Corrective feedback guides human perceptual decision-making by informing about the world state rather than rewarding its choice

Short title: Corrective feedback on perceptual decisions

Hyang-Jung Lee[1], Heeseung Lee[1], Chae Young Lim[2], Issac Rhim[3], Sang-Hun Lee[1*]

[1]Department of Brain and Cognitive Sciences, Seoul National University, Seoul, South Korea

[2]Department of Statistics, Seoul National University, Seoul, South Korea

[3]Institute of Neuroscience, University of Oregon, Eugene, OR, USA

*Corresponding author

E-mail: visionsl@snu.ac.kr

Abbreviations

PDM, perceptual decision-making; VDM, value-based decision-making; RL, reinforcement learning; BDT, Bayesian decision theory; *toi*, trial of interest; PSE, point of subjective equality; BMBU, Bayesian model of boundary updating; AICc, Akaike information criterion corrected for sample size

Keywords

Feedback; History effects; Reinforcement learning; Perceptual decision-making; Bayesian inference; Binary classification

Abstract

Corrective feedback received on perceptual decisions is crucial for adjusting decision-making strategies to improve future choices. However, its complex interaction with other decision components, such as previous stimuli and choices, challenges a principled account of how it shapes subsequent decisions. One popular approach, based on animal behavior and extended to human perceptual decision-making, employs 'reinforcement learning,' a principle proven successful in reward-based decision-making. The core idea behind this approach is that decision-makers, although engaged in a perceptual task, treat corrective feedback as rewards from which they learn choice values. Here, we explore an alternative idea, which is that humans consider corrective feedback on perceptual decisions as evidence of the actual state of the world rather than as rewards for their choices. By implementing these 'feedback-as-reward' and 'feedback-as-evidence' hypotheses on a shared learning platform, we show that the latter outperforms the former in explaining how corrective feedback adjusts the decision-making strategy along with past stimuli and choices. Our work suggests that humans learn about what has happened in their environment rather than the values of their own choices through corrective feedback during perceptual decision-making.

Introduction

Perceptual decision-making (PDM) means committing to a proposition about an objective world state (e.g., "The temperature today is low."). Decision-makers adjust future commitments based on what they experienced from past commitments, including what they perceived, what they chose, and what the environment gave them in return. Among these *history factors*, *trial-to-trial corrective feedback*—feedback about the correctness of a decision maker's choices on a trial-

to-trial basis—is widely used by experimenters to train subjects on PDM tasks. Despite this clear utility of feedback and a pile of evidence for its impact on subsequent PDM behavior across species and sensory modalities [1–11], much remains elusive about how corrective feedback, in conjunction with other history factors, exerts its trial-to-trial influence on subsequent decisions.

Unlike PDM, value-based decision-making (VDM) involves making choices based on decision makers' subjective preferences (e.g., 'choosing between two drinks based on their tastes'). Reinforcement learning (RL) algorithms have proven effective in explaining how past rewards affect future VDM based on error-driven incremental mechanisms [12–18]. Intriguingly, there have been attempts to explain the impact of past feedback on subsequent PDM by grafting an RL algorithm onto the PDM processes [3,4,8–10]. This grafting premises that decision-makers treat corrective feedback in PDM similarly to reward feedback in VDM. On this premise, this RL-grafting account proposes that decision-makers update the *value* of their choice to minimize the difference between the expected reward and the actual reward received, called 'reward prediction error' (red dashed arrows in Fig 1A). Importantly, the amount of reward prediction error is inversely related to the strength of sensory evidence—i.e., the extent to which a given sensory measurement of the stimulus supports the choice—because the expected value becomes low as the sensory evidence becomes weak. For example, suppose a decision-maker committed to a proposition, "The temperature today is low." Then, $correct$ feedback to that commitment increases the value of the 'low' choice since the positive reward for the 'low' choice leads to the positive reward prediction error, which indicates the need to heighten the value of the 'low' choice. Importantly, the amount of value-updating is greater when the experienced temperature is moderately cold (e.g., −2°c, weak sensory evidence for the 'low' choice) compared to when it is very cold (e.g., −15°c, strong sensory evidence for the 'low' choice) because the expected reward is smaller in the former, which leads to a greater level of reward

prediction error compared to the latter (as illustrated in the left panel of Fig 1B). A recent study [9] referred to this sensory evidence-dependent impact of feedback as “confidence-guided choice updating” based on the tight linkage between decision confidence and sensory evidence. This RL-grafting account, referred to as the *value-updating scenario* hereinafter, appears natural given that corrective feedback is typically provided as physical rewards such as juice or water in animal PDM experiments [4,5,8–10,19–21]. The value-updating scenario seems plausible from the perspective that PDM and VDM might share common mechanisms [22], as suggested by some empirical studies [23,24].

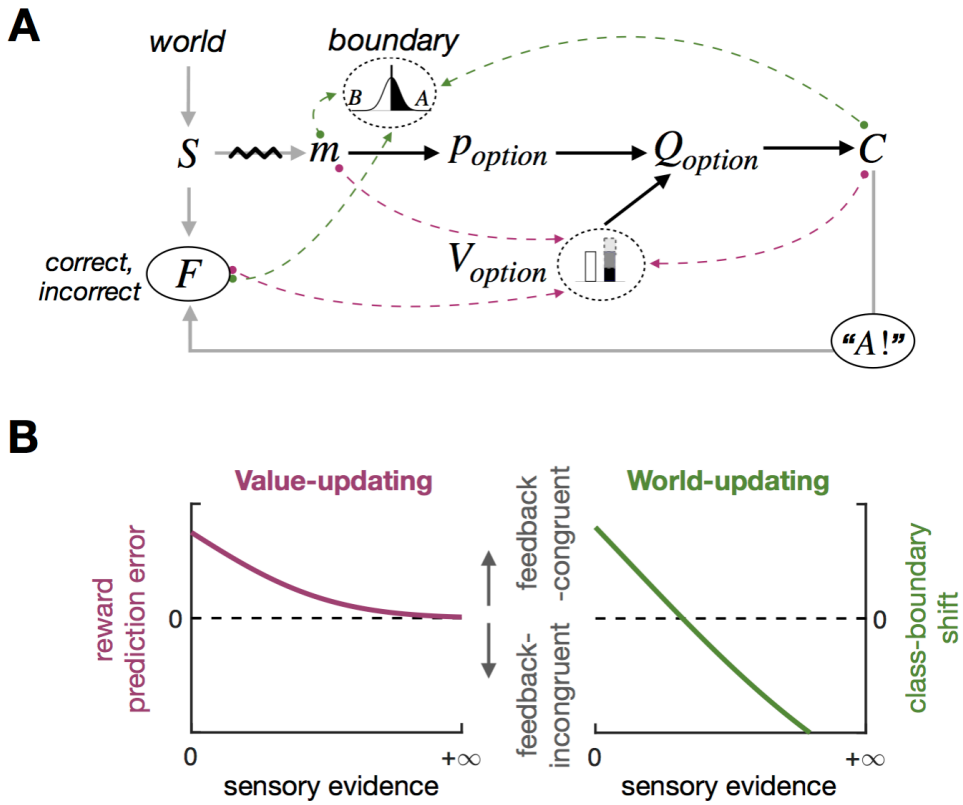


Fig 1. Two possible scenarios for what humans learn from feedback for PDM and their distinct predictions of feedback effects. (A) Decision-making platform for perceptual binary classification. The gray arrows depict how a sensory measurement m and feedback F are generated from a stimulus S, which is sampled from the $world$, and a choice C. The black arrows depict the computational process, where, for a given choice $option$, a decision-maker computes its expected value Q_{option} by multiplying the probability that the choice is correct p_{option} given m and the class boundary B with the value of that choice V_{option} and make a choice C based on Q_{option}. In principle, the decision-maker may update either V_{option} (red dashed arrows; value-updating) or $world$ (green dashed arrows; world-updating) from m, C, and F. **(B)** Distinct sensory-evidence-dependent feedback effects predicted by the value-updating and

world-updating scenarios. According to the value-updating scenario (left), as sensory evidence becomes stronger, p_{option} increases, and accordingly, so does Q_{option}. As a result, reward prediction errors become smaller but remain in the direction congruent with feedback, which predicts that feedback effects on subsequent trials diminish asymptotically as a function of the strength of sensory evidence. According to the world-updating scenario (right), as sensory evidence becomes stronger, the stimulus distribution, and accordingly B too, becomes shifted farther towards the stimulus in the direction counteracting the influence of feedback. As a result, the direction of feedback effects is the same as that predicted by the value-updating scenario for weak sensory evidence but eventually reverses to the direction incongruent with feedback as sensory evidence becomes stronger.

Nevertheless, value-updating might not be the only route through which feedback effects transpire in PDM, especially for humans receiving corrective feedback without any physical rewards. Alternatively, decision-makers may treat feedback not as rewards but as a logical indicator of whether the proposition they committed to is true or false in the world. In this scenario, decision-makers update their belief about world statistics (i.e., stimulus distribution) by combining the information about the trueness of their choice, which is informed by feedback, and the information about the stimulus, which is informed by a sensory measurement (dashed arrow from m in Fig 1A). Suppose you have recently arrived in Canada for the first time in the winter and felt the chilly air. You remarked, "The temperature today is low." Your friend, who has lived for long in Canada, may agree or disagree with you, and this will provide you with information on the typical temperature distribution during the Canadian winter. The $incorrect$ feedback from your friend (e.g., "Actually, it's not low at all today.") indicates that the temperature experienced today falls on the higher side of the actual distribution, making you adjust your belief about the distribution to the lower side. On the contrary, the $correct$ feedback (e.g., “Yes, it’s low today.”) will lead you to adjust your belief about the distribution to the higher side. It is important to note that, besides the feedback from your friend, the temperature felt by yourself also informs you of the statistical distribution of temperature since it is a sample from that distribution. For instance, if the temperature felt moderately cold (e.g., –2°c), your belief about the temperature distribution will only slightly shift towards the lower side. However, if it felt very cold (e.g., –15°c), your belief will shift towards the same lower side, but with a much

greater amount, which can counteract the impact of the $correct$ feedback on your belief (i.e., adjusting your belief to the higher side).

Therefore, according to this alternative scenario, referred to as the *word-updating scenario* hereinafter, $correct$ feedback to "The temperature today is low." will increase the tendency to classify the next day's temperature as 'low,' just like the value-updating scenario. However, unlike the value-updating scenario, the world-updating scenario implies that when sensory evidence is too strong, such a tendency can be reversed, leading to a counterintuitive increase in the tendency to classify the next day's temperature as 'high,' (as illustrated in the right panel of Fig 1B). The world-updating scenario is conceptually parsimonious because it does not require any component outside the PDM processes, such as the RL algorithms developed in the VDM. Especially in Bayesian Decision Theory (BDT) [25,26], which has been providing compelling accounts for PDM behavior, world statistics is a crucial knowledge that is required to infer a world state in PDM [27–30].

Here we tested which of the two scenarios better explains the effects of corrective feedback—without any physical reward—on humans' PDM. To do so, we implemented the value-updating and world-updating scenarios into a variant of RL model [9] and a Bayesian model, respectively, and directly compared the two models' accountability for the feedback effects on humans' PDM behavior. As a PDM task, we opted for a binary classification task, one most widely used PDM task in which decision-makers sort items into two discrete classes by setting a boundary since the two scenarios make distinct predictions about the stimulus-dependent feedback effects in this task. As was described intuitively above and will be explained rigorously later, the value-updating scenario predicts that feedback, which acts like rewards, "uni-directionally" fosters and suppresses the rewarded ($correct$) and unrewarded ($incorrect$) choices, respectively, in subsequent trials while diminishing its impact asymptotically as sensory evidence becomes stronger, due to the reduction in reward prediction error (the red

curve in Fig 1B). By contrast, the world-updating scenario predicts that the feedback effects not just diminish but eventually become reversed to the opposite side as sensory evidence becomes stronger, as the shift of the class boundary towards the previous stimulus counteracts the boundary shift due to feedback (the green curve in Fig 1B).

We found the world-updating model superior to the value-updating model in explaining human history effects of corrective feedback on PDM. Critically, the value-updating model fails to account for the observed stimulus-dependent feedback effects. Our findings suggest that humans are likely to treat corrective feedback in PDM as logical indicators of the trueness of the proposition to which they committed, rather than as rewards, and update their knowledge of world statistics, rather than the values of their choices, based on feedback in conjunction with the other history factors—previous stimuli and choices.

Results

Quantifying the retrospective and prospective history effects of feedback on binary classification

To study the stimulus-dependent feedback effects in PDM, we acquired long sequences (170 trials/sequence) of binary choices ($C \in \{small, large\}$) many times (30 sequences/participant) from each of 30 human participants while varying the ring size ($S \in \{-2, -1, 0, 1, 2\}$) and providing corrective feedback ($F \in \{correct, incorrect\}$) (Fig 2A). On each trial, participants viewed a ring, judged whether its size is $small$ or $large$ as accurately as possible while receiving feedback, which indicated by color whether the choice was correct or incorrect (Fig 2B). We ensured the ring size varied sufficiently—including the ones very easy and difficult for classification—so that the two scenarios' distinct predictions on the stimulus-dependent feedback effects could be readily compared. Also, we used stochastic feedback, where $correct$ and $incorrect$ feedback

was occasionally given to incorrect and correct choices, respectively, to cover the entire 3D space of decision-making episodes defined orthogonally over ‘stimulus’, ‘choice’, and ‘feedback’ ($5 \times 2 \times 2 = 20$ episodes; Fig 2C; Materials and methods).

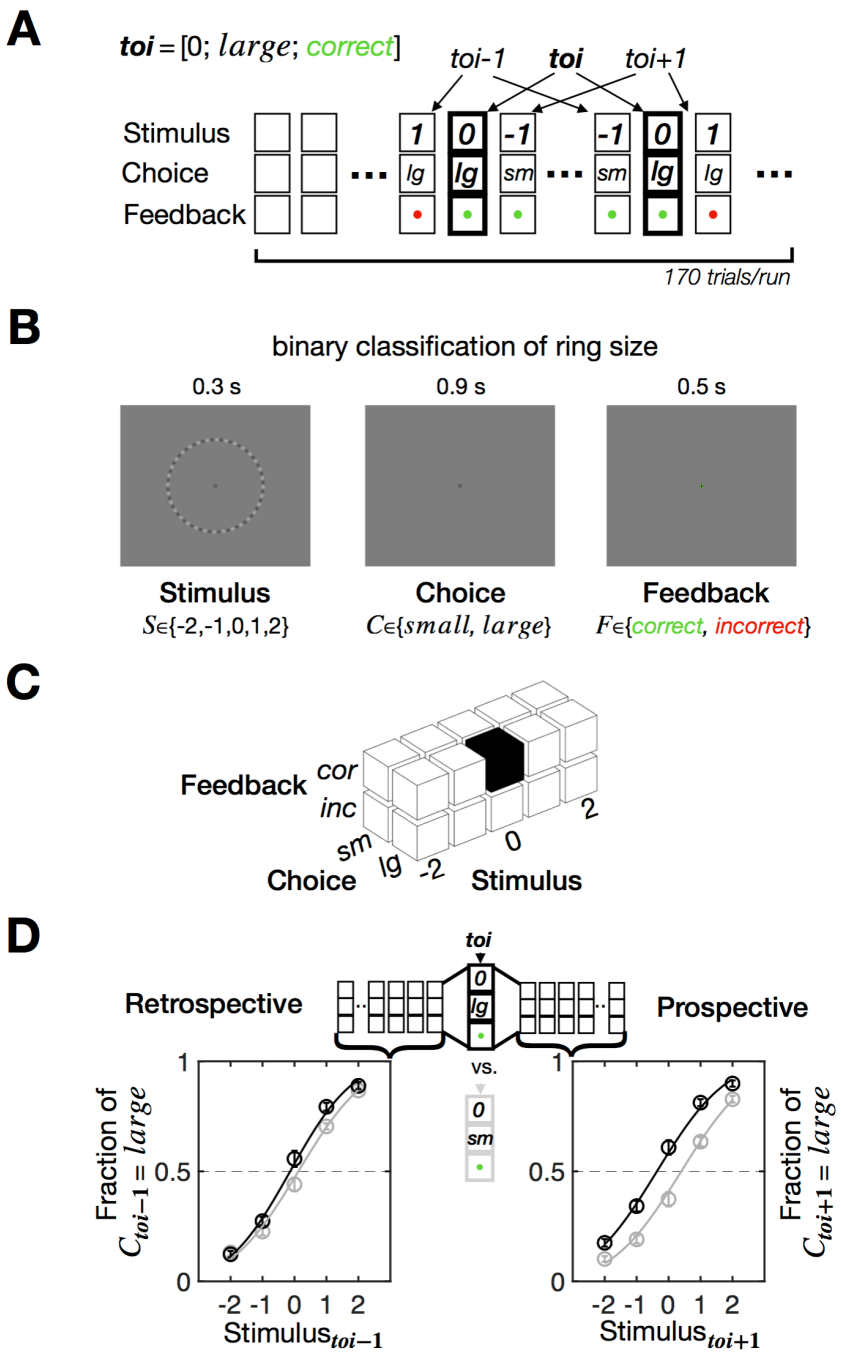


Fig 2. Experimental design and definition of retrospective and prospective history effects. (A) A chain of PDM episodes over a single sequence of trials. Each trial sequence consists of 170 column vectors of PDM episode [stimulus; choice; feedback]. In this example, the trial of interest (*toi*) is characterized by an episode vector $[0;\ large;\ correct]$ and demarcated by thick outlines. The trials that

precede and follow *toi* can be labeled as *toi-1* and *toi+1*, respectively. **(B)** Trial structure. Participants viewed a randomly sampled ring with their eyes fixed, classified its size, and then received feedback indicating whether the classification was correct or incorrect by the color around the fixation. **(C)** The 3D state space of the PDM episodes in the experiment. The example episode of *toi* in **(A)** is marked by the black cube. **(D)** Definition of retrospective and prospective history effects. As illustrated in **(A)** and **(C)**, for any given episode of *toi*, all the trials labeled with *toi-1* and *toi+1* are stacked and used to derive the psychometric curves, respectively. The PSEs estimated for the *toi-1* and *toi+1* psychometric curves quantify the retrospective and prospective history effects, respectively. In this example, the black and gray curves were defined for *toi* = $[0;\ large;\ correct]$ and *toi* = $[0;\ small;\ correct]$, respectively, with circles and bars representing the mean and s.e.m. across 30 participants, respectively. The data underlying this figure **(D)** can be found in S1 Data.

To rigorously evaluate the correspondence between model prediction and human behavior, we quantified the history effects in both retrospective and prospective directions of time, as follows (Fig 2D). First, we localized the trials in which a PDM episode of interest occurred (trial of interest, *toi*) and stacked the trials that preceded (the retrospective block of trials, *toi-1*) and those that followed (the prospective block of trials, *toi+1*) the *toi*. Second, we derived the two psychometric curves from the retrospective and prospective blocks of trials, respectively, and fit the cumulative normal distribution function to these curves to estimate the point-of-subjective-equality (PSE) measures, which have previously been used [19–21] and known to reliably estimate the history-dependent choice biases in PDM [31]. Thus, the PSEs of the retrospective and prospective trials quantify the choice biases that exist *before* and *after* the PDM episode of interest occurs, respectively, with negative and positive values signifying that choices are biased to $large$ and $small$, respectively.

Decision-making processes for binary classification

As a first step of evaluating the value-updating and world-updating scenarios, we constructed a common platform of decision-making for binary classification where both scenarios play out. This platform consists of three processing stages (Fig 3A). At the stage of ‘perception’, the decision-maker infers the class probabilities, i.e., the probabilities that the ring size (S) is larger and smaller, respectively, than the class boundary (B) given a noisy sensory measurement (m), as follows:

$$p(CL = large) = p(S > B|m) = \int_B^{\infty} p(S|m)dS;$$

$$p(CL = small) = 1 - p(CL = large),$$

where CL stands for the class variable with the two ($small$ and $large$) states.

At the stage of 'valuation', the decision-maker forms the expected values for the two choices (Q_{large} and Q_{small}) by multiplying the class probabilities by the learned values of the corresponding choices (V_{large} and V_{small}) as follows:

$$Q_{large} = p(CL = large) \times V_{large};$$

$$Q_{small} = p(CL = small) \times V_{small}.$$

Lastly, at the stage of 'decision', the decision-maker commits to the choice whose expected value is greater than the other. In this platform, choice bias may originate from the perception or valuation stage. Suppose the decision-maker's belief about size distribution at the perception stage is not fixed but changes depending on previous PDM episodes (Fig 3B, top). Such changes lead to the changes in PSE of the psychometric curve because the class probabilities change as the class boundary changes (Fig 3B, bottom). Alternatively, suppose the decision-maker's learned values of the choices are not fixed but change similarly (Fig 3C, top). These changes also lead to the changes in PSE of the psychometric curve because the expected values change as the choice values change (Fig 3C, bottom).

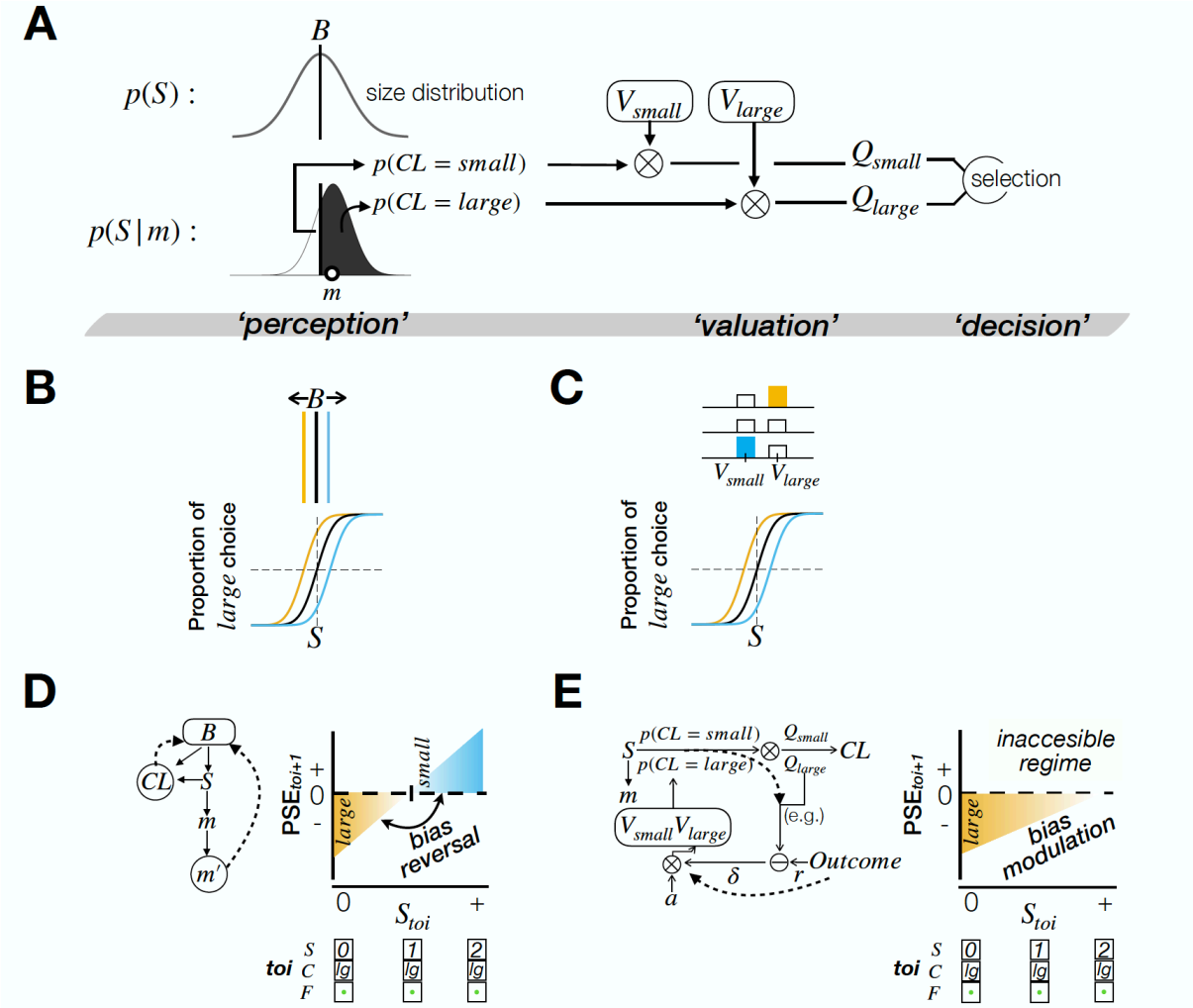


Fig 3. Implementation of the value-updating and world-updating scenarios into computational models in a common PDM platform. (A) Computational elements along the three stages of PDM for binary classification. At the '*perception*' stage, the probabilities that the class variable takes its binary states $small$ and $large$—$p(CL = large)$ and $p(CL = small)$—are computed by comparing the belief on the stimulus size $p(S|m)$ against the belief on the class boundary B—the mean of the belief on stimulus distribution in the world $p(S)$. At the '*valuation*' stage, the outcomes of the perception stage are multiplied by the learned values Vs to produce the expected values Qs. At the '*decision*' stage, the choice with the greater expected value is selected. **(B, C)** Illustration of two potential origins of choice biases, one at the '*perception*' stage **(B)** and the other at the '*valuation*' stage **(C)**. The color indicates the direction of choice bias (yellow for bias to $large$; balck for no bias; blue for bias to $small$). **(D, E)** Illustration of the architectures (left panels) and predictions on the stimulus-dependent feedback effects (right panels) of BMBU **(D)** and the belief-based RL model **(E)**. In the left panels, the dashed arrows represent the ways the history factors (feedback and stimulus) exert their contribution to choice bias. In the right panels, PSE_{toi+1}, which quantifies the choice bias in the trials following a certain PDM episode at *toi*=[0; $large$; *correct*], is plotted as a function of the stimulus size at *toi*. The color indicates the direction of choice bias, as in **(B)** and **(C)**.

The belief-based RL model

To implement the value-updating scenario, we adapted the belief-based RL model [9] to the current experimental setup. Here, feedback acts like a reward by positively or negatively reinforcing the value of choice ($V_{large(small)}$) with the deviation of the reward outcome (r) from the expected value of that choice ($Q_{large(small)}$), as follows:

$$V_{large(small)} \leftarrow V_{large(small)} + \alpha\delta;$$

$$\delta = r - Q_{large(small)} = r - p(CL = large(small)) \times V_{large(small)},$$

where α, δ, and r are the learning rate, the reward prediction error, and the reward, respectively. The state of feedback determines the value of r: $r = 1$ for $correct$; $r = 0$ for $incorrect$. Note that δ has the statistical decision confidence at the perception stage, i.e., $p(CL = large(small))$, as one of its three arguments. As stressed by the authors who developed this algorithm [9], this feature makes the strength of sensory evidence—i.e., statistical decision confidence—modulate the degree to which the decision-maker updates the chosen value based on feedback (Fig 3E, left). Hence, this belief (confidence)-based modulation of value-updating underlies the stimulus-dependent feedback effects: the amount of feedback effects decreases as sensory evidence becomes stronger since the reward prediction error decreases as a function of $p(CL = large(small))$, which is proportional to sensory evidence (Fig 3E, right).

The Bayesian model of boundary-updating (BMBU)

To implement the world-updating scenario, we developed BMBU, which updates the class boundary based on the previous PDM episode in the framework of BDT. Specifically, given 'a state of the class variable that is indicated jointly by feedback and choice', CL, and 'a noisy memory recall of the sensory measurement (, which will be referred to as "mnemonic measurement" hereinafter)', m', BMBU infers the mean of the size distribution (i.e., class boundary), B, by updating its prior belief about B, $p(B)$, with the likelihood of B, $p(m', CL|B)$, by

inverting its learned generative model of how m' and CL are generated (Fig 3D, left; Equations 3-6 in Materials and methods for the detailed formalisms for the learned generative model), as follows:

$$p(B|m', CL) \propto p(m', CL|B)p(B) \equiv p(m', C, F|B)p(B).$$

This inference uses multiple pieces of information from the PDM episode just experienced, including the mnemonic measurement, choice, and feedback, to update the belief about the location of the class boundary (refer to Equations 8-14 in Materials and methods for more detailed formalisms for the inference). In what follows, we will explain why and how this inference leads to the specific stimulus-dependent feedback effects predicted by the world-updating scenario (Fig 3D, right), where world knowledge is continuously updated.

Suppose a decision-maker currently believes that the size distribution is centered around 0. Let's first consider a case where the decision-maker experiences a PDM episode with an ambiguous stimulus: the ring with size 0 is presented and produces a sensory measurement m that is only slightly greater than 0 (through the stochastic process where m is generated from S; Equation 5), which leads to the $large$ choice since the inferred S from such m is greater than the center of the size distribution (Equation 4 and 7), and then followed by $correct$ feedback. BMBU predicts that after this PDM episode, the decision-maker will update the belief about the size distribution by shifting it towards the smaller side. Hence, the choice in the next trial will be biased towards the larger option, resulting in a negatively biased PSE for the psychometric curve defined by the trials following the episode of interest. This is because the impact of the mnemonic measurement on boundary-updating is minimal whereas that of the informed class variable is substantial. After the above episode, the decision-maker's noisy mnemonic measurement m' is also likely to be slightly larger than 0 since m' is an unbiased random sample of the sensory measurement m (Equation 6). Thus, the impact of m' on boundary

updating is minimal because m' is close to 0 and thus only slightly attracts the class boundary. On the contrary, the impact of the informed state of the class variable CL on boundary updating is relatively substantial, pushing the class boundary towards the regime consistent with the informed state of CL (Equations 9-12), which is the smaller side. As a result, the class boundary is negatively (towards-$small$-side) biased, which leads to the negative bias in the PSE of the psychometric curve defined from the trials following the episode of interest (as depicted by the left (yellow) regime in the plot of Fig. 3D).

Next, to appreciate the stimulus-dependent nature of feedback effects in the world-updating scenario, let's consider another case where the decision-maker experiences a PDM episode with an unambiguous stimulus: the ring with size 2 is presented and produces a sensory measurement m that falls around 2, which leads to the $large$ choice and then followed by $correct$ feedback. After this episode, as in the previous case with an ambiguous stimulus, the informed state of the class variable ($CL = large$) shifts the class boundary to the smaller side. However, unlike the previous case, the impact of the mnemonic measurement m' on boundary-updating, which is likely to be around 2, is substantial, resulting in a shift of the boundary towards the far larger side. Consequently, the class boundary becomes positively (towards-$large$-side) biased. Here, the mnemonic measurement and the informed state of the class variable exert conflicting influences on boundary updating. Since the mnemonic measurement increases as the stimulus size grows (e.g., $S = 0 \rightarrow 1 \rightarrow 2$), the relative impact of the mnemonic measurement on boundary-updating is increasingly greater as the stimulus size grows, eventually overcoming the counteracting influence of the informed state of the class variable (S1 Fig). As a result, the bias in the class boundary is initially negative but is progressively reversed to be positive as the stimulus size grows, which leads to the bias reversal in the PSE of the psychometric curve defined from the trials following the episode of interest (as depicted by the right (blue) regime in the plot of Fig 3D).

We stress that this 'stimulus-dependent bias reversal' is a hallmark of the world-updating scenario's prediction of the history effects in PDM. Specifically, the direction of bias reversal is always from $small$ to $large$ as long as the feedback in conjunction with the choice indicates $CL = small$ (e.g., $\{S = 0 \rightarrow -1 \rightarrow -2, C = small, F = correct\}$ or $\{S = 0 \rightarrow -1 \rightarrow -2, C = large, F = incorrect\}$) and always from $large$ to $small$ as long as the feedback in conjunction with the choice indicates $CL = large$ (e.g., $\{S = 0 \rightarrow 1 \rightarrow 2, C = large, F = correct\}$ or $\{S = 0 \rightarrow 1 \rightarrow 2, C = small, F = incorrect\}$). Critically, the value-updating scenario does not predict the bias reversal (Fig 3E, right). It predicts that the feedback effects only asymptotically decrease as a function of sensory evidence but never switch to the other direction. This is because the decision confidence, $p(CL = large(small))$, only modulates the amount of value-updating but never changes the direction of value-updating.

Ex ante simulation of the feedback effects under the two scenarios

Above, we have conceptually explained why and how the two scenarios imply the distinct patterns of stimulus-dependent feedback effects. Though this implication seems intuitively apparent, it must be confirmed under the experimental setting of the current study. Moreover, there are good reasons to expect any history effect to exhibit complex dynamics over trials. First, sensory and mnemonic measurements are subject to stochastic noises, which propagates through decision-making and value/boundary-updating processes to subsequent trials (e.g., a sensory measurement that happens to fall on a relatively $small$ side is likely to lead to a $small$ choice, which affects the subsequent value/boundary-updating process, and so on). Second, provided that any deterministic value/boundary-updating processes are presumed to be at work, the PDM episode on a given trial must, in principle, be probabilistically conditioned on the episodes in past trials (e.g., the current $small$ choice on the ring of $S = 0$ is likely to have

followed the previous episodes leading to 'boundary-updating in the $large$ direction' or 'positive value-updating of the $small$ choice'). Third, two steps of deterministic value/boundary-updating occur between what can be observed at $toi - 1$ and at $toi + 1$ (as indicated by the psychometric curves in Fig 4A), once following the episode at $toi - 1$ (U_{toi-1} in Fig 4A) and next following the episode at toi (U_{toi} in Fig 4A). Thus, the differences between the retrospective and prospective history effects should be construed as reflecting not only U_{toi} but also U_{toi-1}. The nuanced impacts of this hidden updating on the history effects must be complicated, and thus be inspected with realistic simulations. Further, considering that these multiple stochastic and deterministic events interplay to create diverse temporal contexts, history effects are supposed to reveal themselves in multiplexed dynamics.

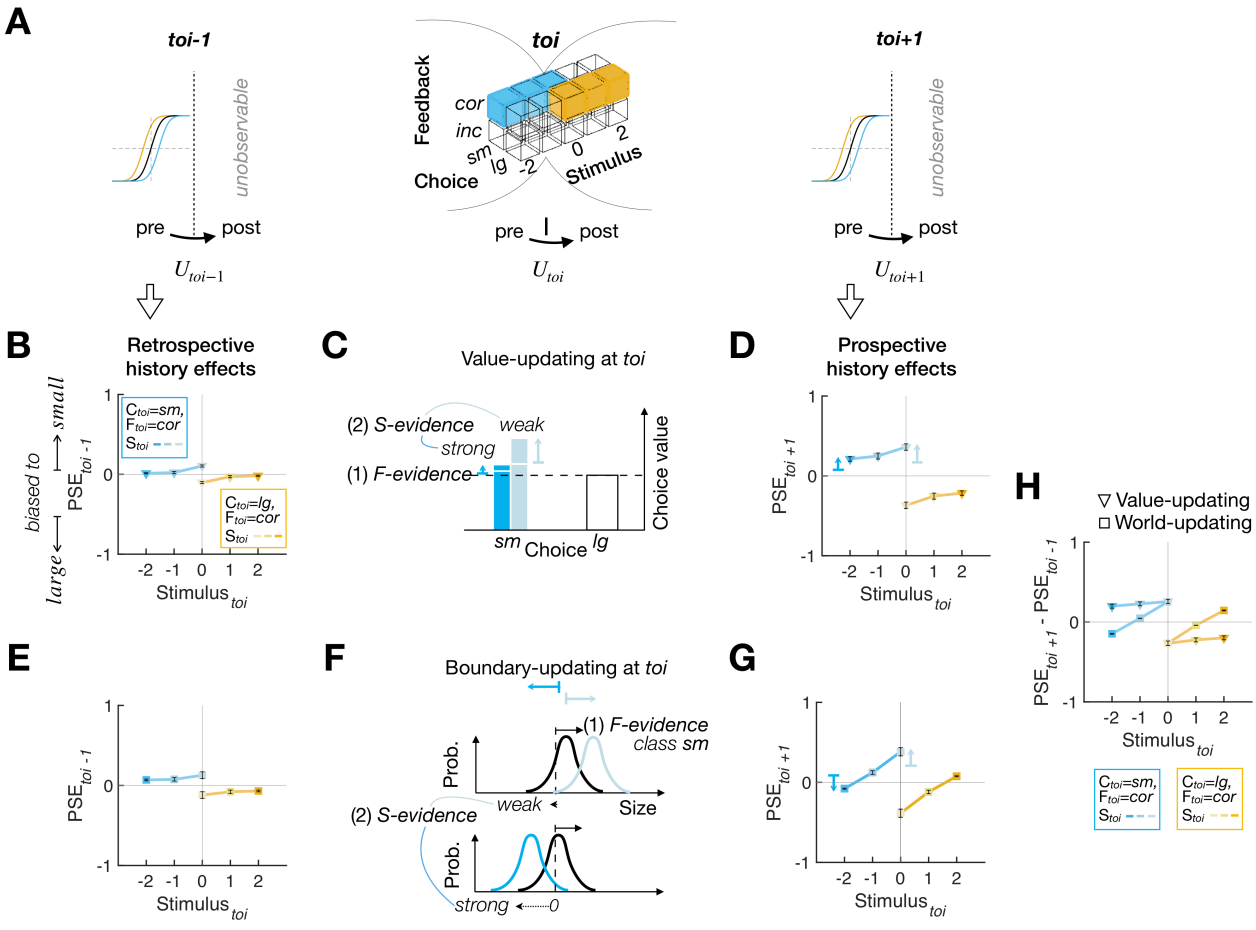


Fig 4. *Ex ante* simulation results for the PDM episodes with $correct$ feedback. (A) Illustration of how the retrospective (left) and prospective (right) history effects relate to the value updates and boundary

updates (bottom) occurring over the trials overarching the trial of interest. While the updating occurs latently at every trial (as indicated by U_{toi-1}, U_{toi}, U_{toi+1}), its behavioral consequences are observable only at the pre-updating phase at *toi-1* and *toi+1*. **(B-D)** The observable retrospective **(B)** and prospective **(D)** history effects and latent value-updating processes **(C)** for the value-updating model agent. **(C)** Since $correct$ feedback is treated as a positive reward, the chosen value is updated positively while the amount of value-updating varies depending on the strength of sensory evidence, as indicated by the length of the vertical arrows in different colors (weak sensory evidence, pale blue; strong sensory evidence, dark blue). The short horizontal bars and arrow heads of the colored arrows indicate the chosen values before and after U_{toi}, respectively. **(E-G)** The observable retrospective **(E)** and prospective **(G)** history effects and latent boundary-updating processes **(F)** for the world-updating model agent. **(F)** Since $correct$ feedback is treated as a logical indicator of the true state of the class variable (i.e., the true inequality between the class boundary and the stimulus), the class boundary shifts as a joint function of feedback and sensory evidence, where the boundary shift due to sensory evidence (solid black arrows) counteracts that due to feedback (dotted black arrows), as indicated by the arrows in different colors (weak sensory evidence, pale blue; strong sensory evidence, dark blue). The short vertical bars and arrow heads of the colored arrows at the top indicate the class boundary before and after U_{toi}, respectively. **(H)** Juxtaposition of the differences between the retrospective and prospective history effects displayed by the two model agents. **(C, F)** The contributions of both sensory and feedback evidence are indicated by *S-evidence* and *F-evidence*, respectively. **(B, D, E, G)** Data points are the means and s.e.m.s across the parameter sets used in *ex ante* simulations (see Materials and methods). The data underlying this figure **(B, D, E, G, H)** can be found in S1 Data.

Hence, we simulated *ex ante* the two models over a reasonable range of parameters by making the model agents perform the binary classification task on the sequences of stimuli that will be used in the actual experiment (Table A in S1 Appendix, S4 Fig, and Materials and methods). The simulation results confirmed our intuition, as summarized in Fig 4, which shows the retrospective and prospective history effects for the PDM episodes with $correct$ feedback. Notably, the retrospective history effects indicate that both value-updating and world-updating agents were already slightly biased to the choice they are about to make in the—following—*toi* (Fig 4B and 4E). One readily intuits that such retrospective biases are more pronounced when conditioned on the *toi* with weak sensory evidence because the stochastic bias consistent with the choice that would be made in the *toi* is required more in those trials. This testifies to the presence of the complex dynamics of history effects discussed above and is also consistent with what has been previously observed (e.g., see Figure 2 of the previous study [9]). Importantly, in line with our conceptual conjecture (Fig 3D and 3E), the two agents evidently disagree on the prospective history effects. While the value-updating agent always exhibits the feedback-congruent bias but never reverses the direction of bias, the world-updating agent

shows the feedback-congruent bias after viewing the ambiguous stimulus but progressively reversed the direction of bias as the stimulus evidence supporting the decision becomes stronger (Fig 4C, 4D, and 4F-4H).

Next, Fig 5 summarizes the history effects for the PDM episodes with $incorrect$ feedback. The retrospective history effects show that both agents exhibit the choice bias consistent with the choice they will make next trial, as in the case for $correct$ feedback, but the amounts of bias are much greater compared to those in the $correct$-feedback condition (Fig 5B and 5E). These pronounced retrospective effects conditioned on the $incorrect$-feedback episodes are intuitively understood as follows: the value-updating agent's value ratio or the world-updating agent's class boundary was likely to be somehow "unusually and strongly" biased before the *toi,* given that they make an $incorrect$—thus "unusual"—choice in the *toi*. Supporting this intuition, the retrospective bias increases as sensory evidence increases, since the prior value ratio or class boundary must be strongly biased to result in that particular $incorrect$ choice despite such strong sensory evidence. Importantly, despite these large retrospective biases, the prospective history effects indicate that both agents adjust their value and class boundary, respectively, in their own manners identical to those for the $correct$-feedback episodes (Fig 5C, 5D, 5F, and 5G). Thus, as in the case of the $correct$-feedback episodes, the direction reversal is displayed only by the world-updating agent, but not by the value-updating agent (Fig 5H).

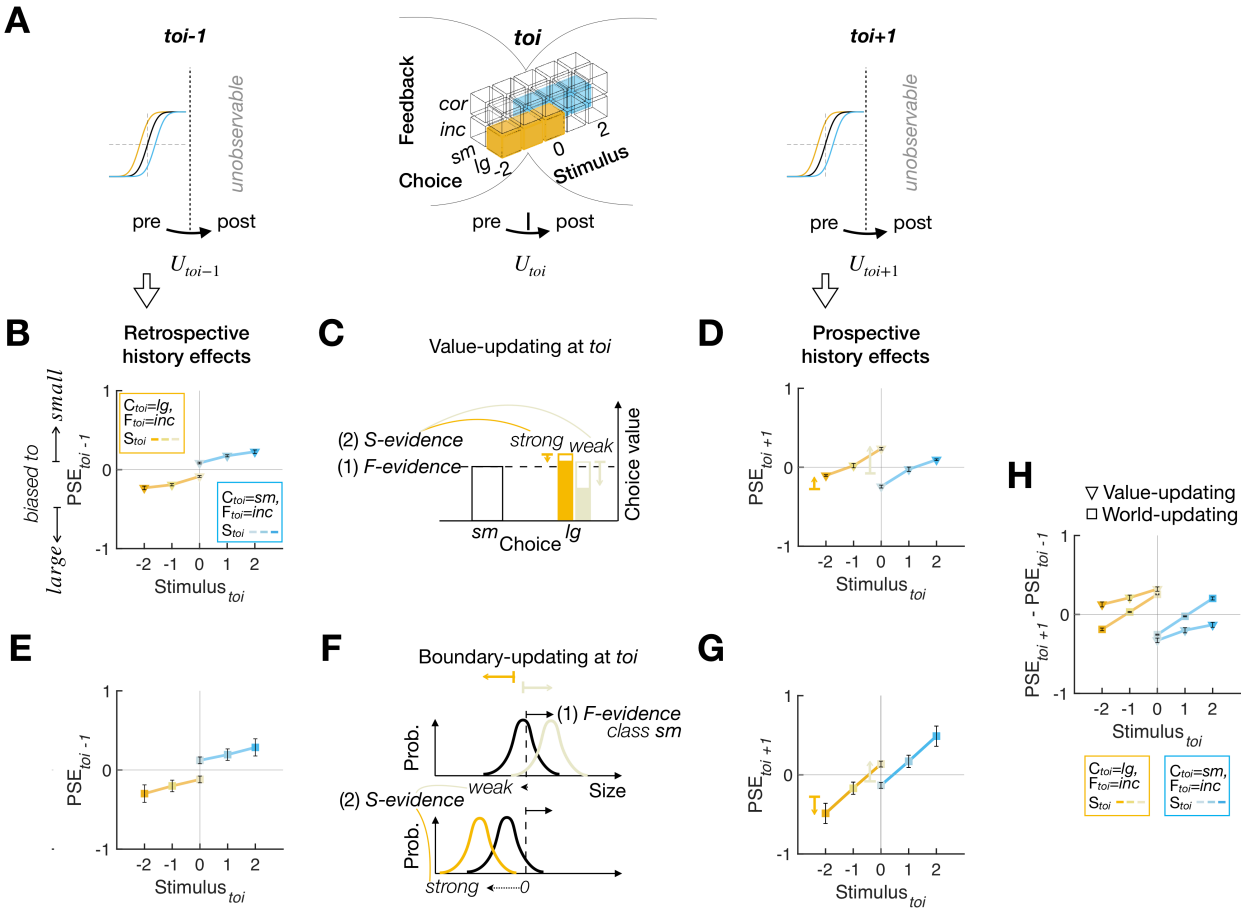


Fig 5. *Ex ante* simulation results for the PDM episodes with $incorrect$ feedback. The format is identical to that in Fig 4. The data underlying this figure **(B, D, E, G, H)** can be found in S1 Data.

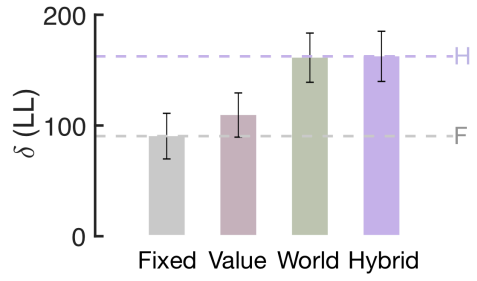
In sum, the *ex ante* simulation confirmed that the bias reversal of the stimulus-dependent feedback effects occurs only under the world-updating scenario but not under the value-updating scenario, regardless of the ($correct$ or $incorrect$) states of feedback. The simulation results also confirmed that, with the current experimental setting, we can empirically determine which of the two scenarios provides a better account of feedback effects.

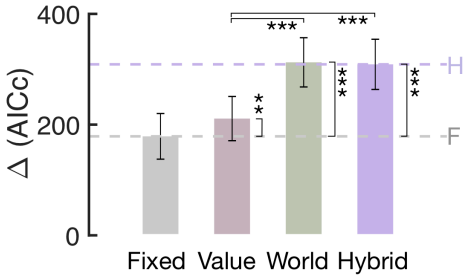
Evaluating the two scenarios for the goodness of fit to human decision-making data

Having confirmed the distinct predictions of the two scenarios via *ex ante* simulation, we evaluated their goodness of fit to human data. As points of reference for evaluation in the model space (Fig 6A), we created three reference models. The 'Base' model sets the class boundary at the unbiased value ($B = 0$) and does not update any choice values, thus incorporating neither arbitrary choice preference nor adaptive updating. The 'Fixed' model is identical to the Base model except that it incorporates arbitrary choice preference by fitting the constant class boundary to the data. The 'Hybrid' model incorporated both value-updating and world-updating algorithms. We quantified the models' ability to predict human classification choices using log likelihood (Fig 6B) and compared their abilities using the Akaike information criterion corrected for sample size (AICc [32]; Fig 6C)).

A

Model	intrinsic choice bias?	adaptive updating ?	params. #
Base	No	No	1
Fixed	Yes	No	2
Value	Yes	*value*	5
World	Yes	*boundary*	6
Hybrid	Yes	*value & boundary*	9





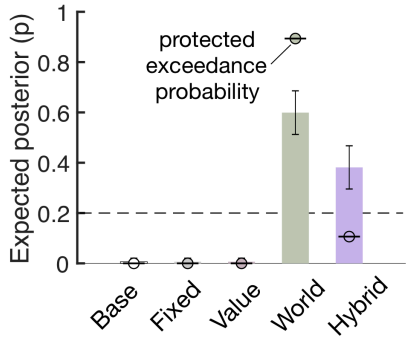


Fig 6. Model goodness of fit to human choice behavior. (A) Specification of the models constituting the model space. The color labels also apply to the rest of the panels in **(B-D)**. **(B-C)** Model comparisons in goodness of fit in terms of log likelihood **(B)** and AICc **(C).** The height of bars represents the across-

participant average differences from the goodness-of-fit measures of the Base model (N=30, mean±s.e.m.). Both difference measures indicate a better fit for higher values. Dashed lines in purple (Hybrid model) and gray (Fixed model) provide the reference points for evaluating the value-updating and world-updating models' accountability of the trial-to-trial choice variability (see Main text for their exact meanings). Pair-wise model comparisons were performed using paired one-tailed *t*-tests (asterisks indicate significance: *, $P<0.05$; **, $P<0.005$; ***, $P<10^{-8}$) **(D)** Model comparisons in the hierarchical Bayesian Model Selection measures. Height of bars, expected posterior probabilities; error bars, standard deviation of posterior probabilities. Dots marked with short dashes, protected exceedance probability. Dashed lines, chance level (p = 0.2), indicating the probability that a model is favored over others in describing the data by random chance. Bayesian omnibus risk (BOR), the estimated probability that observed differences in model frequencies may be due to chance, is reported (BOR = 1.7636×10^{-10}). The data underlying this figure **(B, C, D)** can be found in S1 Data.

The Fixed model's performance relative to the Base model's (gray dashed lines in Fig 6B, 6C) reflects the fraction of choice variability that is attributed to arbitrary choice preference. On the other hand, the Hybrid model's performance relative to the Base model's (purple dashed lines in Fig 6B, 6C) reflects the maximum fraction of choice variability that can be potentially explained by either the value-updating model, the world-updating model, or both. Thus, the difference in performance between the Hybrid and Fixed models (the space spanned between the gray and purple dashed lines in Fig 6B, 6C) quantifies the meaningful fraction of choice variability that the two competing models of interest are expected to capture. Prior to model evaluation, we confirmed that the two competing models (the value-updating and world-updating models) and two reference models (the Base and Hybrid models) are empirically distinguishable by carrying out a model recovery test (S3 Fig).

With this target fraction of choice variability to be explained, we evaluated the two competing models by comparing them against the Fixed and Hybrid models' performances while taking into account model complexity with AICc. The value-updating model was moderately better than the Fixed model (paired one-tailed *t*-test, $t(29) = -2.8540$, $P = 0.0039$) and substantially worse than the Hybrid model (paired one-tailed *t*-test, $t(29) = 7.6996$, $P = 8.6170\times10^{-9}$) and the world-updating model (paired one-tailed *t*-test, $t(29) = 8.3201$, $P = 1.7943\times10^{-9}$). By contrast, the world-updating model was substantially better than the Fixed model (paired one-tailed *t*-test, $t(29) = -10.3069$, $P = 1.6547\times10^{-11}$) but not significantly better

than the Hybrid model (paired one-tailed *t*-test, $t(29) = -1.0742$, $P = 0.1458$). These results indicate (i) that the world-updating model is better than the value-updating model in accounting for the choice variability and (ii) that adding the value-updating algorithm to the world-updating algorithm does not improve the accountability of the choice variability.

To complement the above pair-wise comparisons, we took the hierarchical Bayesian Model Selection approach [33–35] using AICc model evidence, to assess how probable it is that each of the five models prevails in the population (expected posterior probability; vertical bars in Fig 6D) and how likely it is that any given model is more frequent than the other models (protected exceedance probability; dots with horizontal bars in Fig 6D). Both measures corroborated the outcomes of the pair-wise comparisons: the world-updating model predominated in expected posterior probability (0.5992) and protected exceedance probability (0.8938).

In sum, the world-updating scenario was superior to the value-updating scenario in predicting the choice behavior of human participants performing the binary classification task.

Ex post simulation of the feedback effects under the two scenarios

The goodness-of-fit results summarized above simply indicate that the world-updating model is better than the value-updating model in predicting the trial-to-trial variability in choice behavior while taking into account model complexity. Our study aims to examine whether these two competing models of interest can account for the stimulus-dependent feedback effects observed in human decision-makers. To do so, we carried out *ex post* simulations based on the goodness-of-fit results [36] by testing whether the value-updating and world-updating models can reproduce the observed stimulus-dependent feedback effects.

The *ex post* simulation was identical to the *ex ante* simulation except that each decision-maker's best-fit model parameters were used (Table B in S1 Appendix; Materials and methods). We assessed how well the models reproduce the human history effects of feedback in two different ways. First, we compared the models and the humans similarly to the *ex ante* simulation (Fig 7A-7C). We included the PDM episodes with non-veridical feedback (symbols with dotted lines in Fig 7A-7C) though those episodes infrequently occurred (12.09±0.02% (mean±s.e.m.) out of total *toi* episode trials; bars with dotted outlines in Fig 7D). As a result, we inspected the retrospective and prospective history effects, and their differences, for all the possible combinations of ‘stimulus,’ ‘choice,’ and ‘feedback’ (20 PDM episodes in total), which resulted in a total of 60 PSE pairs to compare. The PSEs simulated by the world-update model closely matched the human PSEs, in both pattern and magnitude (Fig 7A and 7C), whereas those by the value-update model substantively deviated from the human PSEs (Fig 7A and 7B). The statistical comparison (paired two-tailed *t*-tests with Bonferroni correction) indicates that the value-updating model’s PSEs significantly deviated from the corresponding human PSEs for almost half of the entire pairs (29 out of 60 pairs) whereas none of the world-updating model’s PSEs significantly differed from the human PSEs (0 out of 60 pairs). Notably, most mismatches occurred because the value-updating model does not reverse the direction of feedback effects as sensory evidence becomes stronger while humans do so (compare the third columns of Fig 7A and 7B).

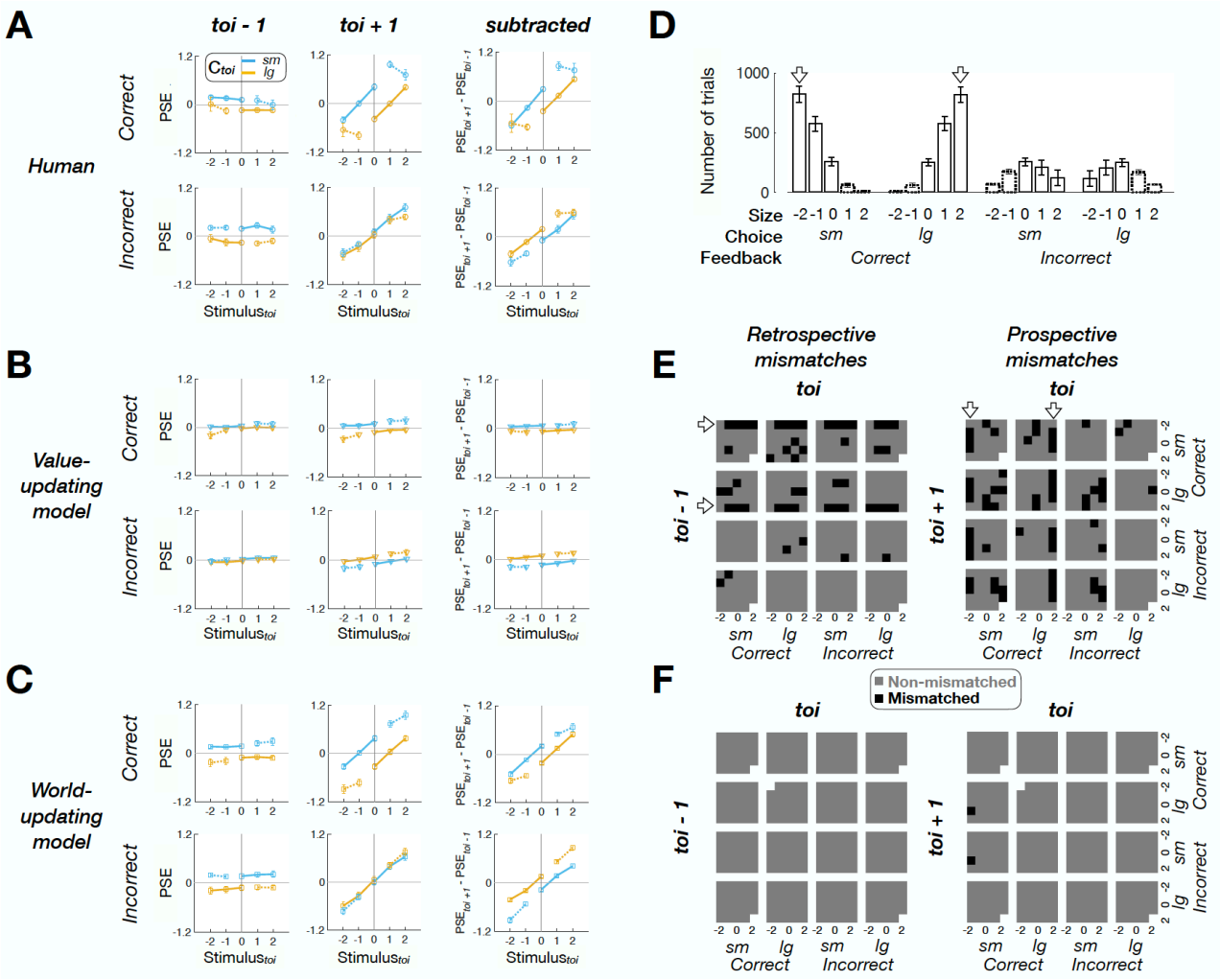


Fig 7. *Ex post* simulation results. (A-C) Retrospective (left columns), prospective (middle columns), and subtractive (right columns) history effects in PSE for the human **(A)**, value-updating **(B)**, and world-updating **(C)** decision-makers. Top and bottom rows in each panel show the PSEs associated with the *toi* episodes involving $correct$ and $incorrect$ feedback. Symbols with error bars, mean±s.e.m. across 30 decision-makers. See S5 Fig for the results from the Hybrid model decision-makers. **(D)** Frequency of PDM episodes in the human data (mean and SD across participants). **(E, F)** Maps of significant deviations of the value-updating **(E)** and world-updating (**F**) model agents from the human decision-makers in the retrospective (left) and prospective (right) history effects. Gray and black cells of the maps mark the insignificant and significant deviations (paired two-tailed *t*-tests with the Bonferroni correction for multiple comparisons). Empty cells are data points with NaN values due to insufficient trials. The data underlying this figure **(A, B, C, D, E, F)** can be found in S1 Data.

Second, we compared the models and the humans in the probability distribution of retrospective and prospective episodes conditioned on each episode of *toi* (Fig 7D-7F). This comparison allows us to assess the models' reproducibility not just for feedback effects but also for the history effects in general and to explore the origin of the value-based model's failure. By

collapsing all the preceding and following trials onto each of the 20 *toi* episodes (the columns of Fig 7E and 7F) and computing their probability distributions across—again—the 20 *toi-1* and 20 *toi+1* episodes (the rows of Fig 7E and 7F), respectively, we could create 400 joint-probability cells.

We carried out repeated *t*-tests with Bonferroni correction to see where the model-human mismatches occur (data were missing for a few cells—mostly those including non-veridical-feedback episodes, as indicated by the empty cells in Fig 7E and 7F, because those episodes were too rare (Fig 7D) to occur for all participants). For the remaining cells, the world-updating model showed a remarkable level of correspondence with the humans, deviating from the humans at only two cells (out of 790 cells, 0.25%; Fig 7F). By contrast, the value-updating model failed to match the humans for 94 cells (out of 792 cells, 11.87%; Fig 7E). Here, the mismatches occurred systematically: they were frequent when the preceding episode defining any given cell (i.e., episodes at *toi-1* for the retrospective cells or episodes at *toi* for the prospective cells) was featured with strong sensory evidence (as indicated by the arrows in Fig 7E). This systematic deviation precisely reflects the incapability of the value-updating model to reverse the direction of feedback effects as sensory evidence strengthens.

In sum, the stimulus-dependent history effects of feedback observed in humans could be reproduced by the world-updating scenario but not by the value-based scenario.

Discussion

Here, we explored the two possible scenarios for what humans learn from corrective feedback in a PDM task. We implemented the value-updating scenario with the belief-based RL model

545 [9,10], originally developed to account for the stimulus-dependent effects of reward feedback on
546 animals' PDM. As an alternative, we implemented the world-updating scenario with BMBU,
547 where decision-makers continuously update their internal knowledge about stimulus distribution
548 based on sensory measurements and corrective feedback. The latter excels over the former in
549 predicting the choice behavior and reproducing the stimulus-dependent feedback effects in
550 human PDM, suggesting that humans update their knowledge about world statistics upon
551 corrective feedback for PDM.

552 Given RL models' success in VDM and the presence of physical rewards, it is not
553 surprising for the belief-based RL model to be considered as an account of the feedback effects
554 in animals' PDM. The original work [9] supported this model using six datasets, including one
555 human dataset [37]. However, the current work indicates that the way humans learn from
556 corrective feedback—without any physical or monetary reward—in PDM deviates from the
557 value-updating scenario. The critical deviation occurred for the PDM episodes with strong
558 sensory evidence: past $correct$ feedback should, albeit weakly, reinforce the choice made in the
559 past according to the value-updating scenario, whereas humans made the opposite choice
560 more frequently. In fact, the human dataset previously analyzed in the study [9] exhibits the
561 same deviations (see their Figure 8c,d). When this dataset was analyzed in our way, it
562 displayed the patterns almost identical to those of our dataset (S7A Fig). For that matter,
563 another published human dataset [31] substantially deviated from the value-updating scenario
564 (S7B Fig). We remain cautious about the possibility that even animals may demonstrate such
565 deviations as well. However, this possibility seems worth exploring though, given that the main
566 dataset from the 16 rats engaged in an olfactory PDM task also exhibited patterns similar to
567 those found in humans when corrected for the bias present in previous trials (see Figure 2i in
568 the study [9]). Notably, in these studies [9,31,37], the class boundary existed either implicitly
569 (e.g., a perfectly balanced odor mixture [9]) or explicitly (e.g., a reference stimulus presented in

another interval [37]). This suggests the possibility that the bias reversal of feedback effects may be a general phenomenon that can be observed in diverse types of binary classification tasks. However, further empirical tests are required to confirm this possibility. The bias reversal of feedback effects should not be treated lightly as a nuisance because any variant of the RL algorithm cannot reverse the direction of reinforcement in principle, as demonstrated in our work and in the modeling results of the same study [9] (shown in their Figure 3). By contrast, BMBU provides a principled account of these effects by treating $correct$ and $incorrect$ feedback as what they supposedly mean, a teaching signal indicating the true state of the class variable.

To be sure, the idea of shifting the decision or class boundary toward past stimuli *per se* is not new and has been previously hypothesized [38,39] or implemented into various models [40–44]. However, BMBU goes beyond these efforts by offering a normative formalism of incorporating $correct$ and $incorrect$ feedback as evidence for the class boundary such that it has an equal footing as sensory evidence in PDM tasks. This integration of feedback and sensory evidence within the framework of BDT advances the current computational account of the history effects because it addresses the history factors in the complete dimensions of PDM (‘stimulus’, ‘choice’, and ‘feedback’), which is important given the multiplexed nature of history effects emphasized by prior studies [8–11,31,45]. Our modeling work joins recent computational and empirical efforts of incorporating feedback in the normative evidence accumulation model [6,46], a framework commonly employed in various classic PDM tasks, such as a random-dot motion task. Furthermore, a study on rats' binary classification behavior has shown that rats can use information about the correct class state (referred to as "second-order prior" by the authors) by integrating their own choices with feedback (reward outcome) and that the population neural activity in the orbitofrontal cortex represents this information [11]. Together with these studies, our work supports a general view that decision-makers use corrective feedback as evidence for updating their world knowledge pertinent to the PDM task engaging them. Having mentioned the

general view on the role of feedback in human PDM, future efforts are needed to further verify the stimulus-dependent feedback effects under various sensory modalities and PDM tasks.

Previously, the so-called “Anna Karenina” account was presented to describe the seemingly idiosyncratic $incorrect$ feedback effects [9]. The Anna-Karenina account leaves the crucial aspect of feedback effects—the different consequences of $correct$ vs $incorrect$ feedback—unexplained. Since the belief-based RL model predicts the specific pattern of feedback effects for incorrect trials, as shown via *ex ante* simulation, endorsing the Anna-Karenina account admits that the belief-based RL model fails to account for the effects of $incorrect$ feedback observed in animals. For that matter, past studies on the history effects in PDM paid little attention to incorrect trials because they are, owing to their infrequency, considered too noisy and unreliable to be properly analyzed. By contrast, BMBU accounts for the effects of feedback in a principled way, regardless of whether the feedback is $correct$ or $incorrect$. Furthermore, BMBU explains why the feedback effects appear different between the correct and incorrect trials on the surface (compare the prospective history effects between Fig 4 and Fig 5): the correct and incorrect trials share the same deterministic boundary-updating process but had different histories of their own stochastic events, which led to correct versus incorrect choices, respectively.

As mentioned earlier, the history effects are dynamic and multiplexed in nature. This calls for an effort to establish a rigorous framework to probe behavioral data for the history effects. Several recent studies made such efforts by taking various approaches, yet all emphasizing the presence of distinct sources of biases. One study [47] assumed two sources with differing time scales and took a regression-based approach to separate their influences on choice bias by incorporating them as independent regressors to predict choices. Another group of researchers [6,9] also noted the presence of slow fluctuations and raised a concern about the conventional practice of inspecting only the prospective history effects because non-systematic

slow fluctuations in the decision-making strategy may cause the observed effects. This group dealt with this concern by subtracting the retrospective history effects from the prospective ones. A more recent study [48] shared this concern but disagreed about its remedy by showing that the subtraction method cannot fairly recover diverse systematic updating strategies. Alternatively, they took a model-based approach to separate any given updating strategy from random drifts in decision criteria. We acknowledge the importance of the efforts by these studies and share the same concern. But we emphasize that BMBU successfully reproduced human history effects in both directions of time without incorporating any non-systematic components arising from random drifts. BMBU's concurrent reproduction of the retrospective and prospective history effects was confirmed not just for the summary statistics (the PSEs in Fig 7C) but also for the individual data points spanning almost the entire space of PDM episode pairs (Fig 7F). This suggests that it is an empirical matter of whether the decision criterion slowly drifts or not, raising another concern that systematic history effects might be explained away as non-existing slow drifts. In this sense, we propose that researchers should treat the retrospective history effects not as a baseline or control condition but as what must be explained, the phenomenon equally important as the prospective history effects, before resorting to any non-systematic sources. We believe that such a treatment is the way historians treat historical events [49], and that our approach showcases its one rigorous example.

Materials and methods

Ethics statement

The study protocol was approved by the Seoul National University Institutional Review Board (No. 1310/001-020). All the experiments were conducted in accordance with the principles expressed in the Declaration of Helsinki. All subjects gave prior written informed consent to participate in the experiments.

Participants

All participants (13 females and 17 males, aged 18−30 years) were recruited from the Seoul National University (SNU) community and were compensated approximately $10/h.

Procedure

Stimuli. The stimulus was a thin (.07 deg in visual angle), Gaussian-noise filtered, black-and-white ring flickering at 20 Hz on a gray luminance background. On each trial, a fixation first appeared for 0.5 s on average (fixation duration uniformly jittered from 0.3 s to 0.7 s on a trial-to-trial basis) before the onset of a ring stimulus. Five different ring sizes (radii of 3.84, 3.92, 4.00, 4.08, 4.16 deg in visual angle (d.v.a.), denoted by −2, −1, 0, 1, 2, respectively, in the main text) were randomized within every block of 5 trials.

Task. Participants performed a binary classification task on ring size with trial-to-trial corrective feedback. Each individual participated in 5 daily sessions, each consisting of 6 runs, each consisting of 170 trials, ended up performing a total of 5,100 trials. In any given trial, participants viewed one of the five rings and indicated its class ($small$ or $large$) within 1.2 s after stimulus onset by pressing one of the two keys using their index and middle fingers. The assignment of

computer keys for $small$ and $large$ choices alternated between successive sessions to prevent any unwanted choice bias possibly associated with finger preference. The response period was followed by a feedback period of 0.5 s, during which the color of the fixation mark informed the participants of whether their response was correct (green) or not (red). In case no response had been made within the response period, the fixation mark turned yellow, reminding participants that a response must be made in time. These late-response trials comprised 0.5418% of the entire trials across participants and were included in data analysis. Meanwhile, the trials on which a response was not made at all comprised 0.0948% of the entire trials. These trials were excluded from analysis and model fitting. As a result, the number of valid trials per participant ranged from 5,073 to 5,100 with an average of 5,095.2 trials. Before each run, we showed participants the ring stimulus of the median size (4.00 d.v.a. in radius) on the screen for 15 s while instructing them to use that ring as a reference for future trials, i.e., to judge whether a test ring is smaller or larger than this reference ring. This procedure was introduced for the purpose of minimizing any possible carryovers from the belief they formed about the class boundary in the previous session. Participants were encouraged to maximize the fraction of correct trials.

Feedback Manipulation. We provided participants with stochastic feedback using a 'virtual' criterion sampled from a normal distribution $N(\mu_{True}, \sigma_{True})$. σ_{True} was always fixed at 1.28 throughout the entire runs. In each run, μ_{True} was initially (up to 40−50 trials) set to 0 and then to one of the three values ($\mu_{True} = \{-0.4, 0, 0.4\}$) with the equal proportion (10 runs for each value) for the rest of trials. The stochastic feedback was introduced this particular way to create PDM episodes with (occasional) non-veridical feedback while mimicking a real-world situation where references are slightly noisy and biased in an unnoticeable manner.

Data analysis

For any given PDM episode at a trial of interest (*toi*), we quantified the retrospective and prospective history effects by probing the psychometric curves at the trials before and after *toi*, respectively. The psychometric function ($\psi(x)$) was estimated by fitting the cumulative Gaussian distribution (F) to the curves using *Psignifit* package [50–52] (https://github.com/wichmann-lab/psignifit), as follows:

$$\psi(x;\mu,\sigma) = F(x;\mu,\sigma),$$

where μ and σ are the mean and standard deviation of F. By finding the best-fitting value of μ, we defined the point of subjective equality (PSE; the stimulus level with equal probability for a $small$ or $large$ choice), which was used as the summary statistics that quantifies the history effects associated with a given PDM episode. To ensure reliable PSE estimates, we acquired bootstrap samples (N=5,000) of psychometric curves based on the binomial random process and took their average as the final estimate for each PDM episode. In our main data analysis, the results of which are displayed in Fig 7, we chose not to include the parameters for guess or lapse rates in estimating PSEs. This was done to prevent unfair overfitting problems from occurring in infrequent episode types with small numbers of trials available for fitting. On the other hand, to preclude any potential confounding problem related to the task difficulty associated with PDM episode types, we also repeated the above PSE estimation procedure with guess (γ) and lapse (λ) rates included as free parameters: $\psi(x;\mu,\sigma,\gamma,\lambda) = \gamma + (1-\gamma-\lambda)F(x;\mu,\sigma)$. The results did not differ between the original estimation procedure without the lapse and guess rates and the procedure with the lapse and guess rates (Bonferroni-corrected P = 0.2023 ~ 1.000; paired two-tailed *t*-tests; see S2 Data for detailed statistical information).

Value-updating model

As a model of the value-updating scenario, we used the belief-based RL model proposed in the previous work [9,10]. This model incorporates RL algorithm into the conventional Bayesian

formalism of decision confidence—also known as statistical decision confidence using a partially observable Markov decision process (Fig 3E). In this model, the decision-maker, given sensory measurement m, computes the probability that the stimulus belongs to '$large$' (p_L) or '$small$' ($p_S = 1 - p_L$) class (hereinafter the p-computation), where $p_L = \int_{\mu_0}^{\infty} p(S|m)dS$. This probability will be referred to as a 'belief-state', as in the original work [9,10]. Here, the probability distribution $p(S|m)$ is defined as a normal distribution with mean m and standard deviation σ_m. Whereas μ_0 was assumed to be zero in the original work, we set μ_0 free as a constant parameter to allow the belief-based RL model to deal with any potential individuals' idiosyncratic choice bias, as we will allow the world-updating model (BMBU) to do so (see below). Next, the expected values of the two choices Q_S and Q_L can be obtained by p_S and p_L multiplied with the learned values of the options of $small$ and $large$, V_S and V_L, respectively. Accordingly, the expected value Q_C is also defined separately for the choice made between $small$ and $large$: Q_S and Q_L.

In the original work, the argmax rule was applied to determine the choice (i.e., the higher Q determines the choice C). Instead, here we applied the softmax rule, which selects $large$ with probability $\frac{\exp(\beta Q_L)}{\exp(\beta Q_S) + \exp(\beta Q_L)}$ (the higher Q preferentially selects C) where β is an inverse temperature. This feature did not exist in the original model but was introduced here to allow the belief-based RL model to deal with stochastic noise at the decision stage, as we allow the world-updating model (BMBU) to do so.

The initial values of $small$ and $large$ choices were set identically as a free parameter V_{init}. Upon receiving feedback on the decision, the decision-maker updates the value of the selected choice V_C by the reward prediction error δ with learning rate α:

$$V_C \leftarrow V_C + \alpha\delta. \quad (1)$$

No temporal discounting is assumed for simplicity. Since the decision-maker treats corrective feedback as rewards ($correct$: $r = +1$, $incorrect$: $r = 0$), the reward prediction error δ is computed as the deviation of the reward from the expected value:

$$\delta = r - Q_C = r - p_C V_C. \tag{2}$$

Note that the belief state p_C (i.e., statistical decision confidence) modulates δ such that δ increases as p_C decreases, which is the crucial relationship constraining the belief-based RL model's key prediction on the stimulus-dependent feedback effects. Specifically, upon $correct$ feedback, δ will take a positive value and reinforce the choice value. However, as p_C increases, the magnitude of such reinforcement will decrease. Critically, despite the decrease of reinforcement as a function of p_C, the sign of reinforcement will never be reversed until the expected value Q reaches the maximum reward value ($r = 1$). Based on the same ground, the sign of reinforcement will never be reversed either in the case of $incorrect$ feedback. The free parameters of the value-updating model are $\theta = \{\mu_0, \sigma_m, \alpha, \beta, V_{init}\}$.

World-updating model

As a model of the world-updating scenario, we developed the Bayesian model of boundary-updating (BMBU). BMBU shares the same platform for PDM with the belief-based RL model (as depicted in Figs 1A and 3A) but, as a BDT model, makes decisions using its "learned" generative model while continually updating its belief about the class boundary B, the key latent variable of that internal model (as depicted in the left panel of Fig 3D).

"Learned" generative model. In BDT, the learned generative model refers to the decision-maker's subjective internal model that relates task-relevant variables (m, m', and B in the left panel of Fig 3D) to external stimuli and behavioral choices (S and CL, respectively, in the left panel of Fig 3D). As previously known [53,54], the decision-maker's internal model is likely to deviate from the "actual" generative model that accurately reflects how the experimenter

759 generated external stimuli due to one's limitations in the sensory and memory apparatus. In the
760 current experimental setup, we assumed that the internal model of the decision-maker deviates
761 from that of the experimenter in the following aspect: due to the noise in the sensory and
762 memory encoding processes, the decision-maker is likely to believe that many rings of different
763 sizes are presented, although the experimenter used only five discrete-size rings. The post-
764 experiment interviews supported this: none of the participants reported perceiving discrete
765 stimuli during the experiment. A deviation like this is known to occur commonly in
766 psychophysical experiments where a discrete number of stimuli were used [40,54,55].

767 We incorporated the above deviation into the decision-maker's internal model by
768 assuming that the stimulus at any given trial is randomly sampled from a Gaussian distribution
769 with mean B and variance σ_S^2 (as depicted by $B \rightarrow S$ in Fig 3D):

770 $$p(S|B) = N(S; B, \sigma_S^2), \tag{3}$$

771 which defines the probability distribution of stimuli conditioned on the class boundary, where σ_S^2
772 corresponds to the extent to which a given decision-maker assumes that stimuli are distributed.
773 Next, the inequality between the class boundary and the stimulus determines the state of the
774 class CL (as depicted by the converging causal relations involving the class variable, $B \rightarrow CL \leftarrow$
775 S, in Fig 3D):

776 $$CL = large\ (small)\ if\ S > (<)\ B, \tag{4}$$

777 which defines the correct answer of the perceptual task. On the other hand, the sensory
778 measurement m at any given trial is randomly sampled from a Gaussian distribution with mean
779 S and variance σ_m^2 (as depicted by $S \rightarrow m$ in Fig 3D):

780 $$p(m|S) = N(m; S, \sigma_m^2), \tag{5}$$

781 which defines the probability distribution of sensory measurements conditioned on the stimulus,
782 where σ_m^2 corresponds to the extent to which the decision-maker's sensory system is noisy.

Lastly, the mnemonic measurement m' at any given trial is randomly sampled from a Gaussian distribution with mean m and variance $\sigma_{m'}^2$ (as depicted by $m \rightarrow m'$ in Fig 3D):

$$p(m'|m) = N\left(m'; m, \sigma_{m'}^2\right), \tag{6}$$

which defines the probability distribution of mnemonic measurements conditioned on the sensory measurement, where $\sigma_{m'}^2$ corresponds to the extent to which the decision-maker's working memory system is noisy. This generative process ($m \rightarrow m'$) is required because the sensory evidence of the stimulus is no longer available in the sensory system—due to a brief (0.3 sec; Fig 2B) stimulus duration—at the moment of updating the state of the class boundary (as will be shown below in the subsection titled "Boundary-updating") and instead must be retrieved from the working memory system. The mnemonic recall of the stimulus is known to be noisy, becoming quickly deteriorated right away after stimulus offset, especially for continuous visual evidence such as color and orientation [56,57]. The generative process relating m to m' has been adopted for the same reason by recent studies [58,59], including our group [55], and is consistent with the non-zero levels of memory noise in the model-fit results ($\sigma_{m'}^2$ = [1.567, 5.606]). The substantial across-individual variability of the fitted levels of $\sigma_{m'}^2$ is also consistent with the previous studies [55,58,59].

With the learned generative model defined above, the decision-maker commits to a decision by inferring the current state of the class variable CL from the current sensory measurement m and then updates the current state of the boundary variable from both the current mnemonic measurement m' and the current feedback F.

Decision-making. As for decision-making, BMBU, unlike the belief-based RL model, does not consider the choice values but completely relies on the p-computation by selecting the $large$ class if $p_L > 0.5$ and the $small$ class if $p_L < 0.5$. The p-computation is carried out by propagating the sensory measurement m within its learned generative model:

$$p_L = \int_{\hat{B}}^{\infty} p(S|m)dS, \tag{7}$$

where the finite limit of the integral is defined by the inferred state of the boundary $\hat{B}$, which is continually updated on a trial-to-trial basis (as will be described below). This means that the behavioral choice can vary depending on $\hat{B}$ even for the same value of m (as depicted in the 'perception' stage of Fig 3A and Fig 3B).

Boundary-updating. After having experienced a PDM episode in any given trial t, BMBU (i) computes the likelihood of the class boundary by concurrently propagating the mnemonic measurement m_t' and the "informed" state of the class variable CL_t, which can be informed by feedback F_t and choice C_t in the current PDM episode, within its learned generative model ($p(m_t', CL_t|B_t)$) and then (ii) forms a posterior distribution of the class boundary ($p(B_t|m_t', CL_t)$) by combining that likelihood with its prior belief about the class boundary at the moment ($p(B_t)$), which is inherited from the posterior distribution formed in the previous trial $t-1$ ($p(B_{t-1}|m_{t-1}', CL_{t-1})$). Intuitively put, as BMBU undergoes successive trials, its posterior belief in the previous trial becomes the prior in the current trial, being used as the class boundary for decision-making and then being combined with the likelihood to be updated as the posterior belief in the current trial. Below, we will first describe the computations for (i) and then those for (ii). As explained above (Equation 6), we stress that the likelihood computation must be based not on the sensory measurement m_t but on the mnemonic measurement m_t' because m_t is no longer available at the moment of boundary-updating.

As for the boundary likelihood computation (i), BMBU posits that the decision-maker infers how likely the current PDM episode—i.e., the combination of the mnemonic measurement m_t', the choice C_t, and the corrective feedback F_t—is generated by hypothetical values of the class boundary ($p(m_t', C_t, F_t|B_t)$). Since the "true" state of the class variable CL_t is deduced from any given pair of C_t and F_t states in binary classification as follows,

$$CL_t = large\ if\ C_t = large\ and\ F_t = correct\ or\ if\ C_t = small\ and\ F_t = incorrect;$$

$$CL_t = small\ otherwise,$$

the likelihood can be defined using only m'_t and CL_t: $p(m'_t, C_t, F_t|B_t) \equiv p(m'_t, CL_t|B_t)$. Hence, the likelihood of the class boundary is computed by propagating m'_t and CL_t inversely over the learned generative model (as defined by Equations 3-6):

$$p(m'_t, CL_t|B_t) = \int p(m'_t, CL_t, S_t|B_t)dS_t = \int p(m'_t|S)p(CL_t|S_t, B_t)p(S_t|B_t)dS_t, \tag{8}$$

which entails the marginalization over every possible state of S_t, a variable unknown to the decision-maker. Here, since the binary states of CL_t ($CL_t \in \{small, large\}$) indicates the inequality between S_t and B_t (Equation 4), B_t is used as the finite limit of the integrals to decompose the original integral into the one marginalized over the range of S_t satisfying $CL_t = small$ and the other marginalized over the range of S_t satisfying $CL_t = large$:

$$\int p(m'_t|S)p(CL_t|S_t, B_t)p(S_t|B_t)dS_t =$$

$$= \int_{-\infty}^{B_t} p(m'_t|S_t)p(CL_t|S_t, B_t)p(S_t|B_t)dS_t + \int_{B_t}^{+\infty} p(m'_t|S_t)p(CL_t|S_t, B_t)p(S_t|B_t)dS_t, \tag{9}$$

Note that the boundary likelihood function is computed based on CL_t informed by feedback. The right-hand side of Equation 9 can further be simplified for the informed state CL_t by replacing the infinite limits with finite values (Equation S5 in Text in S1 Appendix). For the case of $CL_t = large$, $p(CL_t|S_t, B_t)$ in the left and right integral terms on the right-hand side of Equation 9 becomes 0 and 1, respectively, while becoming 1 and 0 for the case of $CL_t = small$ in the ranges of S_t of the corresponding integrals (Equation S3-S6 in Text in S1 Appendix). Hence, we find the likelihood of the class boundary in a reduced form, separately for $CL_t = large$ and $CL_t = small$, as follows:

$$p(m'_t, CL_t = small|B_t) = \int_{-\infty}^{B_t} p(m'_t|S_t)p(S_t|B_t)dS_t;$$
$$p(m'_t, CL_t = large|B_t) = \int_{B_t}^{+\infty} p(m'_t|S_t)p(S_t|B_t)dS_t \tag{10}$$

where $p(m_t'|S_t) = N(m_t'; S_t, \sigma_{m'}^2 + \sigma_m^2)$, according to the "chain" relations defined in the learned generative model ($S \rightarrow m \rightarrow m'$ in the left panel of Fig 3D; Equation 5-6; see Equation S2 for derivations in Text in S1 Appendix). Equation 10 indicates that BMBU calculates how likely hypothetical boundary states bring about the mnemonic measurement ($B \rightarrow S \rightarrow m \rightarrow m'$) while taking into account the informed state of the class variable ($B \rightarrow CL \leftarrow S$), by constraining the possible range of the stimulus states. To help readers intuitively appreciate these respective contributions of the mnemonic measurement and the informed state of the class variable (feedback) to the boundary likelihood, we further elaborated on how Equation 9 is reduced to Equation 10 depending on the informed state of CL_t (see Text in SI Appendix and S1 Fig).

Lastly, we evaluate the integral for $CL_t = small$ in Equation 10 by substituting $p(S_t|B_t) = N(S_t; B_t, \sigma_S^2)$ and $p(m_t'|S_t) = N(m_t'; S_t, \sigma_{m'}^2 + \sigma_m^2)$, from the defined statistical knowledge in the learned generative model (Equation 3 and Equation 5-6, respectively) and find:

$$p(m_t', CL_t = small|B_t) = \frac{1}{\sqrt{2\pi\left(\frac{\sigma_M^2\sigma_S^2}{(\sigma_M^2+\sigma_S^2)}\right)}} \int_{-\infty}^{B_t} e^{-\frac{\left(S_t - \frac{B_t\sigma_M^2 + m_t'\sigma_S^2}{\sigma_M^2+\sigma_S^2}\right)^2}{2\frac{\sigma_M^2\sigma_S^2}{(\sigma_M^2+\sigma_S^2)}}} dS_t \times \frac{1}{\sqrt{2\pi(\sigma_M^2+\sigma_S^2)}} e^{-\frac{(B_t - m_t')^2}{2(\sigma_M^2+\sigma_S^2)}}. \tag{11}$$

where $\sigma_M^2 = \sigma_{m'}^2 + \sigma_m^2$. For the other state in feedback, we evaluate the integral in the same manner and find:

$$p(m_t', CL_t = large|B_t) = \frac{1}{\sqrt{2\pi\left(\frac{\sigma_M^2\sigma_S^2}{(\sigma_M^2+\sigma_S^2)}\right)}} \int_{B_t}^{\infty} e^{-\frac{\left(S_t - \frac{B_t\sigma_M^2 + m_t'\sigma_S^2}{\sigma_M^2+\sigma_S^2}\right)^2}{2\frac{\sigma_M^2\sigma_S^2}{(\sigma_M^2+\sigma_S^2)}}} dS_t \times \frac{1}{\sqrt{2\pi(\sigma_M^2+\sigma_S^2)}} e^{-\frac{(B_t - m_t')^2}{2(\sigma_M^2+\sigma_S^2)}}. \tag{12}$$

Having calculated the likelihood of B_t, we turn to describe (ii) how BMBU combines that likelihood with a prior distribution on trial t, which forms a posterior distribution of B_t according to Bayes rule:

$$p(B_t|m'_t, CL_t) \propto p(m'_t, CL_t|B_t)p(B_t). \tag{13}$$

We assumed that, at the beginning of the current trial t, the decision-maker recalls the posterior belief $p(B_{t-1}|m'_{t-1}, CL_{t-1})$ formed (Equation 13) from the previous trial—to use it as the prior of B_t—into the current working memory space, and it is thus subject both to decay λ and diffusive noise $\sigma_{diffusion}$ during the recall process. As a result, the prior $p(B_t)$ is basically the recalled posterior, defined as the normal distribution $N(\hat{B}_t, \sigma_{B_t}^2)$ as follows:

$$\hat{B}_t = \lambda \hat{B}_{t-1}^{post} + (1-\lambda)\mu_0;$$

$$\sigma_{B_t}^2 = \lambda \sigma^{2}{}_{t-1}^{post} + \sigma_{diffusion}^2, \tag{14}$$

where $\hat{B}_{t-1}^{post}$ and $\sigma^{2}{}_{t-1}^{post}$ denote mean and variance of the previous trial's posterior distribution.

Note that the decay parameter $\lambda = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^{2}{}_{t-1}^{post}}$ influences the width and location of the belief distribution, and that the diffusive noise of $\sigma_{diffusion} > 0$ helps to keep the width of the distribution over multiple trials, thus avoiding sharpening and stopping the updating process [60]. In this way, λ and $\sigma_{diffusion}$ allows BMBU to address the idiosyncratic choice bias and noise, as we allow the belief-based RL model to so with μ_0 and the sofmax rule.

In sum, BMBU posits that human individuals carry out a sequence of binary classification trials with their learned generative model while continually updating their belief about the location of the class boundary in that generative model. BMBU describes these decision-making and boundary-updating processes using a total of 6 parameter ($\theta = \{\mu_0, \sigma_m, \sigma_s, \sigma_0, \sigma_{m'}, \sigma_{diffusion}\}$), which are set free to account for individual differences.

Reference models

As the references for evaluating the belief-based RL model and BMBU in predicting the variability of human choices, we created three reference models. The 'Base' model captures the choice variability that can be explained by the p-computation with the class boundary fixed at 0 unanimously for all participants and without any value-updating process. Thus, it has only a single free parameter representing the variability of the sensory measurement ($\theta = \{\sigma_m\}$). The 'Fixed' model captures the choice variability that can be explained by the p-computation with the class boundary set free to a fixed constant μ_0 for each participant and without any value-updating process. Thus, it has two free parameters ($\theta = \{\mu_0, \sigma_m\}$). The 'Hybrid' model captures the choice variability that can be explained both by the p-computation with the inferred class boundary by BMBU and by the value-updating process implemented by the belief-based RL model. Thus, it has nine free parameters ($\theta = \{\mu_0, \sigma_m, \sigma_s, \sigma_0, \sigma_{m'}, \sigma_{diffusion}, \alpha, \beta, V_{init}\}$). In Fig 6B-6D, the differential goodness-of-fit measures on the y-axis indicate the subtractions of the performance of the 'Base' model from those of the remaining models.

Model fitting

For each participant, we fitted the models to human choices over N valid trials (N≤170) of M(=10) experimental runs under K(=3) conditions, where invalid trials were the trials in which the participants did not make any response. For any given model, we denote the log likelihood of a set of parameters θ given the data as follows:

$$LL(\theta, model) = \log p(data|\,\theta, model) = \sum_{k=1}^{K_{cond}} \sum_{j=1}^{M_{runs}} \sum_{i=1}^{N_{trials}} \log p\left(C_{i,j,k} \middle|\, \theta, model\right),$$

where $C_{i,j,k}$ denotes the participant's choice ($large$ or $small$) on the i-th trial of the j-th run under the j-th condition. Computation of this LL is analytically intractable given the stochastic nature of choice determination. So, we used inverse binomial sampling (IBS [61]), an efficient way of generating unbiased estimates via numerical simulations. The maximum-likelihood estimate of

the model parameters was obtained with Bayesian Adaptive Direct Search (BADS) [62], a hybrid Bayesian optimization to find the parameter vector θ^* that maximizes the log likelihood, which works well with stochastic target functions. To reduce the risk of being stuck at local optima, we repeated 20 independent fittings by setting the starting positions randomly using Latin hypercube sampling (*lhsdesign_modifed.m* by Nassim Khlaled, https://www.mathworks.com/matlabcentral/fileexchange/45793-latin-hypercube), then picked the fitting with the highest log likelihood. To avoid infinite loops from using IBS, we did not impose individual lapse rates in an arbitrary manner. Instead, we calculated the average of the lapse rate and guess rate from the cumulative Gaussian fit to a given individual's grand mean (based on the entire trials) psychometric curve. With these individual lapse probabilities (mean rate of 0.05, which ranged [0.0051, 0.1714]), trials were randomly designated as lapse trials, in which the choice was randomly determined to be either $small$ or $large$.

Model comparison in goodness-of-fit

We compared the goodness-of-fit of the models using corrected Akaike information criterion (AICc) based on maximum-likelihood estimation fitting, as follows:

$$AICc = -2 \cdot \mathrm{LL}(\theta^*) + 2p + \frac{2p(p+1)}{(N \text{ x } M \text{ x } K) - p - 1},$$

where p is the number of parameters of the model and the total number of trials in the dataset is $N \times M \times K$. Log model evidence was obtained for each participant by multiplying AICc by $-1/2$ [35]. Furthermore, we took a hierarchical Bayesian Model Selection approach that infers the posterior over model frequencies in the population based on log model evidence values in each subject. To conclude whether a given model is the most likely model above and beyond chance, we also reported protected exceedance probabilities for each model (see Fig 6E and 6F). The

random effects model selection at the group level relied on the function *VBA_groupBMC.m* of the Variational Bayesian Analysis toolbox (https://mbb-team.github.io/VBA-toolbox/) [63].

Model recovery analysis

We performed a model recovery analysis to further validate our model fitting pipeline. In the analysis, we considered the two competing models of interest (the world-updating and value-updating models) and the two reference models (the Base and Hybrid models). Using the same parameter set, we generated synthetic data for each participant's true stimulus sequences. For the realistic synthetic data, the parameter values were chosen based on the best-fitting parameter estimates from each individual. We generated 30 sets of synthetic data for each model, with 153,000 trials in each set. We then fit all four models to each synthetic dataset, resulting in 480 fitting problems. We assessed the models using the AICc-based log model evidence and computed exceedance probabilities. Our analysis showed that all models were distinguishable, which confirms the validity of our model fitting pipeline (S3 Fig).

Ex ante and *Ex post* model simulations

We conducted *ex ante* model simulations to confirm and preview the value-updating and world-updating models' distinct predictions on the stimulus-dependent feedback effects under the current experimental setting. Model simulations were conducted using trial sequences (i.e., stimulus order and correct answers) identical to those administered to human participants. The model parameters used in the *ex ante* simulation are summarized in the Table A in S1 Appendix. Note that the 25 levels (uniformly-spaced [0.15, 3.27]) of σ_m, the only parameter common to the two models, were used. As for the other parameters specific to each model, we selected the values that generated human-level task performances (see S4 Fig for details and statistical results). Simulations were repeated 100 times, resulting in the $100 \times N \times M \times K =$

507,300~510,000 trials per participant. For simplicity, we assumed neither lapse trials nor any arbitrary choice bias.

The procedure of *ex post* model simulations was identical to that of *ex ante* model simulations except that the best-fitting model parameters and lapse trials were used.

Statistical tests

Unless otherwise mentioned, the statistical comparisons were performed using paired *t*-tests (two-tailed, N=30). To test the reversed feedback effects under conditions of strong sensory evidence, we applied one-sample *t*-tests (one-tailed, N=27 for S7A Fig, N=8 for S7B Fig). Repeated *t*-tests on PSEs between data and model (Fig 7B-C, S5 Fig) were performed (two-tailed, N=30). In Table D in S1 Appendix, we reported the number of test conditions of significant deviation from the data (Bonferroni-corrected threshold; *: $P<0.00083$, **: $P<0.000167$, ***: $P<0.0000167$). Additionally, Wilcoxon signed-rank tests were performed with the same threshold applied (Table D in S1 Appendix). Repeated *t*-tests on each cell of episode frequency maps between the data and the models (Fig 7E and 7F; S6 Fig) were performed, and *P*-values were subjected to Bonferroni correction (two-tailed, N=30; value-updating, $P<0.0000631$; world-updating, $P<0.0000633$). Task performances between human agents (N=30) and model agents with different sets of parameters (N=25) were compared based on unpaired *t*-tests (two-tailed, S4 Fig).

Acknowledgments

The authors are grateful to Daeyeol Lee for his insightful comments and inspiring conversations concerning the prior version of the manuscript.

References

1. Gold JI, Law CT, Connolly P, Bennur S. The relative influences of priors and sensory evidence on an oculomotor decision variable during perceptual learning. J Neurophysiol. 2008;100(5):2653–68.

2. Hwang EJ, Dahlen JE, Mukundan M, Komiyama T. History-based action selection bias in posterior parietal cortex. Nat Commun. 2017;8(1):1242.

3. Abrahamyan A, Silva LL, Dakin SC, Carandini M, Gardner JL. Adaptable history biases in human perceptual decisions. Proc National Acad Sci. 2016;113(25):E3548–57.

4. Busse L, Ayaz A, Dhruv NT, Katzner S, Saleem AB, Scholvinck ML, Zaharia AD, Carandini M. The detection of visual contrast in the behaving mouse. J Neurosci. 2011;31(31):11351–61.

5. Scott BB, Constantinople CM, Erlich JC, Tank DW, Brody CD. Sources of noise during accumulation of evidence in unrestrained and voluntarily head-restrained rats. eLife. 2015;4:e11308.

6. Mendonça AG, Drugowitsch J, Vicente MI, DeWitt EEJ, Pouget A, Mainen ZF. The impact of learning on perceptual decisions and its implication for speed-accuracy tradeoffs. Nat Commun. 2020;11(1):2757.

7. Fernberger SW. Interdependence of judgments within the series for the method of constant stimuli. J Exp Psychol. 1920;3(2):126–50.

8. Lak A, Nomoto K, Keramati M, Sakagami M, Kepecs A. Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. Curr Biol. 2017;27(6):821–32.

9. Lak A, Hueske E, Hirokawa J, Masset P, Ott T, Urai AE, Donner TH, Carandini M, Tonegawa S, Uchida N, Kepecs A. Reinforcement biases subsequent perceptual decisions when confidence is low, a widespread behavioral phenomenon. eLife. 2020;9:e49834.

10. Lak A, Okun M, Moss MM, Gurnani H, Farrell K, Wells MJ, Reddy CB, Kepecs A, Harris KD, Carandini M. Dopaminergic and prefrontal basis of learning from sensory confidence and reward value. Neuron. 2020;105(4):700-711.e6.

11. Nogueira R, Abolafia JM, Drugowitsch J, Balaguer-Ballester E, Sanchez-Vives MV, Moreno-Bote R. Lateral orbitofrontal cortex anticipates choices and integrates prior with current information. Nat Commun. 2017;8(1):14823.

12. Lee D, Seo H, Jung MW. Neural basis of reinforcement learning and decision making. Annu Rev Neurosci. 2012;35(1):287–308.

13. Daw ND, Doya K. The computational neurobiology of learning and reward. Curr Opin Neurobiol. 2006;16(2):199–204.

14. Sutton R, Barto A. Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press; 1998.

15. Corrado GS, Sugrue LP, Seung HS, Newsome WT. Linear-nonlinear-poisson models of primate choice dynamics. J Exp Anal Behav. 2005;84(3):581–617.

16. Lau B, Glimcher PW. Dynamic response-by-response models of matching behavior in rhesus monkeys. J Exp Anal Behav. 2005;84(3):555–79.

17. Gläscher J, Daw N, Dayan P, O'Doherty JP. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron. 2010;66(4):585–95.

18. Lee SW, Shimojo S, O'Doherty JP. Neural computations underlying arbitration between model-based and model-free learning. Neuron. 2014;81(3):687–99.

19. Odoemene O, Pisupati S, Nguyen H, Churchland AK. Visual evidence accumulation guides decision-making in unrestrained mice. J Neurosci. 2018;38(47):10143–55.

20. Hermoso-Mendizabal A, Hyafil A, Rueda-Orozco PE, Jaramillo S, Robbe D, Rocha J de la. Response outcomes gate the impact of expectations on perceptual decisions. Nat Commun. 2020;11(1):1057.

21. Akrami A, Kopec CD, Diamond ME, Brody CD. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. Nature. 2018;554(7692):368–72.

22. Summerfield C, Tsetsos K. Building bridges between perceptual and economic decision-making: neural and computational mechanisms. Front Neurosci. 2012;6:70.

23. Kahnt T, Grueschow M, Speck O, Haynes JD. Perceptual learning and decision-making in human medial frontal cortex. Neuron. 2011;70(3):549–59.

24. Polanía R, Krajbich I, Grueschow M, Ruff CC. Neural oscillations and synchronization differentially support evidence accumulation in perceptual and value-based decision making. Neuron. 2014;82(3):709–20.

25. Körding KP, Wolpert DM. Bayesian decision theory in sensorimotor control. Trends Cogn Sci. 2006;10(7):319–26.

26. Trommershäuser J, Maloney LT, Landy MS. Decision making, movement planning and statistical decision theory. Trends Cogn Sci. 2008;12(8):291–7.

27. Griffiths TL, Tenenbaum JB. Optimal predictions in everyday cognition. Psychol Sci. 2006;17(9):767–73.

28. Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND. How to grow a mind: statistics, structure, and abstraction. Science. 2011;331(6022):1279–85.

29. Glaze CM, Filipowicz ALS, Kable JW, Balasubramanian V, Gold JI. A bias–variance trade-off governs individual differences in on-line learning in an unpredictable environment. Nat Hum Behav. 2018;2(3):213–24.

30. Ma WJ. Bayesian decision models: A primer. Neuron. 2019;104(1):164–75.

31. Hachen I, Reinartz S, Brasselet R, Stroligo A, Diamond ME. Dynamics of history-dependent perceptual judgment. Nat Commun. 2021;12(1):6036.

32. Burnham KP, Anderson DR. Model Selection and Inference, A Practical Information-Heuristic Approach. New York: Springer-Verlag; 2002.

33. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. Neuroimage. 2009;46(4):1004–17.

34. Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group studies — Revisited. Neuroimage. 2014;84:971–85.

35. Cao Y, Summerfield C, Park H, Giordano BL, Kayser C. Causal inference in the multisensory brain. Neuron. 2019;102(5):1076-1087.e8.

36. Palminteri S, Wyart V, Koechlin E. The importance of falsification in computational cognitive modeling. Trends Cogn Sci. 2017;21(6):425–33.

37. Urai AE, Braun A, Donner TH. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. Nat Commun. 2017;8(1):14637.

38. Zariwala HA, Kepecs A, Uchida N, Hirokawa J, Mainen ZF. The limits of deliberation in a perceptual decision task. Neuron. 2013;78(2):339–51.

39. Renart A, Machens CK. Variability in neural activity and behavior. Curr Opin Neurobiol. 2014;25:211–20.

40. Qamar AT, Cotton RJ, George RG, Beck JM, Prezhdo E, Laudano A, Tolias AS, Ma WJ. Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. Proc National Acad Sci. 2013;110(50):20332–7.

41. Summerfield C, Behrens TE, Koechlin E. Perceptual classification in a rapidly changing environment. Neuron. 2011;71(4):725–36.

42. Norton EH, Fleming SM, Daw ND, Landy MS. Suboptimal criterion learning in static and dynamic environments. PLoS Comput Biol. 2017;13(1):e1005304.

43. Treisman M, Williams TC. A theory of criterion setting with an application to sequential dependencies. Psychol Rev. 1984;91(1):68–111.

44. Lages M, Treisman M. A criterion setting theory of discrimination learning that accounts for anisotropies and context effects. Seeing Perceiving. 2010;23(5):401–34.

45. Fritsche M, Mostert P, Lange FP de. Opposite effects of recent history on perception and decision. Curr Biol. 2017;27(4):590–5.

46. Drugowitsch J, Mendonça AG, Mainen ZF, Pouget A. Learning optimal decisions with confidence. Proc National Acad Sci. 2019;116(49):24872–80.

47. Mochol G, Kiani R, Moreno-Bote R. Prefrontal cortex represents heuristics that shape choice bias and its integration into future behavior. Curr Biol. 2021;31(6):1234-1244.e6.

48. Gupta D, Brody CD. Limitations of a proposed correction for slow drifts in decision criterion. arXiv preprint arXiv:220510912. 2022.

49. Carr EH. What is History? London: University of Cambridge & Penguin Books; 1961.

50. Wichmann FA, Hill NJ. The psychometric function: I. Fitting, sampling, and goodness of fit. Percept Psychophys. 2001;63(8):1293–313.

51. Wichmann FA, Hill NJ. The psychometric function: II. Bootstrap-based confidence intervals and sampling. Percept Psychophys. 2001;63(8):1314–29.

52. Schütt HH, Harmeling S, Macke JH, Wichmann FA. Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. Vision Res. 2016;122:105–23.

53. Ma WJ. Organizing probabilistic models of perception. Trends Cogn Sci. 2012;16(10):511–8.

54. Haefner RM, Berkes P, Fiser J. Perceptual decision-making as probabilistic inference by neural sampling. Neuron. 2016;90(3):649–60.

55. Lee H, Lee HJ, Choe KW, Lee SH. Neural evidence for boundary updating as the source of the repulsive bias in classification. J Neurosci. 2023;43(25):4664–83.

56. Wilken P, Ma WJ. A detection theory account of change detection. J Vis. 2004;4(12):11.

57. Bays PM, Gorgoraptis N, Wee N, Marshall L, Husain M. Temporal dynamics of encoding, storage, and reallocation of visual working memory. J Vis. 2011;11(10):6–6.

58. Luu L, Stocker AA. Post-decision biases reveal a self-consistency principle in perceptual inference. eLife. 2018;7:e33334.

59. Luu L, Stocker AA. Categorical judgments do not modify sensory representations in working memory. PLoS Comput Biol. 2021;17(6):e1008968.

60. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. Nature. 2006;441(7095):876–9.

61. Opheusden B van, Acerbi L, Ma WJ. Unbiased and efficient log-likelihood estimation with inverse binomial sampling. PLoS Comput Biol. 2020;16(12):e1008483.

62. Acerbi L, Ma WJ. Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. Advances in Neural Information Processing Systems. 2017;30:1836–46.

63. Daunizeau J, Adam V, Rigoux L. VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. PLoS Comput Biol. 2014;10(1):e1003441.

Supporting Information

S1 Fig. Schematic illustration of BMBU's account of how the joint contribution of the sensory and feedback evidence to boundary updating leads to the reversal of choice bias as a function of sensory evidence strength.

(A) Reversal of subsequent choice bias—expressed in PSE—as a function of sensory evidence strength and boundary inference—expressed in likelihood computation—based on a PDM episode. Left panel: The circles with different colors (indicated by (b-d), which points to the corresponding panels below **(B-D)**) represent the PSEs associated with the boundary updating for three example PDM episodes, where the stimulus (S_t) varies from 0 to 2 while the choice (C_t) and feedback (F_t) are $large$ and $correct$, respectively. Right panel: At the core of boundary inference is the computation of the likelihood of the class boundary based on the mnemonic measurement (m'_t) and the informed state of the class variable (CL_t), where CL_t is jointly determined by F_t and C_t (see **Materials and methods** for the full computation of boundary inference in BMBU). **(B-D)** The likelihoods of the class boundary given the three example PDM episodes defined in **(A)**, where sensory evidence varies from the low **(B)**, to the intermediate **(C)**, and to the high **(D)** level. To help understand why and how, given the same feedback evidence, the direction of boundary updating reverses as the sensory evidence strengthens, we visualize the boundary likelihoods as a product of two functions (Equation 12), indicated by sub-panels marked as (1) and (2). Top row: As indicated by (1), we plot each boundary likelihood when only the mnemonic measurement is considered, assuming that no feedback is provided. Note that these likelihood functions are centered around the values of m'_t, by attracting the class boundary toward themselves, driving a shift towards the $large$ side (i.e. positive side on the boundary axis). Middle-Bottom rows: When the feedback evidence is given—i.e., when the informed state of CL_t is revealed as $large$—in addition to the mnemonic measurement, an additional piece of information about the class boundary arises. As indicated by (1)×(2), we plot each boundary likelihood (defined in **(A)**). As indicated by (2), we plot each function (Middle row), as the result of (Bottom row) divided by (Top row). The complementary cumulative distribution functions shown here are also centered around m'_t because the $large$ state of CL_t means that the class boundary is located somewhere smaller than m'_t. Note that these skewed distributions push the inferred class boundary away from the state of CL_t informed by feedback, driving a shift towards the $small$ side (i.e. negative side on the boundary axis). Consequently, the influences from the sensory evidence and the feedback evidence counteract each other (Bottom row). Note that the likelihood functions are centered in the $small$ side when the sensory evidence is weak **(B)**, in the neutral side when intermediate **(C)**, and in the large side when strong **(D)**. These systematic shifts of the class-boundary likelihood as a function of the strength of sensory evidence predict that the PSE of the psychometric curve for the subsequent trial (*t+1*) reverses its sign from negative to positive as a function of the stimulus size, as shown in **(A)**.

(TIF)

S2 Fig. Example trial courses of estimated class boundary. (A) An example trial history to show how a temporal trajectory of the class boundary inferred by BMBU. For example, at trial #1 (x-axis), a physical stimulus (symbol x) was 0, a sensory measurement (symbol o) was a positive value when the boundary belief (solid black bar; y-axis) was centered at 0. BMBU's choice was $large$ (symbol square on the top of y-axis), and correct feedback (same square filled with green color) was provided, which indicates that the class variable at trial #1 CL_1 was $large$ (arrow's direction indicates the effect of the trial class variable on the subsequent boundary-updating). BMBU updates one's belief based on evidence from stimulus (colored symbol o) and feedback (CL_1), available at the time of boundary-updating. To illustrate cases where the bias reversal we defined in Fig 3D in the main text happen and do not happen, same examples were intentionally used as those we used in S1 Fig where we further detailed on the model's mechanisms. Depending on colors, sensory evidence is weak (yellow symbol o) or strong (purple symbol o), which leads to whether or not the reversal happens. Trial cases featured in a red box indicates that the "Reinforcement" principle is held (predicting subsequent choices to repeat $large$ choice) while those featured in a green box indicates that the "Reversal" happens (predicting subsequent choices to reverse the previously made $large$ choice). **(B)** Temporal trajectories of the class boundary when the same 6-trial sequence of physical stimuli in **(A)** was simulated for 100 times. This means different m and m' were realized. The data underlying this figure **(A, B)** can be found in S1 Data.

(TIF)

S3 Fig. Model recovery analysis. Each square represents exceedance probability p_{exc} from model recovery procedure. The 'ground-truth' model to simulate synthetic behavior was correctly recovered with p_{exc} >0.9 for all 4 models considered in the study. The light shade of the diagonal squares indicates that the ground-truth model was the best-fitting model, leading to a successful model recovery. Numerical values can also be found in S1 Data.

(TIF)

S4 Fig. Histograms of classification accuracies of the human participants and their model partners in the *ex ante* simulations.

(A, B) Across-individual distributions of the classification accuracy of the belief-based RL model **(A)** and BMBU **(B)** overlaid on those of the human participants. The models' choices were generated via *ex ante* simulations with a specific set of model parameters (Table A in S1 Appendix), the results of which are depicted in Fig 4 and Fig 5. The classification accuracy is measured by calculating the percentage of the trials in which the choice matched the feedback used in the actual experiment. The empty bars correspond to the histogram of human performances, the range of which is demarcated by the dashed vertical lines ([min, max]=[60.65%, 73.94%]). The average human classification accuracy was 67.85%. **(A)** Comparison of classification accuracy between the belief-based RL model's simulation (red color) and the human choices. The model's *ex ante* simulation accuracy was not different from the human accuracy ($t(53) = 1.4429$, $P = 0.1549$; Null hypothesis: model's performance vector and humans' performance

vector come from populations with equal means, unpaired two-tailed *t*-test). **(B)** Comparison of classification accuracy between BMBU's simulation (green color) and the human choices. The model's *ex ante* simulation accuracy was not different from the human accuracy ($t(53) = 0.9707$, $P = 0.3361$, unpaired two-tailed *t*-test). There was no significant difference in classification accuracy between the value-updating model and BMBU ($t(48) = 0.5733$, $P = 0.5691$, unpaired two-tailed *t*-test). The data underlying this figure **(A, B)** can be found in S1 Data.

(TIF)

S5 Fig. Retrospective (left columns), prospective (middle columns), and subtractive (right columns) history effects in PSE for the 'Hybrid' model's *ex post* model simulations.

Top and bottom rows in each panel show the PSEs associated with the *toi* episodes involving $correct$ and $incorrect$ feedback at *toi*. Symbols with error bars, mean±s.e.m. across the 30 model agents, which correspond to their 30 human partners. The colors of the symbols and lines label choices (blue: $small$ and yellow: $large$). The data underlying this figure can be found in S1 Data.

(TIF)

S6 Fig. Maps of frequency deviations of the value-updating (A) and world-updating (B) model agents' classifications in the *ex post* simulations from the human decision-makers in the retrospective (left) and prospective (right) history effects.

Each cell represents a pair of PDM episodes, as specified by the column and row labels. At each cell, the color represents how much the episode frequency observed in the model agents deviates from that observed in the corresponding human decision-makers. The results of statistical tests on these deviations are summarized in Fig 7E and 7F. The data underlying this figure **(A, B)** can be found in S1 Data.

(TIF)

S7 Fig. Retrospective (left columns), prospective (middle columns), and subtractive (right columns) history effects in PSE for the human classification performances of Urai et al. (2017)'s work [37] **(A) and Hachen et al. (2021)'s work** [31] **(B).**

(A, B) We downloaded both publicly available datasets, analyzed them in the same way that we analyzed human observers in our work, and plotted the results in the same format used for Fig 7A. Top and bottom rows in each panel show the PSEs associated with the *toi* episodes involving $correct$ and $incorrect$ feedback. Symbols with error bars, mean±s.e.m. across human observers. The colors of the symbols and lines label choices (blue: $small$ and yellow: $large$). The overall patterns of the PSEs plotted here appear similar to those plotted in Fig 7A, displaying the reversals in direction of stimulus-dependent feedback effects. When the same statistical tests used in our work were carried out, some of the data points at the stimuli with strong sensory evidence at *toi* significantly deviated from zero in the direction opposite to the feedback effect predicted by the value-updating scenario, as indicated by the asterisks. **(A)** Sequential

features of human observers (N=27) analyzed in our way from human dataset that once had been published [37], which is openly available (http://dx.doi.org/10.6084/m9.figshare.4300043), then analyzed in the previous study [9]. In this study, the participants performed a binary classification task on the difference in motion coherence by sorting the pairs of random-dot-kinematogram stimuli shown in two intervals (s1 and s2) into one of the two classes ('s1<s2' vs. 's1>s2') over consecutive trials. The presented stimuli were taken from 3 sets of difficulty levels (the difference between motion coherence of the test and the reference stimulus; easy: [2.5, 5, 10, 20, 30], medium: [1.25, 2.5, 5, 10, 30], hard: [0.625, 1.25, 2.5, 5, 20]). As done in the original study [9], we binned the trials into 8 levels by merging the trials of two neighboring coherence levels (e.g., the coherence levels of [0.625, 1.25]) into a single bin. Note that the coherence bins of [20, 35, 45, 48.75, 51.25, 55, 65, 80] (%s1) on the x-axis (50% represents the equal coherence between s1 and s2) are matched to the x-axis in Figure 8 of the previous study in which the same dataset had been used. Asterisks mark the significance of one-sample *t*-tests (uncorrected $P<0.05$, one-tailed in the direction of feedback effects) on the panel *toi+1* (stimulus 80%: $t(26) = 2.0138$, $P = 0.0272$) and on the panel *subtracted* (stimulus 20%: $t(26) = -3.1900$, $P = 0.0018$, stimulus 80%: $t(26) = 3.8810$, $P = 0.0003$). **(B)** Sequential features of human observers (N=8) published in another previous study [31]. We used the human dataset openly available as part of the repository (https://osf.io/hux4n). In this study, the participants performed a binary classification task on the speed of vibrotactile stimuli by classifying the speed of the presented vibration as 'low-speed (weak)' or 'high-speed (strong)'. Note that the nine-level stimuli of [−4,−3,−2,−1,0,1,2,3,4] on the x-axis followed how data were encoded by the original study [31]. Asterisks mark the significance of one-sample *t*-tests (uncorrected $P<0.05$, one-tailed in the direction of feedback effects) on the panel *toi+1* (stimulus −4: $t(7) = -3.6757$, $P = 0.004$, stimulus −3: $t(7) = -3.5252$, $P = 0.0048$, and stimulus −2: $t(7) = -2.0325$, $P = 0.04$) and on the panel *subtracted* (stimulus −4: $t(7) = -1.9848$, $P = 0.044$). The data underlying this figure **(A, B)** can be found in S1 Data.

(TIF)

S1 Appendix. Supporting details

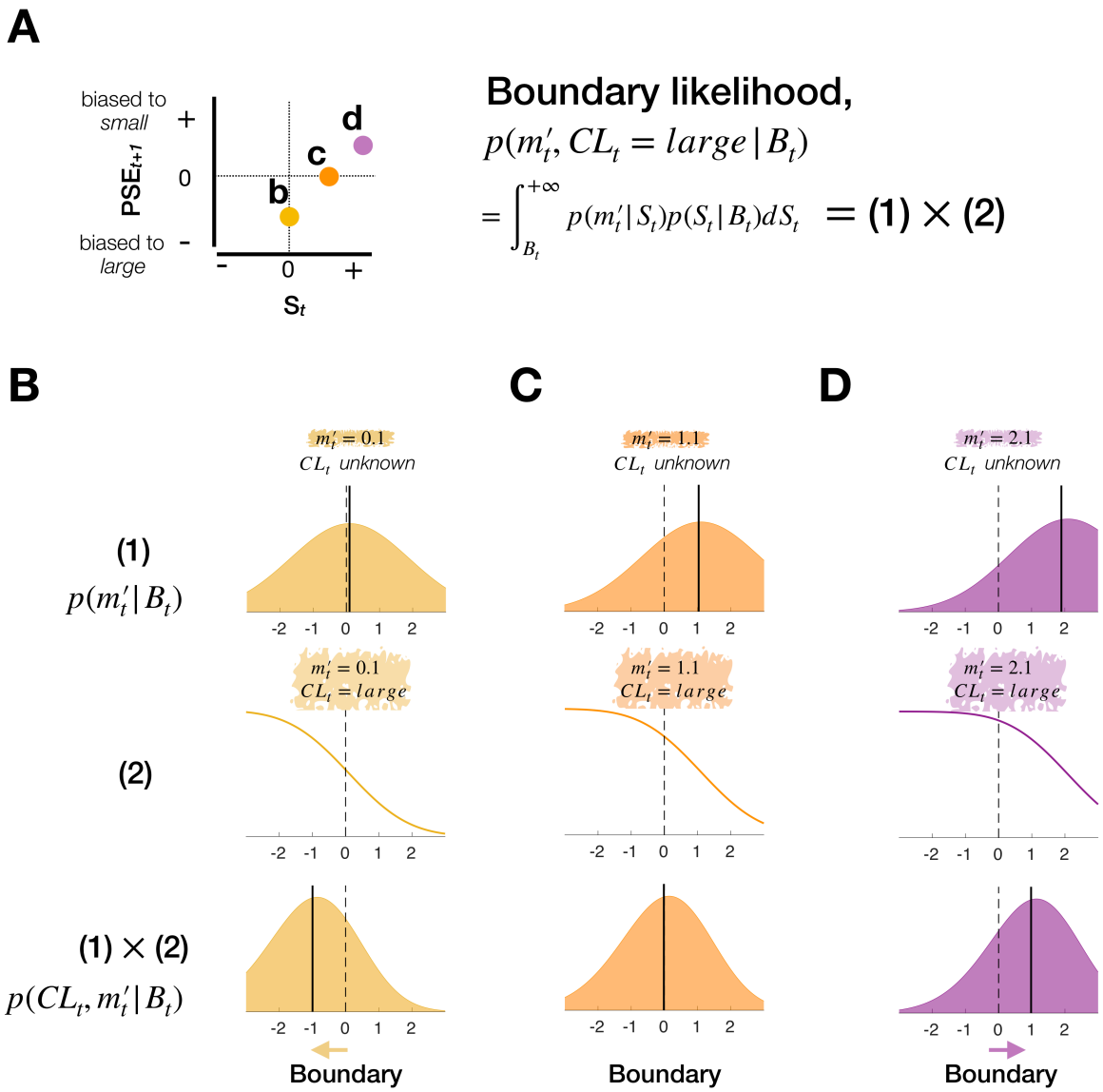
Supplemental details (Text) on additional model specifications of BMBU are provided. **Supplementary tables (A-D Tables) to support the Results section** are provided. **Table A. Parameters used for *ex ante* simulations. Table B. Parameters recovered from fitting the main models, world-updating and value-updating models, to human choices (**N=30**). Table C. Parameters recovered from fitting the rest of the models to human choices (**N=30**). Table D. Statistical results on model behavior versus human behavior in terms of PSE measures.**

(DOCX)

S1 Data. Excel spreadsheet containing, in separate sheets, the underlying numerical data for Figs 2D, 4B, 4D, 4E, 4G, 4H, 5B, 5D, 5E, 5G, 5H, 6B, 6C, 6D, 7A, 7B, 7C, 7D, 7E, 7F, S2A, S2B, S3, S4A, S4B, S5, S6A, S6B, S7A, and S7B.

(XLSX)

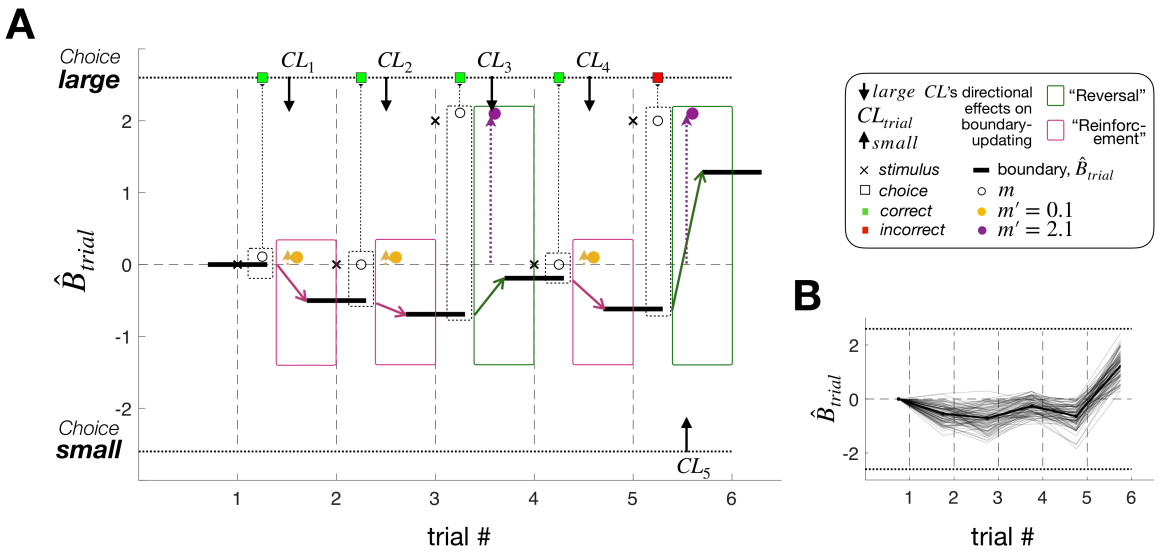
S2 Data. Excel spreadsheet containing detailed statistical information comparing alternative PSE estimation methods.
(XLSX)



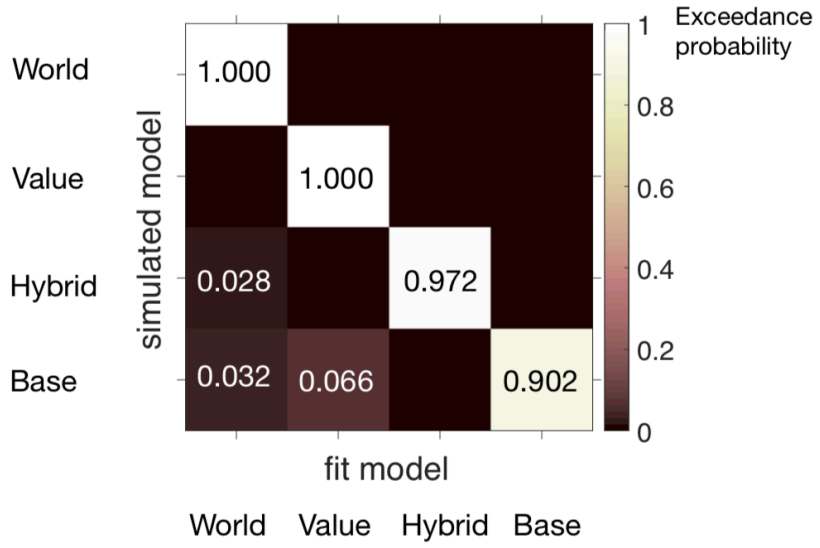
S1 Fig. Schematic illustration of BMBU's account of how the joint contribution of the sensory and feedback evidence to boundary updating leads to the reversal of choice bias as a function of sensory evidence strength.

(A) Reversal of subsequent choice bias—expressed in PSE—as a function of sensory evidence strength and boundary inference—expressed in likelihood computation—based on a PDM episode. Left panel: The circles with different colors (indicated by (b-d), which points to the corresponding panels below **(B-D)**) represent the PSEs associated with the boundary updating for three example PDM episodes, where the stimulus (S_t) varies from 0 to 2 while the choice (C_t) and feedback (F_t) are $large$ and $correct$, respectively. Right panel: At the core of boundary inference is the computation of the likelihood of the class boundary based on the mnemonic measurement (m'_t) and the informed state of the class variable (CL_t), where CL_t is jointly determined by F_t and C_t (see **Materials and methods** for the full computation of boundary inference in BMBU). **(B-D)** The likelihoods of the class boundary given the three example PDM

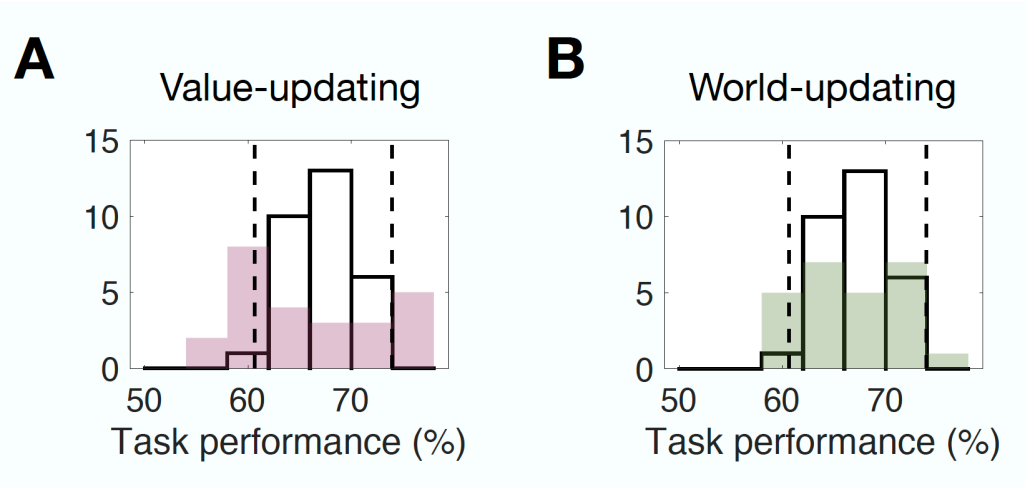
episodes defined in **(A)**, where sensory evidence varies from the low **(B)**, to the intermediate **(C)**, and to the high **(D)** level. To help understand why and how, given the same feedback evidence, the direction of boundary updating reverses as the sensory evidence strengthens, we visualize the boundary likelihoods as a product of two functions (Equation 12), indicated by sub-panels marked as (1) and (2). Top row: As indicated by (1), we plot each boundary likelihood when only the mnemonic measurement is considered, assuming that no feedback is provided. Note that these likelihood functions are centered around the values of m_t', by attracting the class boundary toward themselves, driving a shift towards the $large$ side (i.e. positive side on the boundary axis). Middle-Bottom rows: When the feedback evidence is given—i.e., when the informed state of CL_t is revealed as $large$—in addition to the mnemonic measurement, an additional piece of information about the class boundary arises. As indicated by (1)×(2), we plot each boundary likelihood (defined in **(A)**). As indicated by (2), we plot each function (Middle row), as the result of (Bottom row) divided by (Top row). The complementary cumulative distribution functions shown here are also centered around m_t' because the $large$ state of CL_t means that the class boundary is located somewhere smaller than m_t'. Note that these skewed distributions push the inferred class boundary away from the state of CL_t informed by feedback, driving a shift towards the $small$ side (i.e. negative side on the boundary axis). Consequently, the influences from the sensory evidence and the feedback evidence counteract each other (Bottom row). Note that the likelihood functions are centered in the $small$ side when the sensory evidence is weak **(B)**, in the neutral side when intermediate **(C)**, and in the large side when strong **(D)**. These systematic shifts of the class-boundary likelihood as a function of the strength of sensory evidence predict that the PSE of the psychometric curve for the subsequent trial (*t+1*) reverses its sign from negative to positive as a function of the stimulus size, as shown in **(A)**.



S2 Fig. Example trial courses of estimated class boundary. (A) An example trial history to show how a temporal trajectory of the class boundary inferred by BMBU. For example, at trial #1 (x-axis), a physical stimulus (symbol x) was 0, a sensory measurement (symbol o) was a positive value when the boundary belief (solid black bar; y-axis) was centered at 0. BMBU's choice was $large$ (symbol square on the top of y-axis), and correct feedback (same square filled with green color) was provided, which indicates that the class variable at trial #1 CL_1 was $large$ (arrow's direction indicates the effect of the trial class variable on the subsequent boundary-updating). BMBU updates one's belief based on evidence from stimulus (colored symbol o) and feedback (CL_1), available at the time of boundary-updating. To illustrate cases where the bias reversal we defined in Fig 3D in the main text happen and do not happen, same examples were intentionally used as those we used in S1 Fig where we further detailed on the model's mechanisms. Depending on colors, sensory evidence is weak (yellow symbol o) or strong (purple symbol o), which leads to whether or not the reversal happens. Trial cases featured in a red box indicates that the "Reinforcement" principle is held (predicting subsequent choices to repeat $large$ choice) while those featured in a green box indicates that the "Reversal" happens (predicting subsequent choices to reverse the previously made $large$ choice). **(B)** Temporal trajectories of the class boundary when the same 6-trial sequence of physical stimuli in **(A)** was simulated for 100 times. This means that different m and m' were realized. The data underlying this figure **(A, B)** can be found in S1 Data.

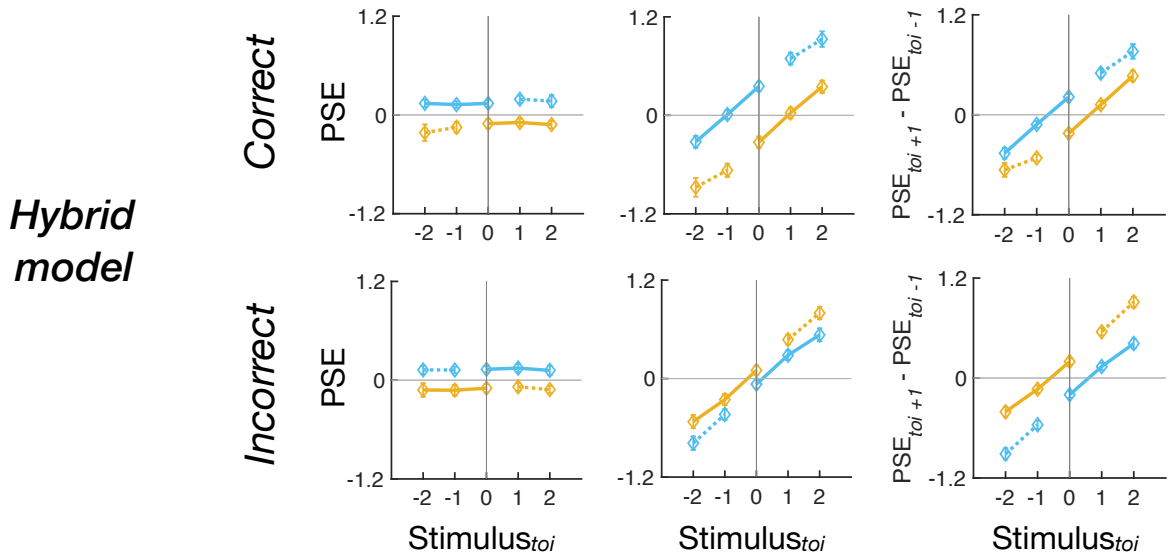


S3 Fig. Model recovery analysis. Each square represents exceedance probability p_{exc} from model recovery procedure. The 'ground-truth' model to simulate synthetic behavior was correctly recovered with p_{exc} >0.9 for all 4 models considered in the study. The light shade of the diagonal squares indicates that the ground-truth model was the best-fitting model, leading to a successful model recovery. Numerical values can also be found in S1 Data.



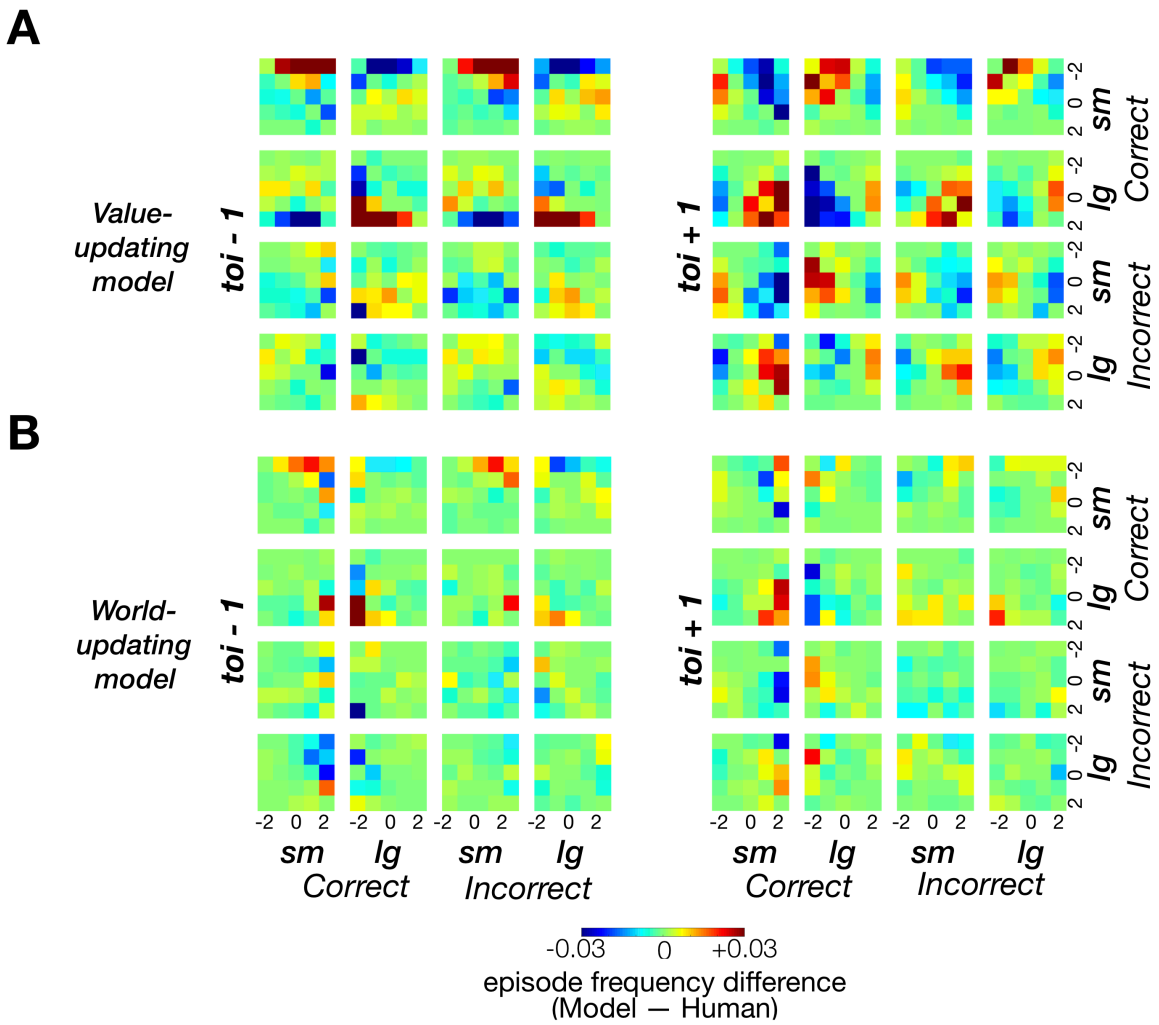
S4 Fig. Histograms of classification accuracies of the human participants and their model partners in the *ex ante* simulations.

(A, B) Across-individual distributions of the classification accuracy of the belief-based RL model **(A)** and BMBU **(B)** overlaid on those of the human participants. The models' choices were generated via *ex ante* simulations with a specific set of model parameters (Table A in S1 Appendix), the results of which are depicted in Fig 4 and Fig 5. The classification accuracy is measured by calculating the percentage of the trials in which the choice matched the feedback used in the actual experiment. The empty bars correspond to the histogram of human performances, the range of which is demarcated by the dashed vertical lines ([min, max]=[60.65%, 73.94%]). The average human classification accuracy was 67.85%. **(A)** Comparison of classification accuracy between the belief-based RL model's simulation (red color) and the human choices. The model's *ex ante* simulation accuracy was not different from the human accuracy ($t(53) = 1.4429$, $P = 0.1549$; Null hypothesis: model's performance vector and humans' performance vector come from populations with equal means, unpaired two-tailed *t*-test). **(B)** Comparison of classification accuracy between BMBU's simulation (green color) and the human choices. The model's *ex ante* simulation accuracy was not different from the human accuracy ($t(53) = 0.9707$, $P = 0.3361$, unpaired two-tailed *t*-test). There was no significant difference in classification accuracy between the value-updating model and BMBU ($t(48) = 0.5733$, $P = 0.5691$, unpaired two-tailed *t*-test). The data underlying this figure **(A, B)** can be found in S1 Data.



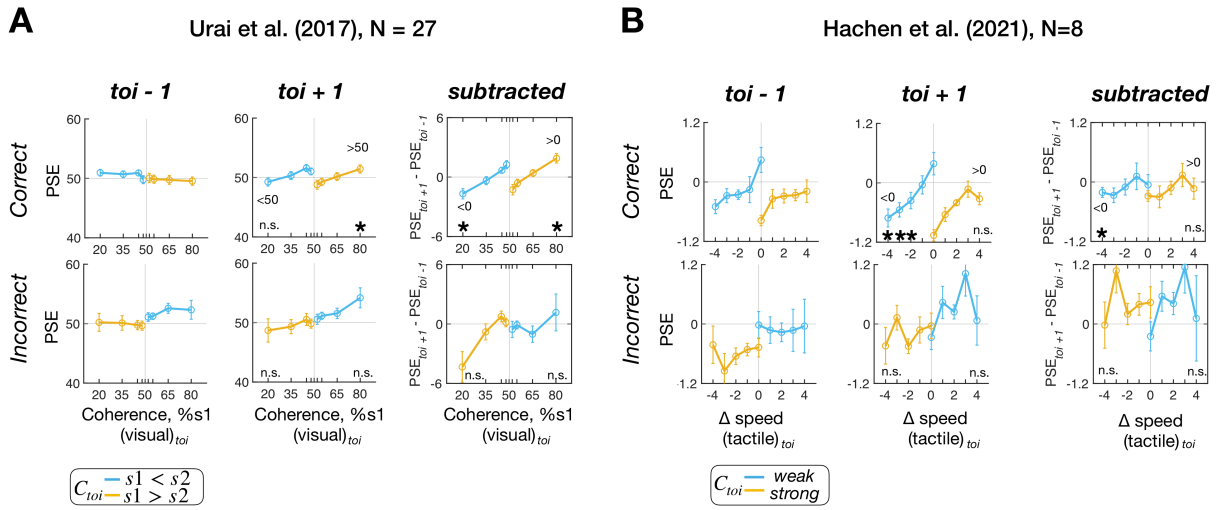
S5 Fig. Retrospective (left columns), prospective (middle columns), and subtractive (right columns) history effects in PSE for the 'Hybrid' model's *ex post* model simulations.

Top and bottom rows in each panel show the PSEs associated with the *toi* episodes involving $correct$ and $incorrect$ feedback at *toi*. Symbols with error bars, mean±s.e.m across the 30 model agents, which correspond to their 30 human partners. The colors of the symbols and lines label choices (blue: $small$ and yellow: $large$). The data underlying this figure can be found in S1 Data.



S6 Fig. Maps of frequency deviations of the value-updating (A) and world-updating (B) model agents' classifications in the *ex post* simulations from the human decision-makers in the retrospective (left) and prospective (right) history effects.

Each cell represents a pair of PDM episodes, as specified by the column and row labels. At each cell, the color represents how much the episode frequency observed in the model agents deviates from that observed in the corresponding human decision-makers. The results of statistical tests on these deviations are summarized in Fig 7E and 7F. The data underlying this figure **(A, B)** can be found in S1 Data.



S7 Fig. Retrospective (left columns), prospective (middle columns), and subtractive (right columns) history effects in PSE for the human classification performances of Urai et al. (2017)'s work [37] **(A) and Hachen et al. (2021)'s work** [31] **(B).**

(A, B) We downloaded both publicly available datasets, analyzed them in the same way that we analyzed human observers in our work, and plotted the results in the same format used for Fig 7A. Top and bottom rows in each panel show the PSEs associated with the *toi* episodes involving $correct$ and $incorrect$ feedback. Symbols with error bars, mean±s.e.m. across human observers. The colors of the symbols and lines label choices (blue: $small$ and yellow: $large$). The overall patterns of the PSEs plotted here appear similar to those plotted in Fig 7A, displaying the reversals in direction of stimulus-dependent feedback effects. When the same statistical tests used in our work were carried out, some of the data points at the stimuli with strong sensory evidence at *toi* significantly deviated from zero in the direction opposite to the feedback effect predicted by the value-updating scenario, as indicated by the asterisks. **(A)** Sequential features of human observers (N=27) analyzed in our way from human dataset that once had been published [37], which is openly available (http://dx.doi.org/10.6084/m9.figshare.4300043), then analyzed in the previous study [9]. In this study, the participants performed a binary classification task on the difference in motion coherence by sorting the pairs of random-dot-kinematogram stimuli shown in two intervals (s1 and s2) into one of the two classes ('s1<s2' vs. 's1>s2') over consecutive trials. The presented stimuli were taken from 3 sets of difficulty levels (the difference between motion coherence of the test and the reference stimulus; easy: [2.5, 5, 10, 20, 30], medium: [1.25, 2.5, 5, 10, 30], hard: [0.625, 1.25, 2.5, 5, 20]). As done in the original study [9], we binned the trials into 8 levels by merging the trials of two neighboring coherence levels (e.g., the coherence levels of [0.625, 1.25]) into a single bin. Note that the coherence bins of [20, 35, 45, 48.75, 51.25, 55, 65, 80] (%s1) on the x-axis (50% represents the equal coherence between s1 and s2) are matched to the x-axis in Figure 8 of the previous study in which the same dataset had been used. Asterisks mark the significance of one-sample *t*-tests (uncorrected $P<0.05$, one-tailed in the direction of feedback effects) on the panel *toi+1* (stimulus 80%: $t(26) = 2.0138$, P =

0.0272) and on the panel *subtracted* (stimulus 20%: $t(26) = -3.1900$, $P = 0.0018$, stimulus 80%: $t(26) = 3.8810$, $P = 0.0003$). **(B)** Sequential features of human observers (N=8) published in another previous study [31]. We used the human dataset openly available as part of the repository (https://osf.io/hux4n). In this study, the participants performed a binary classification task on the speed of vibrotactile stimuli by classifying the speed of the presented vibration as 'low-speed (weak)' or 'high-speed (strong)'. Note that the nine-level stimuli of [−4,−3,−2,−1,0,1,2,3,4] on the x-axis followed how data were encoded by the original study [31]. Asterisks mark the significance of one-sample *t*-tests (uncorrected $P<0.05$, one-tailed in the direction of feedback effects) on the panel *toi+1* (stimulus −4: $t(7) = -3.6757$, $P = 0.004$, stimulus −3: $t(7) = -3.5252$, $P = 0.0048$, and stimulus −2: $t(7) = -2.0325$, $P = 0.04$) and on the panel *subtracted* (stimulus −4: $t(7) = -1.9848$, $P = 0.044$). The data underlying this figure **(A, B)** can be found in S1 Data.

Corrective feedback guides human perceptual decision-making by informing about the world state rather than rewarding its choice

S1 Appendix
Supporting details

Hyang-Jung Lee[1], Heeseung Lee[1], Chae Young Lim[2], Issac Rhim[3], Sang-Hun Lee[1]

[1]Department of Brain and Cognitive Sciences, Seoul National University, Seoul, South Korea
[2]Department of Statistics, Seoul National University, Seoul, South Korea
[3]Institute of Neuroscience, University of Oregon, Eugene, OR, USA

*Correspondence: visionsl@snu.ac.kr

Supplemental details (Text) *on additional model specifications of BMBU*

Here, we will elaborate on "**World-updating model**", a subsection found in the **Materials and methods** of the main text. We will provide more details on the boundary-updating operation of BMBU by explaining how Equation 10 is derived from Equation 9 in the main text.

Equation 8 in the main text defines how the likelihood function of the class boundary B_t given two pieces of evidence, the mnemonic measurement m_t'—i.e., noisy memory recall of the sensory measurement m_t in working memory—and the state of the class variable CL_t informed by feedback. The boundary likelihood, which represents the probabilities of those two observations under any hypothesized state of B_t, is calculated through integration (marginalization) over all possible states of the unknown variable S_t, and can be re-written as:

$$p(m_t', CL_t|B_t) = \int p(m_t'|S_t)p(CL_t|S_t, B_t)p(S_t|B_t)dS_t = \int L_{m_t'}(S_t)L_{CL_t}(S_t, B_t)p(S_t|B_t)dS_t, \quad (S1)$$

where each integrand is elaborated in the following.

The factor $p(m_t'|S_t)$ in Equation S1 corresponds to the information about the stimulus offered by the mnemonic measurement m_t'. The likelihood of any hypothetical state of S_t given m_t' refers to the

probability of m_t' if that hypothetical state is true. Here, we denote $p(m_t'|S_t)$ by a likelihood function, $L_{m_t'}(S_t)$.

Through marginalization over m_t from the learned generative model, which is described by Equations 3-6 in the main text, $L_{m_t'}(S_t) \equiv p(m_t'|S_t)$ can be expressed in a form of the Gaussian function, as follows:

$$p(m_t'|S_t) = \int p(m_t'|m_t)p(m_t|S_t)dm_t = \int N\left(m_t'; m_t, \sigma_{m'}^2\right)N(m_t; S_t, \sigma_m^2)dm_t$$
$$= \int \frac{1}{\sqrt{2\pi\sigma_{m'}^2}} e^{-\frac{(m_t'-m_t)^2}{2\sigma_{m'}^2}} \frac{1}{\sqrt{2\pi\sigma_m^2}} e^{-\frac{(m_t-S_t)^2}{2\sigma_m^2}} dm_t = N\left(m_t'; S_t, \sigma_{m'}^2 + \sigma_m^2\right), \qquad (S2)$$

where the expressions $p(m_t'|m_t)p(m_t|S_t)$ are substituted by the two Gaussian noise distributions as defined in the learned generative model (Equation 5-6 in the main text) and integrated over the possible states of the sensory measurement variable m_t, which is now unknown—inaccessible any longer. Thus, we find that $L_{m_t'}(S_t)$, the likelihood of S_t, follows $N\left(S_t; m_t', \sigma_{m'}^2 + \sigma_m^2\right)$.

Next, the factor $p(CL_t|S_t, B_t)$ in Equation S1 corresponds to the information about the stimulus and the class boundary offered by the state of the class variable CL_t informed by feedback, either $small$ or $large$. The likelihood of any hypothetical joint states of S_t and B_t given CL_t refers to the probability of CL_t if that hypothetical state is true. Here, we denote $p(CL_t|S_t, B_t)$ by a likelihood function, $L_{CL_t}(S_t, B_t)$. $L_{CL_t}(S_t, B_t)$ is determined depending on the inequality between S_t and B_t: in the case of $CL_t = small$,

$$L_{CL_t=small}(S_t, B_t) = p(CL_t = small|S_t, B_t) = \begin{cases} 1, & S_t < B_t \\ 0, & S_t > B_t \\ 0.5, & S_t = B_t \end{cases} \qquad (S3)$$

; in the case of $CL_t = large$,

$$L_{CL_t=large}(S_t, B_t) = p(CL_t = large|S_t, B_t) = \begin{cases} 0, & S_t < B_t \\ 1, & S_t > B_t \\ 0.5, & S_t = B_t \end{cases}. \qquad (S4)$$

Now, let us get back to Equation S1 and denote the boundary likelihood function in the case of $CL_t = small$ (abbreviated as s) by $L_{m'_t,CL_t=s}(B_t)$, which can further be decomposed into two integrals with finite limits (as similarly done for Equation 9 in the main text), as follows:

$$L_{m'_t,CL_t=s}(B_t) \equiv p(m'_t, CL_t = s|B_t)$$

$$= \int_{S_t=-\infty}^{S_t=B_t} p(m'_t|S_t)p(CL_t = s|S_t, B_t)p(S_t|B_t)dS_t + \int_{S_t=B_t}^{S_t=+\infty} p(m'_t|S_t)p(CL_t = s|S_t, B_t)p(S_t|B_t)dS_t,$$

which can be rewritten as follows:

$$= \lim_{r\to\infty, b\to B_t^-} \int_{-r}^{b} p(m'_t|S_t)p(CL_t = s|S_t, B_t)p(S_t|B_t)dS_t$$

$$+ \lim_{r\to\infty, b\to B_t^+} \int_{b}^{r} p(m'_t|S_t)p(CL_t = s|S_t, B_t)p(S_t|B_t)dS_t. \tag{S5}$$

Since the last term on the right-hand side of Equation S5 becomes zero by Equation S3 (for any value ranges of S_t larger than B_t, $p(CL_t = s|S_t, B_t) = 0$), the boundary likelihood function in the case of $CL_t = small$ is reduced as follows:

$$L_{m'_t,CL_t=s}(B_t) \equiv p(m'_t, CL_t = s|B_t) = \int_{-\infty}^{B_t} p(m'_t|S_t)p(S_t|B_t)dS_t. \tag{S6}$$

From Equation 3 in the main text and Equation S2, we know the distributions $p(S_t|B_t)$ and $p(m'_t|S_t)$, respectively. Substituting the expression for these distributions gives:

$$L_{m'_t,CL_t=s}(B_t)$$

$$= \int_{-\infty}^{B_t} \frac{1}{\sqrt{2\pi(\sigma_{m'}^2+\sigma_m^2)}} e^{-\frac{(m'_t-S_t)^2}{2(\sigma_{m'}^2+\sigma_m^2)}} \frac{1}{\sqrt{2\pi\sigma_S^2}} e^{-\frac{(S_t-B_t)^2}{2\sigma_S^2}} dS_t = \frac{1}{\sqrt{2\pi\left(\frac{\sigma_M^2\sigma_S^2}{(\sigma_M^2+\sigma_S^2)}\right)}} \int_{-\infty}^{B_t} e^{-\frac{\left(S_t-\frac{B_t\sigma_M^2+m'_t\sigma_S^2}{\sigma_M^2+\sigma_S^2}\right)^2}{2\frac{\sigma_M^2\sigma_S^2}{(\sigma_M^2+\sigma_S^2)}}} dS_t \times \frac{1}{\sqrt{2\pi(\sigma_M^2+\sigma_S^2)}} e^{-\frac{(B_t-m'_t)^2}{2(\sigma_M^2+\sigma_S^2)}}, \tag{S7}$$

where $\sigma_M^2 = \sigma_{m'}^2 + \sigma_m^2$. Equation S7 is equivalent to Equation 11 in the main text.

To aid in intuitive comprehension of $L_{m'_t,CL_t=s}(B_t)$, we can express it as the product of two terms from Equation S7 by rewriting it as follows:

$$L_{m_t',CL_t=s}(B_t) = F_X(B_t)\mathcal{L}_{m_t'}(B_t), \qquad (S8)$$

where F_X denotes the first term on the right-hand side of Equation S7 (to be detailed in Equation S10), and $\mathcal{L}_{m_t'}$ denotes the second term on the right-hand side of Equation S7 (to be detailed in Equation S9), which equals to the likelihood function for B_t given m_t' defined under the assumption that no CL_t variable exists in the generative process. To make this point explicit, we used a different likelihood notation $\mathcal{L}$ from the notation L used throughout the paper.

Note that, according to the learned generative model in our study, the decision-maker acquires the knowledge about B_t by evaluating how probable the two observations, m_t' (sensory evidence in memory; simply as sensory evidence, hereinafter) and CL_t (feedback evidence), are for each possible value of B_t. Contrastingly, $\mathcal{L}_{m_t'}(B_t)$ in Equation S8 can be considered the information about B_t solely based on the sensory evidence by computing $p(m_t'|B_t)$. Thus, $\mathcal{L}_{m_t'}(B_t)$ represents the "sensory influence" that drives the boundary update, by pushing the joint boundary likelihood $L_{m_t',CL_t=s}(B_t)$ towards m_t' on the B_t axis (see the first row in S1B-D Fig), as follows:

$$\mathcal{L}_{m_t'}(B_t) = \frac{1}{\sqrt{2\pi(\sigma_M^2+\sigma_S^2)}} e^{-\frac{(B_t-m_t')^2}{2(\sigma_M^2+\sigma_S^2)}} = N(B_t; m_t', \sigma_M^2+\sigma_S^2). \qquad (S9)$$

On the other hand, the term $F_X(B_t)$ in Equation S8 can be considered the information about B_t jointly based on the feedback and sensory evidence, while parting out the aforementioned "sensory influence" (divided by $\mathcal{L}_{m_t'}(B_t)$). Let X be a random variable with cumulative distribution function (CDF) F_X

$$F_X(B_t) = \int_{-\infty}^{B_t} f_X(S_t)dS_t\,, \qquad (S10)$$

where the probability density function of X, denoted by $f_X(x) = N\left(x; \frac{B_t\sigma_M^2+m_t'\sigma_S^2}{\sigma_M^2+\sigma_S^2}, \frac{\sigma_M^2\sigma_S^2}{(\sigma_M^2+\sigma_S^2)}\right)$. Since this term is derived specifically for $CL_t = small$, we can interpret this term as the "feedback influence" that drives the boundary update with a CDF multiplied, by pushing the boundary likelihood $L_{m_t',CL_t=s}(B_t)$ in a more positive direction on the B_t axis. Owing to this contribution by the feedback evidence, the boundary

likelihood $L_{m'_t,CL_t=s}(B_t = b)$ for a given value b of B_t would support that $b > m'_t$ is more plausible than $b < m'_t$ for the current state of the class boundary that generated m'_t and $CL_t = small$. This aligns with the intuition from the temperature example described in the main text (Fig 1C).

Similarly, from Equations S5-S7, we derive the boundary likelihood for the case $CL_t = large$:

$$L_{m'_t,CL_t=l}(B_t) = \int_{B_t}^{\infty} p(m'_t|S_t)p(S_t|B_t)dS_t = (1 - F_X(B_t))\mathcal{L}_{m'_t}(B_t), \qquad (S11)$$

Contrary to the $CL_t = small$ case, the boundary likelihood $L_{m'_t,CL_t=l}(B_t = b)$ would support that $b < m'_t$ is more plausible than $b > m'_t$ for B_t that generated m'_t and $CL_t = large$, since the multiplication is performed, instead with a complementary CDF (see the second row in S1B-D Fig).

Supplementary tables (A-D Tables) *to support the Results section*

Table A. Parameters used for *ex ante* simulations.

World		Value	
σ_m	25 levels $[0.15, 3.27]$	σ_m	25 levels $[0.15, 3.27]$
μ_0	0	μ_0	0
σ_0	5	α	0.35
σ_s	1.5811	β	5
$\sigma_{m'}$	2.5	V_{init}	0.5
$\sigma_{diffusion}$	0.8		

Table B. Parameters recovered from fitting the main models, world-updating and value-updating models, to human choices (N=30**).**

World	Mean (standard deviation)	Value	Mean (standard deviation)
σ_m	0.86 (0.46)	σ_m	1.20 (0.30)
μ_0	0.04 (0.62)	μ_0	0.02 (0.43)
σ_0	4.27 (1.22)	α	0.17 (0.20)
σ_s	0.84	β	7.43

	(0.92)		(7.01)
$\sigma_{m'}$	3.70 (1.00)	V_{init}	0.92 (0.16)
$\sigma_{diffusion}$	3.20 (1.57)		

Table C. Parameters recovered from fitting the rest of the models to human choices (N=30).

Hybrid	Mean (standard deviation)	Fixed	Mean (standard deviation)	Base	Mean (standard deviation)
σ_m	0.68 (0.33)	σ_m	1.54 (0.39)	σ_m	1.90 (0.50)
μ_0	0.01 (0.61)	μ_0	0.02 (0.31)		
σ_0	3.55 (1.28)				
σ_s	0.86 (1.03)				
$\sigma_{m'}$	3.84 (1.22)				
$\sigma_{diffusion}$	3.77 (1.87)				
α	0.15 (0.11)				
β	6.96 (2.83)				
V_{init}	0.882 (0.15)				

Table D. Statistical results on model behavior versus human behavior in terms of PSE measures.

Model vs. Data	*toi-1* (20 conditions)				*toi+1* (20 conditions)				*(toi+1) - (toi-1)* (20 conditions)				Total (60 conditions)	
Statistical difference	n.s.		significant		n.s.		significant		n.s.		significant		n.s. # conditions	
Test type	*t*	*w*	*t*	*w*	*t*	*w*	*t*	*w*	*t*	*w*	*t*	*w*	*t*	*w*
Value	(13)	(13)	*(1) **(2) ***(4)	*(1) **(5) ***(1)	(10)	(9)	*(1) **(1) ***(8)	*(3) **(3) ***(5)	(8)	(9)	*(1) **(1) ***(10)	**(6) ***(5)	(31)	(31)
World	(20)	(20)	.	.	(20)	(20)	.	.	(20)	(20)	.	.	(60)	(60)
Hybrid	(20)	(20)	.	.	(17)	(19)	*(3)	*(1)	(19)	(20)	*(1)	.	(56)	(59)

Test type, *t*: Paired t-test
Test type, *w*: Wilcoxon signed-rank test

1605 *: $P<0.00083$, Bonferroni-corrected threshold
1606 **: $P<0.000167$,
1607 ***: $P<0.0000167$