:2004 https://doi.org/10.1038/s41467-020-15581-6 | PubMed

Lisi M., Mongillo G., Milne G., Dekker T., Gorea A. (2020) Discrete confidence levels revealed by sequential decisions. *Nature Human Behaviour* **5**:273-280 https://doi.org/10.1038/s41562-020-00953-1 | PubMed

Maniscalco B., Peters M. A. K., Lau H. (2016) Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception, & Psychophysics* **78**:923-937 https://doi.org/10.3758/s13414-016-1059-x | PubMed

Meyniel F. (2020) Brain dynamics for confidence-weighted learning. *PLOS Computational Biology* **16**:e1007935 https://doi.org/10.1371/journal.pcbi.1007935 | PubMed

Meyniel F., Schlunegger D., Dehaene S. (2015) The Sense of Confidence during Probabilistic Learning: A Normative Account. *PLOS Computational Biology* **11**:e1004305 https://doi.org/10.1371/journal.pcbi.1004305 | PubMed

Meyniel F., Sigman M., Mainen Z. F. (2015) Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron* **88**:78-92 https://doi.org/10.1016/j.neuron.2015.09.039 | PubMed

Miyoshi K., Sakamoto Y (2025) Early sensory evidence shapes multichoice confidence bias: A registered report. *Journal of Experimental Psychology: General* https://doi.org/10.1037/xge0001809 | PubMed

Nassar M. R., Wilson R. C., Heasly B., Gold J. I. (2010) An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. *The Journal of Neuroscience* **30**:12366-12378 https://doi.org/10.1523/jneurosci.0822-10.2010 | PubMed

Navajas J., Hindocha C., Foda H., Keramati M., Latham P E., Bahrami B. (2017) The idiosyncratic nature of confidence. *Nature Human Behaviour* **1**:810-818 https://doi.org/10.1038/s41562-017-0215-1 | PubMed

Nunez M. D., Fernandez K., Srinivasan R., Vandekerckhove J. (2024) A tutorial on fitting joint models of M/EEG and behavior to understand cognition. *Behavior Research Methods* **56**:6020-6050 https://doi.org/10.3758/s13428-023-02331-x | PubMed

Palminteri S., Lebreton M. (2022) The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences* **26**:607-621 https://doi.org/10.1016/j.tics.2022.04.005 | PubMed

Palminteri S., Lefebvre G., Kilford E. J., Blakemore S.-J. (2017) Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLOS Computational Biology* **13**:e1005684 https://doi.org/10.1371/journal.pcbi.1005684 | PubMed

Paunov A., L'Hôtellier M., Guo D., He Z., Yu A., Meyniel F. (2024) Multiple and subject-specific roles of uncertainty in reward-guided decision-making. *eLife* **13**:RP103363 https://doi.org/10.7554/eLife.103363

Payzan-LeNestour E., Bossaerts P (2011) Risk, Unexpected Uncertainty, and Estimation Uncertainty: Bayesian Learning in Unstable Settings. *PLoS Computational Biology* **7**:e1001048 https://doi.org/10.1371/journal.pcbi.1001048 | PubMed

Peters M. A. K., Thesen T., Ko Y D., Maniscalco B., Carlson C., Davidson M., Doyle W., Kuzniecky R., Devinsky O., Halgren E., *et al.* (2017) Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour* **1**:0139 https://doi.org/10.1038/s41562-017-0139 | PubMed

Pouget A., Drugowitsch J., Kepecs A. (2016) Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience* **19**:366-374 https://doi.org/10.1038/nn.4240 | PubMed

Quandt J., Figner B., Holland R. W., Veling H. (2022) Confidence in evaluations and value-based decisions reflects variation in experienced values. *Journal of Experimental Psychology: General* **151**:820-836 https://doi.org/10.1037/xge0001102 | PubMed

Rescorla R. A., Wagner. A. R. (1972) A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In: Black A. H.., Prokasy. W. F. (Eds). *Classical Conditioning II: Current Research and Theory* Appleton-Century-Crofts. pp. 64-99

Rausch M., Hellmann S., Zehetleitner M. (2018) Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception, & Psychophysics* **80**:134-154 https://doi.org/10.3758/s13414-017-1431-5 | PubMed

Rausch M., Zehetleitner M. (2019) The folded X-pattern is not necessarily a statistical signature of decision confidence. *PLOS Computational Biology* **15**:e1007456 https://doi.org/10.1371/journal.pcbi.1007456 | PubMed

Salem-Garcia N., Palminteri S., Lebreton M. (2023) Linking confidence biases to reinforcement-learning processes. *Psychological Review* **130**:1017-1043 https://doi.org/10.1037/rev0000424 | PubMed

Sanders J. I., Hangya B., Kepecs A. (2016) Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron* **90**:499-506 https://doi.org/10.1016/j.neuron.2016.03.025 | PubMed

Schulz E., Gershman S. J. (2019) The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology* **55**:7-14 https://doi.org/10.1016/j.conb.2018.11.003 | PubMed

Schulz L., Fleming S. M., Dayan P (2023) Metacognitive computations for information search: Confidence in control. *Psychological Review* **130**:604-639 https://doi.org/10.1037/rev0000401 | PubMed

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**:656-688 https://doi.org/10.1037/xge0001524 | PubMed

Stan Development Team (2025) Stan. version: v2.37.0 https://mc-stan.org

Stephan K. E., Penny W. D., Daunizeau J., Moran R. J., Friston K. J. (2009) Bayesian model selection for group studies. *NeuroImage* **46**:1004-1017 https://doi.org/10.1016/j.neuroimage.2009.03.025 | PubMed

Ting C.-C., Salem-Garcia N., Palminteri S., Engelmann J. B., Lebreton M. (2023) Neural and computational underpinnings of biased confidence in human reinforcement learning. *Nature Communications* **14**:6896 https://doi.org/10.1038/s41467-023-42589-5 | PubMed

Wagenmakers E.-J., Farrell S. (2004) AIC model selection using Akaike weights. *Psychonomic Bulletin & Review* **11**:192-196 https://doi.org/10.3758/BF03206482 | PubMed

Xue K., Shekhar M., Rahnev D. (2024) Challenging the Bayesian confidence hypothesis in perceptual decision-making. *Proceedings of the National Academy of Sciences* **121**:e2410487121 https://doi.org/10.1073/pnas.2410487121 | PubMed

Yeung N., Summerfield C. (2012) Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**:1310-1321 https://doi.org/10.1098/rstb.2011.0416 | PubMed

Zylberberg A., Barttfeld P., Sigman M. (2012) The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience* **6** https://doi.org/10.3389/fnint.2012.00079 | PubMed

Peer reviews

Reviewer #1 (Public review):

Summary:

This study addresses an important question in reinforcement learning and metacognition by distinguishing value confidence from decision confidence and testing how each is computationally represented. The findings are significant because they suggest that value confidence is well captured by Bayesian uncertainty, whereas decision confidence reflects a hybrid computation combining probability correct with broader value certainty. The evidence is promising, supported by multiple datasets and model comparisons.

Strength.

(1) A major strength of the study is that the authors test their hypotheses across multiple datasets, including previously published datasets and newly collected data. This broad empirical approach increases the generality of the findings.

(2) The Bayesian model of value confidence has a clear theoretical basis. The proposed hybrid model of decision confidence is also intuitive. It appears to capture important aspects of the decision confidence data.

(3) The paper provides a useful framework for linking how certainty about value estimates guides the subsequent choice and the corresponding decision confidence.

Weakness

(1) The conceptual link between value confidence and decision confidence is not yet fully established. The manuscript argues that overall value certainty contributes to decision confidence, but this conclusion is based largely on the latent variable that the model infers from the decision confidence experiment alone. A more direct test would require measuring value confidence and decision confidence within the same participants and task, and analysing how these two types of confidence interact.

(2) The individual-difference analyses in Figure 5 are methodologically challenging. The predictors used in these analyses are derived from model fits to the behavioural data and are then correlated to behaviour in the same task. This creates a risk that correlations inevitably arise. Thus, it does not assure that correlations are cognitively meaningful.

(3) The model recovery results suggest that some candidate models are not clearly distinguishable.

(4) The manuscript would benefit from clearer explanations of why specific models capture particular behavioural patterns.

(5) The claim that value confidence modulates the exploration-exploitation trade-off should be interpreted carefully, because the model uses global uncertainty across both options, not option-specific value confidence.

https://doi.org/10.7554/eLife.111820.1.sa2

Reviewer #2 (Public review):

Summary:

In this work, the authors propose a common value-estimation framework based on Bayesian inference and show that it can account for both participants' confidence in their value estimates ("value confidence") and for their confidence in their final choices ("decision confidence").

Strengths:

The study extends several established findings in the confidence and reinforcement-learning literature. In particular, the authors not only examine decision confidence but also directly model value confidence, and they replicate the idea that decision confidence reflects a combination of multiple computations, previously described for categorical decisions (Navajas et al., 2017), in the context of continuous value-based decisions. I therefore consider the work a useful contribution to the field.

Weaknesses:

However, I believe that the scope of the conclusions is overstated relative to the results that are actually presented.

(1) Interaction between value confidence and decision confidence

The abstract and introduction frame the study as addressing a major gap in the literature, namely, the lack of direct investigation of the interaction between value confidence and decision confidence. Yet the manuscript never directly tests the interaction between these two quantities. Instead, the authors show that the reported decision confidence depends not only on the probability of being correct, but also on the precision of the decision variable DV, which is related to the precision of the value estimates underlying value confidence. While this is related to the proposed research question, it is not a direct analysis of the interaction *between* value confidence and decision confidence themselves.

(2) Unified computational framework

Similarly, the claim that the study provides a "unified computational framework" appears somewhat overstated. The proposed models build on standard and well-established Bayesian frameworks and extend them specifically to account for decision confidence. While this demonstrates that both forms of confidence can be expressed within a common Bayesian formalism, the manuscript does not establish a direct computational interaction or shared mechanism between them beyond their dependence on the same underlying uncertainty estimates.

(3) "Phenotypes" interpretation

The interpretation of the observed individual differences as distinct "behavioural phenotypes" also appears overstated. The reported analyses primarily show continuous variability across participants in the relative weighting of different components contributing to confidence reports, rather than evidence for qualitatively distinct categories or computational subtypes of decision-makers.

(4) Decision confidence terminology

I also found some conceptual ambiguity in the terminology used throughout the manuscript. Early in the paper, decision confidence is defined normatively as the subjective probability of having made the correct choice, corresponding to P(DV>0). Later, however, the authors show that participants' confidence reports are better explained by a combination of this probability and the precision of the decision-variable distribution. Despite this distinction, the manuscript continues referring to the reported quantity simply as "decision confidence." Clarifying the distinction between the theoretical construct and the empirical reports (for example, by referring to "reported decision confidence") would improve conceptual clarity.

https://doi.org/10.7554/eLife.111820.1.sa1

Reviewer #3 (Public review):

Summary:

Comay, Solovey, and Barttfeld aim to provide a unified computational account of confidence in reinforcement learning by distinguishing value confidence-the certainty associated with latent value estimates-from decision confidence-the confidence that a particular choice is correct. Across new experiments and reanalyses of previously published datasets, they argue that value confidence is best described by Bayesian posterior precision, that this form of confidence adaptively reduces decision noise as learning progresses, and that decision confidence is better captured by a hybrid model combining Bayesian probability correct with a more global estimate of value certainty. They further propose that individual differences in the relative weighting of these components define "confidence phenotypes" that predict task performance, exploration-exploitation behavior, and metacognitive accuracy.

Strengths:

A major strength of the study is that it addresses an important conceptual distinction that is often blurred in the confidence literature. The paper usefully separates uncertainty about latent environmental states from confidence in an action derived from those latent beliefs. This distinction is especially important in reinforcement learning, where uncertainty is not merely a retrospective judgment about accuracy but can directly shape future sampling, learning, and action selection. The manuscript is therefore well positioned to bridge work on Bayesian confidence in perceptual decision-making with work on uncertainty-guided learning and exploration.

A second strength is the authors' use of multiple datasets and model comparisons. The claim that value confidence tracks Bayesian uncertainty is supported across tasks in which participants explicitly report confidence in value estimates, including datasets where reward variance is manipulated. The latter manipulation is particularly useful because it helps distinguish a Bayesian uncertainty account from simpler models based only on the number of observations. The finding that value confidence modulates the softmax slope and thereby promotes more exploitative choices as uncertainty decreases is also theoretically coherent and supported across several datasets, including a preregistered replication.

The manuscript's most interesting and potentially impactful contribution is the hybrid model of decision confidence. The authors show that a model based only on Bayesian probability correct captures confidence on correct trials better than on incorrect trials, whereas adding an "overall value confidence" term improves the fit. This is a useful result because it suggests that confidence reports in reinforcement learning may not be a pure readout of decision-level discriminability, but instead may combine decision-specific evidence with more global latent-state uncertainty. This could help explain why human confidence often deviates from ideal Bayesian predictions, especially on error trials.

Weaknesses:

However, the interpretation of the hybrid model remains the main weakness of the paper. The second term, overall value confidence, is not equivalent to the precision of the decision variable. It can dissociate from decision difficulty: two options can be far apart but individually uncertain, or nearly identical but individually well estimated. The authors appear to recognize this issue and have reframed the term as "overall value confidence" rather than decision-variable precision. This is a useful clarification, but the conceptual role of the term still requires sharper treatment. In its current form, it is sometimes described as part of a unified confidence computation, but it may be more accurately understood as a

biasing or contextual signal that modulates reported confidence without necessarily improving decision calibration.

A related concern is model identifiability. In many reinforcement-learning tasks, probability correct and overall value confidence both change systematically over the course of learning. As a result, the hybrid model may gain predictive power partly because it captures generic time-on-task or learning-progress effects, rather than because participants explicitly combine two separable uncertainty signals. The manuscript would be stronger if it more clearly demonstrated that the two latent variables are distinguishable in the behavioral data, for example, through model recovery, parameter recovery, cross-validated prediction, and analyses of the correlation between latent regressors across task conditions and individuals.

The link between the decision rule and confidence model also deserves more scrutiny. The authors use value confidence to modulate decision noise in the choice model, and then use a related global value-confidence term in the confidence-report model. This creates an appealing unified architecture, but it also raises the possibility that the same latent variable is doing multiple kinds of explanatory work. The paper would benefit from a clearer separation between uncertainty as a driver of choices, uncertainty as a determinant of confidence reports, and uncertainty as an inferred latent variable extracted from the same behavioral data.

From a computational neuroscience perspective, the manuscript would also benefit from a more explicit discussion of how these confidence quantities might be represented neurally. The current model treats value confidence, probability correct, and overall value confidence as scalar latent variables available to the observer. Yet uncertainty-related computations may be represented nonlinearly in neural population activity rather than as explicit scalar readouts. Work on nonlinear neural decoding and population codes has shown that task-relevant variables can be carried by nonlinear statistics of neural activity, especially when nuisance variables obscure mean tuning, and that behavioral choices can reveal whether such nonlinear information is efficiently decoded. This literature provides a useful framework for connecting the present behavioral model to possible neural implementations of value and decision confidence.

Overall, the authors largely achieve their goal of demonstrating that value confidence and decision confidence are computationally dissociable in reinforcement learning. The evidence for Bayesian value confidence is strong, and the evidence that confidence-guided exploitation improves the account of choice behavior is convincing. The evidence for the hybrid account of decision confidence is promising but would be strengthened by additional analyses clarifying model identifiability, the interpretation of the overall value-confidence term, and the conditions under which the model makes distinct predictions from simpler time-, value-, or evidence-based alternatives. The paper is likely to be useful for researchers interested in computational models of confidence, metacognition, and adaptive behavior under uncertainty.

https://doi.org/10.7554/eLife.111820.1.sa0