What Humans Learn from Corrective Feedback for Perceptual Decision-Making

Hyang-Jung Lee[1], Heeseung Lee[1], Chae Young Lim[2], Issac Rhim[3], and 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]Department of Psychology, The University of Texas Austin, TX, U.S.A.

*To whom correspondence should be addressed (visionsl@snu.ac.kr)

January 11st, 2023

Abstract

Corrective feedback is vital to learning about the environment, exerting its influence on subsequent episodes of perceptual decision-making (PDM) on a trial-to-trial basis. However, its intricate interplay with other history factors in PDM challenges a principled account of the history effects of feedback. One popular account, grounded in animal behavior and extended to human behavior, grafted the principle proven successful in reward-based decision-making—reinforcement-learning—onto the PDM processes. It assumes that decision-makers, although engaged in PDM, treat feedback as *rewards* from which they learn *choice values*. Here we report that human decision-makers do not conform to this assumption. Instead, humans acted like treating feedback as *evidence* for a specific world state and combining it with sensory evidence to learn the *statistical structure* crucial to perform a PDM task. A Bayesian implementation of our account reproduces human history effects, both retrospectively and prospectively, with an unprecedented level of correspondence and finesse.

Introduction

PDM means committing to a certain proposition about an objective state of the world (e.g., 'this mountain is high'). 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 as a return. Among these *history factors*, which are termed 'stimulus', 'choice', and 'feedback', *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 across sensory modalities[1–10], much remains elusive about how corrective feedback, in conjunction with other history factors, exerts its trial-to-trial effects 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'). The reinforcement learning (RL) algorithm has been offering successful accounts of how past rewards affect subsequent VDM based on error-driven incremental mechanisms[11–15]. Intriguingly, there have been attempts to explain the impacts of feedback on PDM by grafting the RL algorithm onto the PDM processes[3,4,8–10]. This grafting premises that decision-makers treat corrective feedback in PDM as reward feedback in VDM. On this premise, decision-makers would form and update the *values* of the choices based on reward prediction error signals (red dashed arrows in Fig. 1a). Suppose a decision-maker committed to a proposition, 'this mountain is large'. Then, $correct$ and $incorrect$ feedback to that commitment increases and decreases, respectively, the value of the 'large' choice. This RL-grafting account, which will be referred to as *value-updating scenario* hereinafter, appears to be a natural scenario given the fact that corrective feedback was typically provided in physical rewards such as juice or water in animal PDM experiments[4,5,8–10,16–18]. The value-updating scenario seems plausible from the perspective that PDM and VDM might share common mechanisms[19], as suggested by some empirical studies[20,21].

Nevertheless, the 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 to which they committed is true or false in the world. For example, $correct$ and $incorrect$ feedback to 'this mountain is large' indicate that 'this mountain' is higher and lower, respectively, than the average size of mountains in the world. Thus, in this alternative scenario, feedback allows decision-makers to learn the world statistics pertinent to a given PDM task (green dashed arrows in Fig. 1a). We will call this *world-updating scenario*. The world-updating scenario is conceptually parsimonious in that it does not require any components outside the PDM processes, such as the RL algorithm developed in the VDM. Especially in the framework of Bayesian Decision Theory (BDT)[22,23], which has been providing compelling accounts for PDM behavior, the world statistics is a crucial part of the internal (generative) model that is required to infer the world state of interest in PDM[24–27].

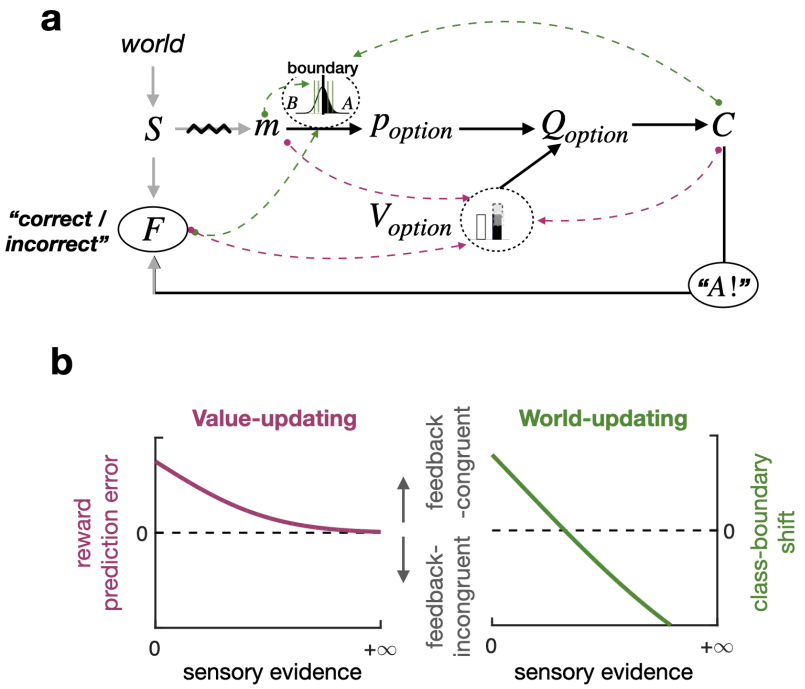


Fig. 1 | Two possible scenarios for what humans learn from feedback for PDM and their distinct predictions of feedback effects. a, Decision platform for perceptual binary classification. The gray arrows depict how a sensory measurement m and feedback F are generated from a stimulus S, which is sampled from the $world$, and a decision d. 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 d 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, d, and F. **b**, Distinct sensory-evidence-dependent feedback effects predicted by the two scenarios. According to the value-updating scenario (left), as the sensory evidence grows, P_{option}, and accordingly Q_{option} too, grows. As a result, reward prediction errors become smaller but remains in the direction congruent with feedback, which predicts that feedback effects on subsequent trials diminish asymptotically as a function of sensory evidence. According to the world-updating scenario (right), as sensory evidence grows, 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 like that predicted by the value-updating scenario for weak sensory evidence but becomes eventually reversed to the direction incongruent with feedback as sensory evidence grows.

Given these two seemingly plausible scenarios for feedback effects, we test which scenario better explains the effects of corrective feedback—without any form of physical reward—on humans' PDM. To do so, we implemented the value-updating and world-updating scenarios into the confidence-dependent RL model[9] and the Bayesian model of boundary-updating, 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 of the most widely used PDM tasks in which decision-makers classify the stimuli that vary continuously in a certain feature dimension (e.g., size) into two discrete classes (e.g., 'small' vs

‘large’), since the two scenarios make distinct predictions about the stimulus-dependent feedback effects in this task. As will be described in detail later, the value-updating scenario predicts that feedback, which acts like rewards, “unidirectionally” fosters and suppresses the rewarded ($correct$) and unrewarded ($incorrect$) choices, respectively, in subsequent trials while such unidirectional feedback effects diminish 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 grows, due to the shift of the class boundary toward 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. 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 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 observe 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 and 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 to vary sufficiently—including the ones very easy and difficult for classification—so that the two scenarios’ distinct predictions on the stimulus-dependent feedback effects can be readily compared. Also, we used stochastic feedback, where $correct$ and $incorrect$ feedback was very occasionally given to incorrect and correct choices, respectively (Experimental methods section), to cover the entire 3D space of decision-making episodes defined over ‘stimulus’, ‘choice’, and ‘feedback’ ($5 \times 2 \times 2 = 20\ episodes$; Fig. 2c).

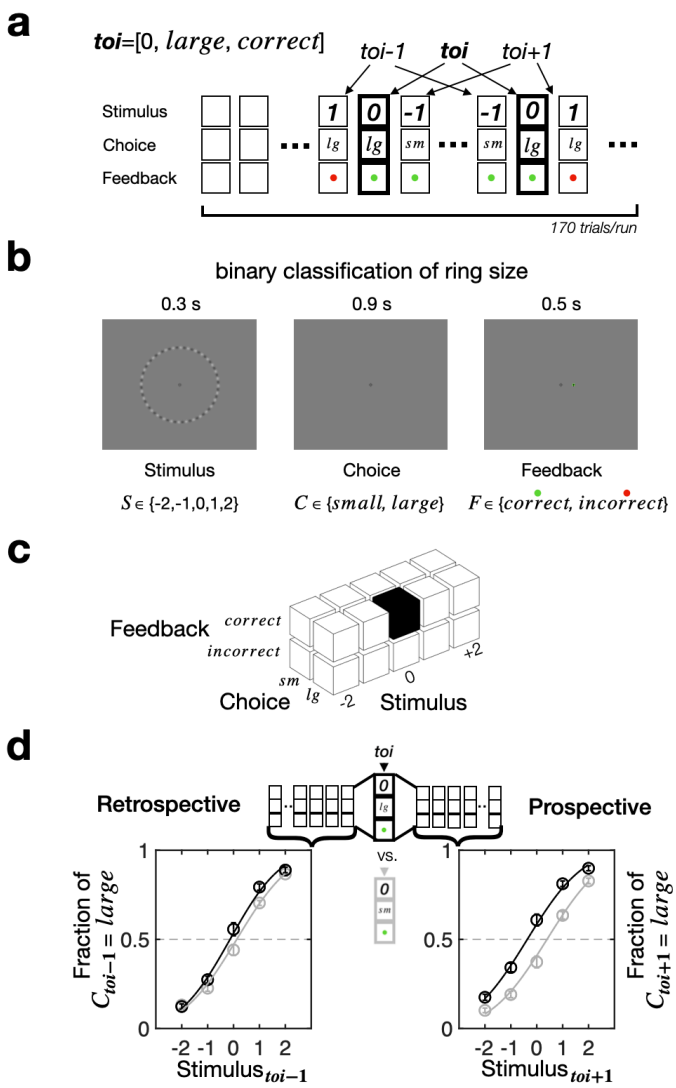


Fig. 2 | Experimental design and definition of retrospective and prospective history effects. a, A chain of PDM episodes over a single sequence of trials. Each trial sequence consists of 170 column vectors of PDM episode [stimulus; choice; feedback]. In this example, the trial of interest (*toi*) is defined as $[0;\ large;\ correct]$ and demarcated by thick outlines. Accordingly, 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, which indicated by the color around the fixation whether the classification was correct or incorrect. **c**, 3D state space of the PDM episodes in the current 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.t.e. across 30 participants, respectively.

To evaluate rigorously the correspondence between model predictions 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 (the trials 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*) *toi*. Second, we derived the two psychometric curves based on the retrospective and prospective blocks of trials, respectively, and fit the cumulative normal distribution function to these curves to estimate the point-of-subjective-inequality (PSE) measures, which have previously been used[16–18] and known[28] to reliably estimate the history-dependent choice biases in PDM. Thus, the PSEs of the retrospective and prospective trials quantify the choice biases that exist *before* and *after*, respectively, the PDM episode of interest occurs, 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 is larger and smaller, respectively, than the class boundary (B) based on 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 can arise at the perception or valuation stage. Suppose the decision-maker's internal model of size distribution at the perception stage is not fixed but changes depending 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).

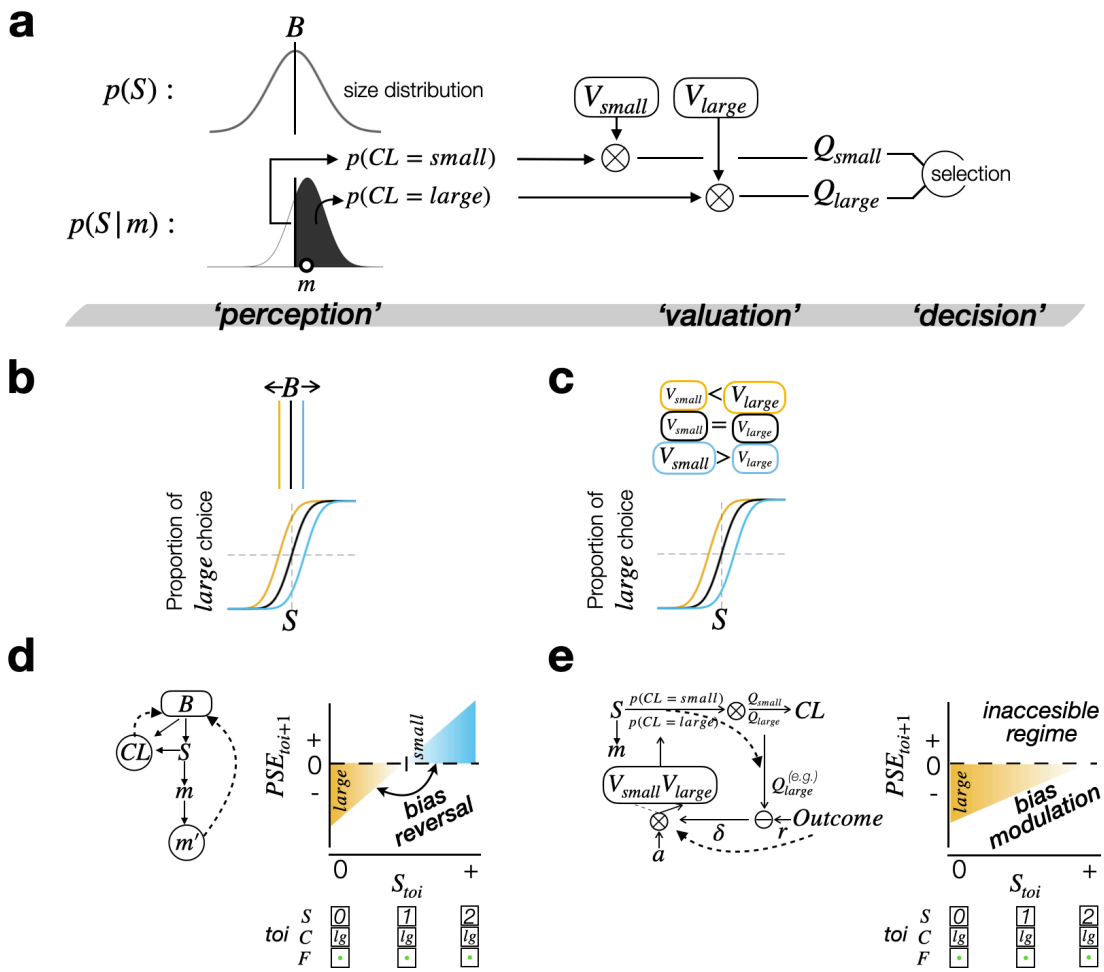


Fig. 3 | Implementation of the value-updating and world-updating scenarios into computational models. a, Computational elements along the three stages of PDM for binary classification. At '*perception'* stage, the probabilities that the class variable takes its binary states $small$ and $large$—$p(CL = large)$ 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 '*valuation'* stage, the outcomes of the perception stage are multiplied by the learned values Vs to produce the expected values Qs. At '*decision'* stage, the choice with the greater expected value is selected. **b**,**c**, Illustration of two potential origins of choice biases, one at '*perception'* stage (**b**) and the other at '*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) exerts 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 *toi*, 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 the 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 this model incorporates the statistical decision confidence at the perception stage, i.e., $p(CL = large(small))$, as one of the three arguments of the reward prediction error function. This feature, as stressed by the authors who developed this algorithm[9], 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 confidence-based modulation of value-updating underlies the stimulus-dependent feedback effects: the amount of feedback effects decreases as the 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 sensory measurement in memory', m', BMBU infers the mean of 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 internal model of how m' and CL are generated (Fig. 3d, left), as follows:

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

This algorithm maps the specific ensembles of 'stimulus', 'choice', and 'feedback' (since C and F jointly determine CL) in the previous trial onto the choice biases with the specific directions and magnitudes in the following trial. Thus, with this algorithm, we can predict how feedback effects depend on stimuli under the world-updating scenario. Suppose a decision-maker currently believes that the size distribution is centered at 0, views a ring with the size of 0, classifies it as $large$—probably owing to a random sensory noise, and receives $correct$ feedback. Following this episode (i.e., $episode = \{S = 0, C = large, F = correct\}$), the choice will be biased to the small side because $correct$ feedback to the $large$ choice (, which means $CL = large$,) will shift the class boundary to the $small$ side whereas the size of 0 will hardly move the class boundary. Next, let us vary only the stimulus variable gradually in the above episode to

appreciate the stimulus-dependent feedback effects in the world-updating scenario. As the size grows (e.g., $S = 0 \rightarrow 1 \rightarrow 2$), the contribution of the size variable to class boundary shift increasingly grows, shifting the boundary farther and farther to the $large$ side, in the direction counteracting that of the feedback contribution (Fig. 3d, right). Hence, because the bias of the class boundary is inversely correlated with that of choice bias, the choice is initially biased to the $large$ side but increasingly biased to the $small$ side as the size grows.

We stress that this 'bias direction reversal as a function of sensory evidence' (we will refer to this as 'bias reversal' in short hereafter) is a hallmark of the world-updating scenario's prediction of the history effects in PDM. Specifically, the direction of 'bias reversal' is always $small$-to-$large$ as long as the feedback in conjunction with the choice indicates $CL = small$ (e.g., $episodes = \{S = 0 \rightarrow -1 \rightarrow -2, C = small, F = correct\}\ or\ \{S = 0 \rightarrow -1 \rightarrow -2, C = large, F = incorrect\}$ and always $large$-to-$small$ as long as the feedback in conjunction with the choice indicates $CL = large$ (e.g., $episodes = \{S = 0 \rightarrow 1 \rightarrow 2, C = large, F = correct\}\ or\ \{S = 0 \rightarrow 1 \rightarrow 2, C = small, F = incorrect\}$). 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 rather conceptually explained why and how the two scenarios imply the distinct patterns of stimulus-dependent feedback effects: as sensory evidence grows, the direction of feedback effects is reversed under the world-updating scenario but not under the value-updating scenario. Though this implication seems intuitively apparent, it must be confirmed with the experimental protocol of the current study. Moreover, there are good reasons to expect that any history effect will exhibit complex dynamics over a sequence of trials. First, sensory evidence is subject to stochastic noises, which propagate 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 value/boundary-updating process, and so on). Second, provided that any deterministic (i.e., 'not stochastic') value/boundary-updating processes are presumed to be at work, the PDM episode on a given trial, in principle, must be probabilistically conditioned on the episodes in past trials (e.g., the current $small$ choice on the ring of $S = 0$ is likely to follow 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 updatings on the history effects must be complicated and thus be inspected with realistic simulation. Further, considering that these multiple stochastic and deterministic events together interplay with one another to create diverse temporal contexts, history effects are supposed to reveal themselves in multiplexed dynamics.

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 (see Supplementary Table 1, Supplementary Fig. 1, and Modeling details section in Methods). The simulation results confirmed our intuition, as summarized in Figure 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 that they are about to make in the—following—*toi* (Fig. 4b,e). 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 the presence of the complex dynamics of history effects discussed above and is also consistent with what has been previously reported (e.g., see Figure 2 of the previous study[9]). Importantly, in line with our conceptual conjecture laid out above (Fig. 3d,e), the two agents evidently disagree on the prospective history effects. While the value-adapting agent always exhibits the feedback-congruent bias but never reverses the direction of bias, the world-adapting 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,d,f,g,h).

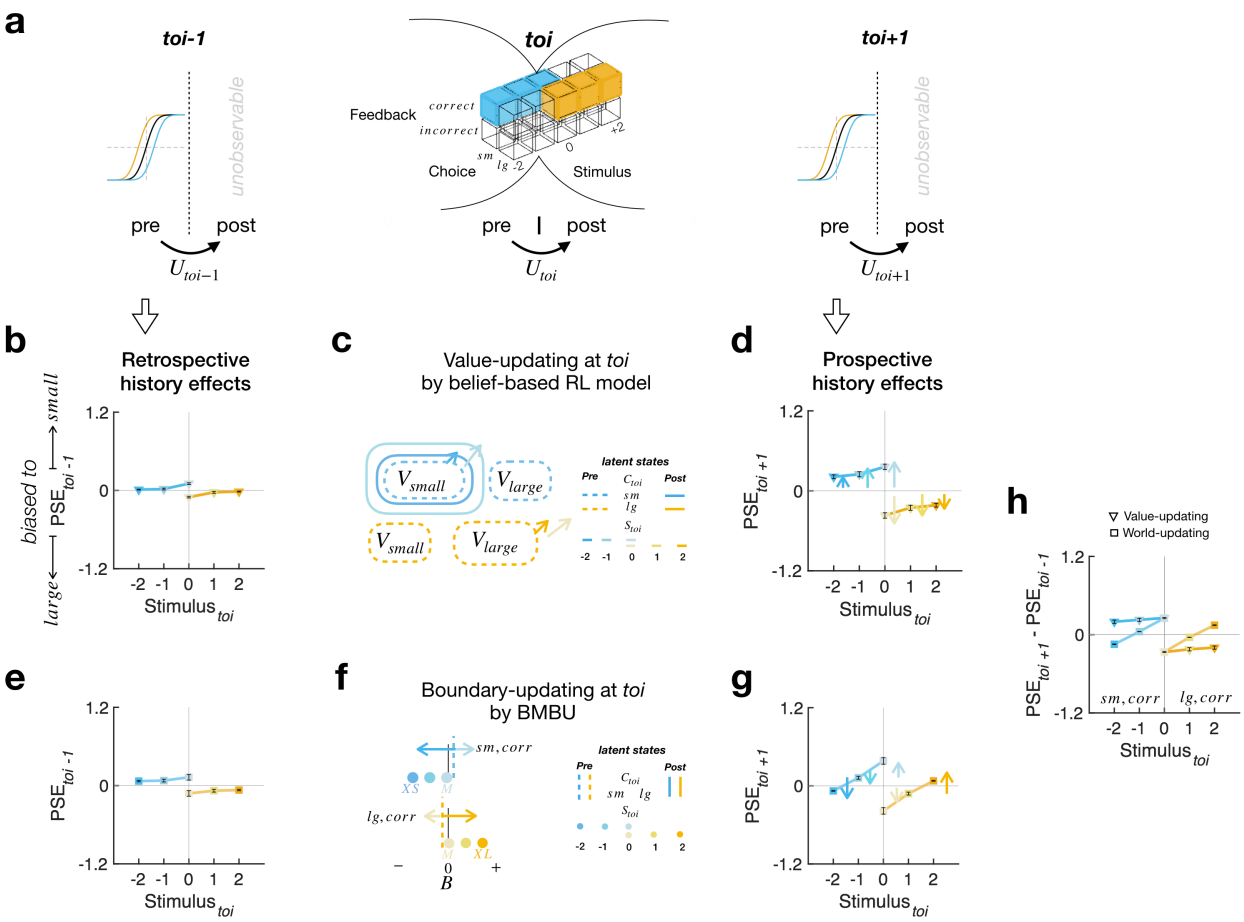


Fig. 4 | *Ex ante* simulation results for the PDM episodes with $correct$ feedback. a, Illustration of how the retrospective (left) and prospective (right) history effects are related to the value/boundary-updatings (bottom) occurring over the trials overarching the trial of interest. While the updating occurs latently at every trial (U_{toi-1}, U_{toi}, U_{toi+1}), one can access its behavioral consequences only at the pre-updating phase at *toi-1* and *toi+1*. **b**,**c**,**d**, The observable retrospective (**b**) and prospective (**d**) history effects and latent value-updating processes (**c**) displayed by the value-updating model agent (v-shaped symbol). Here, only the chosen value is updated positively with the amount dependent on the stimulus size, as indicated by the arrows. **e**,**f**,**g**, The same history effects (**e**,**g**) and latent boundary-updating processes (**f**) displayed by the world-updating model agent (square-shaped symbol). Here, the class boundary is shifted as a joint function of feedback and stimulus size, where the boundary shifts due to stimulus size counteract those due to feedback, as indicated by the arrows. All PSE values were taken from the mean across parameter sets used in *ex ante* simulations (see Methods; error bars show s.e.m.). **h**, Juxtaposition of the difference between the retrospective and prospective history effects displayed by the two model agents.

Next, Figure 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,e). These pronounced retrospective history 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 the sensory evidence increases since the prior value ratio or class boundary must be strongly biased to result in an $incorrect$ 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,d,f,g). Thus, as in the case for the $correct$-feedback episodes, the direction reversal is displayed only by the world-updating agent, but not by the value-updating agent (Fig. 5h).

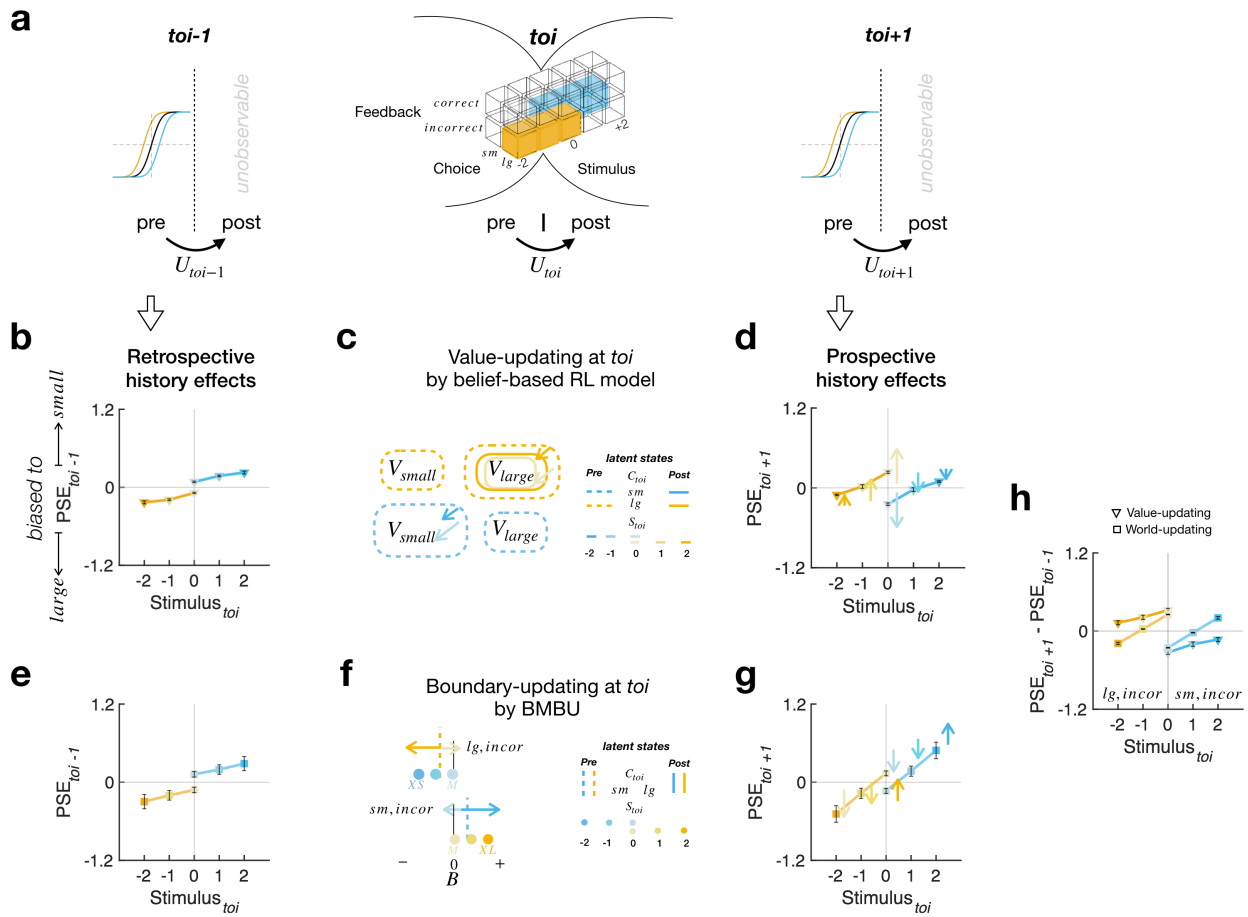


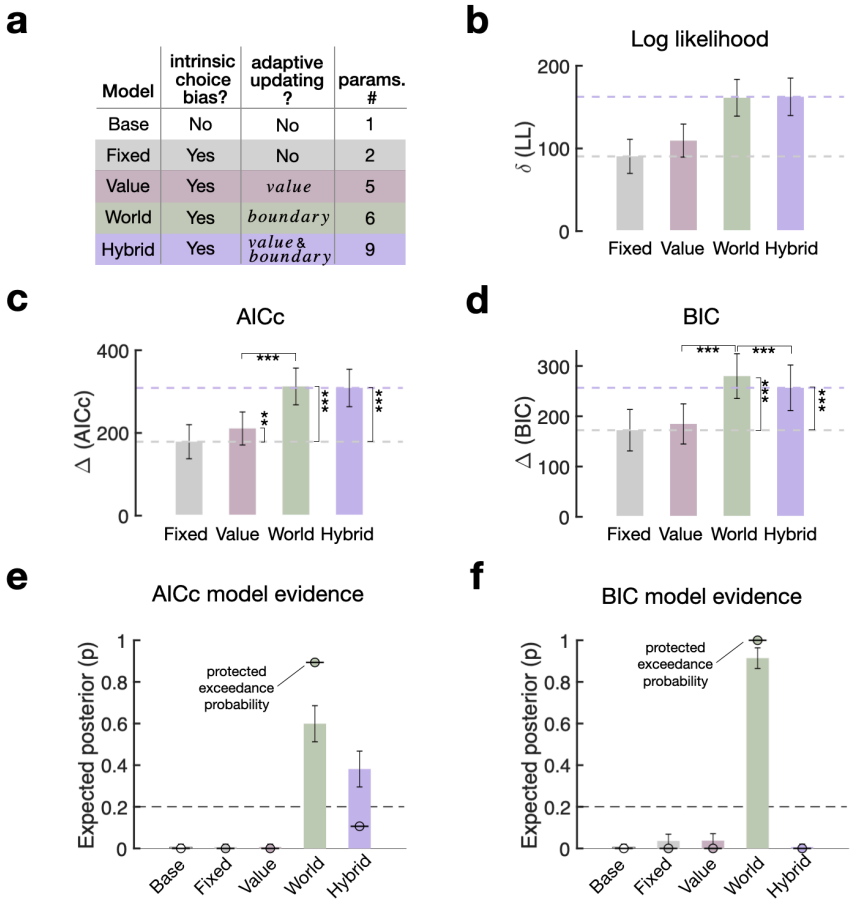
Fig. 5 | *Ex ante* simulation results for the PDM episodes with $incorrect$ feedback. The format is identical to that in Fig. 4.

In sum, the *ex ante* simulation confirmed that the direction 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 protocol, 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 two scenarios' distinct predictions via *ex ante* simulation, we evaluated their goodness of fit to human data. As points of reference for evaluation, we created three more models and included them in a model space. The 'Base' model sets the class boundary at its *true* value ($B = 0$) and does not update any choice values, so that it incorporated 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 class boundary to the data. The 'Hybrid' model incorporated both value-updating and world-updating algorithms. As a result, the model space consists of five models (Fig. 6a). In this model space, we quantified the models' ability to predict human classification choices using log likelihood (LL; Fig. 6b) and compared their abilities using two model-comparison criteria, the Akaike information criterion corrected for sample size (AICc[29]; Fig. 6c) and the Bayesian information criterion (BIC[30]; Fig. 6d).

In each comparison, the performance of the Fixed model relative to that of the Base model (gray dashed lines in Fig. 6b,c,d) captures the fraction of choice variability due to arbitrary choice preference, i.e., an individual's idiosyncratic tendency to prefer $small$ or $large$ choices. The performance of the Hybrid model relative to that of the Fixed model (the space between the gray and purple dashed lines in Fig. 6b,c,d) captures the (hypothetically) maximal fraction of choice variability that (i) cannot be ascribed to arbitrary choice preference but (ii) can be explained either by the value-updating model or by the world-updating model or by both. We found that the world-updating model was indistinguishable from the Hybrid model for AICc (paired *t*-test, $P=0.291$) and even superior to the Hybrid model for BIC (paired *t*-test, $P=1.49\times10^{-7}$). By contrast, the value-updating model was only slightly better than the Fixed model for AICc (paired *t*-test, $P=0.008$) and indistinguishable from the Fixed model for BIC (paired *t*-test, $P=0.28$). These comparisons, especially the BIC results, suggest that the value-updating algorithm only increased model complexity but hardly helped to explain human choice variability other than their arbitrary choice preferences.



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 (**b**-**f**). **b**,**c**,**d**, Model comparisons in goodness of fit in terms of log likelihood (**b**), AICc (**c**), and BIC (**d**). 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). The purple and gray dashed lines provide the reference points (see Main text for their meanings). Pair-wise model comparisons were performed using paired two-tailed *t*-tests (asterisks indicate significance: *, P < 0.05; **, P < 0.01; ***, P < 10^{-6}) **e**,**f**, Model comparisons in the hierarchical Bayesian Model Selection measures in terms of AICc (**e**) and BIC (**f**) values. Height of bars, the across-participant average expected posterior probabilities (N=30, mean±s.e.m). Dots marked with short dashes, protected exceedance probability. Dashed lines, chance level (p=0.2) probability that a model is more likely than the others in describing data due to chance.

To complement the above pair-wise comparisons, we took the hierarchical Bayesian Model Selection (hBMS) approach[31–33], to assess how probable it is that each of the five models prevails in the population (expected posterior probability; vertical bars in Fig. 6e,f) 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. 6e,f). Both hBMS measures corroborated the outcomes of the pair-wise comparisons: the world-updating model predominated in expected posterior probability (*P=0.5992* for AICc; *P=0.9141* for BIC) and protected exceedance probability (*P=0.8938* for AICc; *P=0.9999* for BIC).

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 *ex post* simulations were identical to the *ex ante* simulation except that the best-fit model parameters were used for each decision-maker (Supplementary Table 2-1; section: Computational models in 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 using the same format used in the *ex ante* simulation, i.e., the retrospective and prospective PSEs plotted against the stimuli, choices, and feedback states in the *toi* (Fig. 7a-c). We included the PDM episodes with non-veridical feedback (e.g., giving $correct$ feedback to $small$ choice to the ring size of '1'; indicated by symbols with dashed lines in Fig. 7a-c) though those episodes were rare (12.09±0.02%(mean±SEM) out of total *toi* episode trials; 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 human PSEs, in both pattern and magnitude (Fig. 7a,c), whereas those by the value-update model substantively deviated from human PSEs (Fig. 7a,b). The statistical comparison (paired *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 PSE significantly differed from the human PSEs (0 out of 60 pairs). Notably, most of the mismatches occurred because the value-updating model does not reverse the direction of feedback effects as sensory evidence grows while humans do so (compare the third columns of Fig. 7a and 7b).

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-f). 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,f) and computing their probability distributions across—again—the 20 *toi-1* and 20 *toi+1* episodes (the rows of Fig. 7e,f), respectively, we could create 400 joint-probability cells.

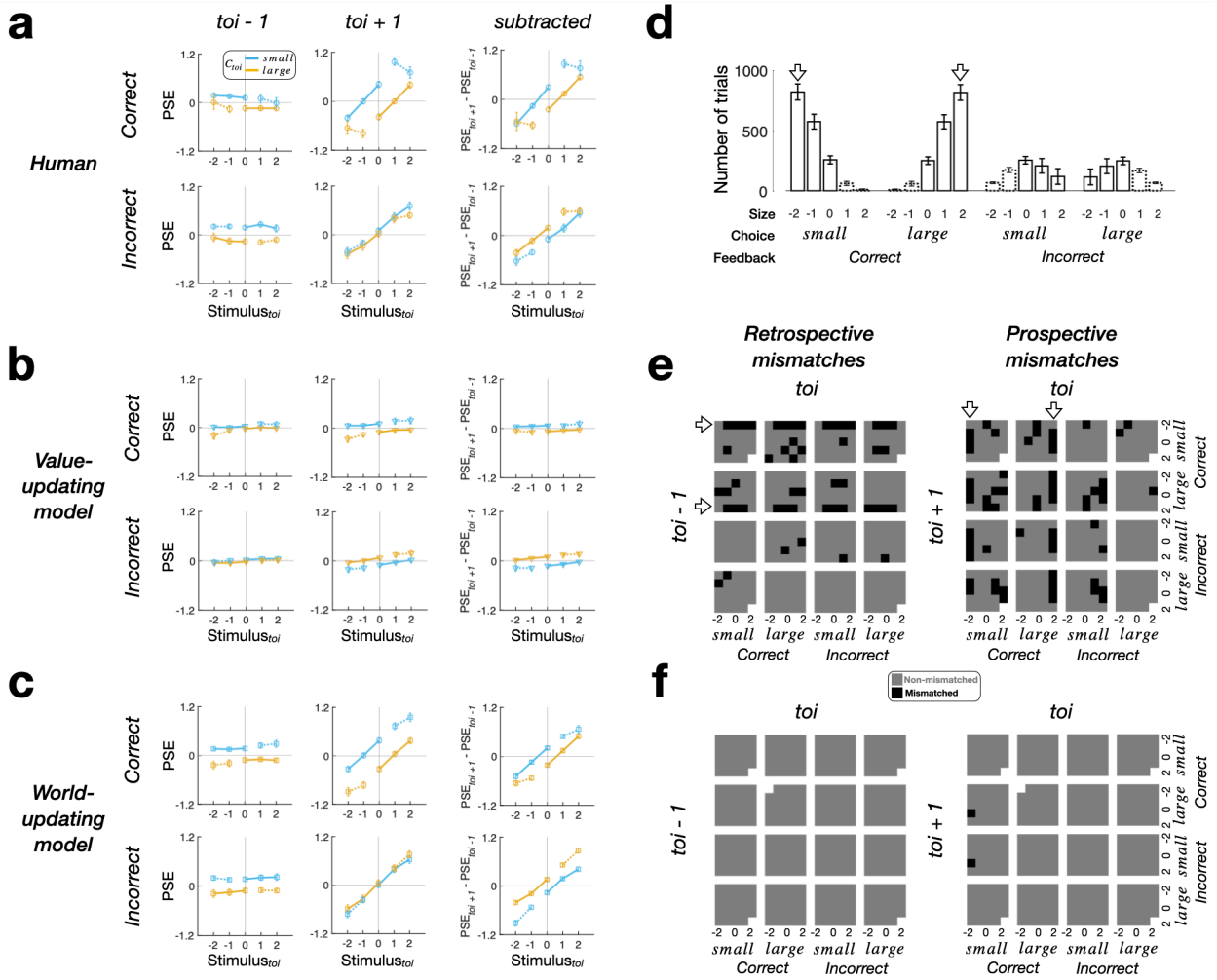


Fig. 7 | *Ex post* simulation results. a-c, Retrospective (left columns), prospective (middle columns), and subtractive (right columns) history effects in PSE for the human (**a**), value-updating (**b**), and world-updating (**c**) decision-makers. Top and bottom rows in each panel show the PSEs associated with the *toi* episodes involving $correct$ and $incorrect$ feedback. Symbols with error bars, mean±s.e.m across 30 decision-makers. **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 (Bonferroni corrected P< 0.05 when paired two-tailed *t*-tests performed). Empty cells, data points with NaN values due to insufficient observations.

We carried out the 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,f, 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 a total of 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 grows in strength.

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

Given the RL model's success in VDM and the presence of physical rewards, it is not surprising for the belief-based RL model to be considered as an account of the feedback effects in animals' PDM. The original work[9] supported this model using the six different datasets including one human dataset[34]. However, the current work indicates that the way humans learn from corrective feedback—without any physical or monetary reward—in PDM deviates from the value-updating scenario. The critical deviation occurred for the PDM episodes with strong sensory evidence: past $correct$ feedback should, albeit weakly, reinforce the choice made in the past according to the value-updating scenario whereas humans tended to switch choices following $correct$ feedback. In fact, the human dataset previously analyzed in the study[9] exhibits the same deviations (see their Figure 8c,d). When this dataset was analyzed in our way, it displayed the patterns of PSE almost identical to those of our dataset (Supplementary Fig. 4a). For that matter, another published human dataset[28] also substantially deviated from the value-updating scenario (Supplementary Fig. 4b). We remain cautious about the possibility that even animals can show such deviations. However, this possibility seems worth exploring though,

given that the main dataset from the 16 rats engaged in an olfactory PDM task also exhibits the patterns like those found in humans when corrected for the bias present in previous trials (see Figure 2i in the study[9]). The reversed feedback effects associated with strong sensory evidence cannot be ignored as a nuisance factor 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 by the same study[9] (shown in their Figure 3) as well. By contrast, BMBU offers a normative account of these effects by treating $correct$ and $incorrect$ feedback as 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[35,36] or implemented into various models[37–41]. However, BMBU goes beyond these efforts by offering a normative formalism of incorporating $correct$ and $incorrect$ feedback as the evidence for the class boundary such that it has an equal footing as sensory evidence in PDM tasks. This normative integration of feedback and sensory evidence within the framework of BDT substantially advances the current computational account of the history effects because it allows for addressing the history factors in the complete dimensions of PDM ('stimulus', 'choice', and 'feedback'), which is important given the multiplexed nature of history effects. For that matter, our modeling work joins recent computational and empirical efforts of incorporating feedback in the normative model of evidence accumulation for a random-dot motion task, another classic PDM task[6,42]. Together with these works, our work supports a general view that human decision-makers use corrective feedback as the 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.

It is worth discussing the so-called "Anna Karenina" account of the seemingly idiosyncratic $incorrect$ feedback effects, which states that "all correct trials are alike; each incorrect trial is incorrect in its own way" paraphrased earlier[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 clearly predicts the specific pattern of feedback effects for incorrect trials as we shown via *ex ante* simulation, endorsing the Anna-Karenina account means 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 offers a

principled account of the effects of feedback regardless of whether 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 highly dynamic and multiplexed in nature. This calls for an effort of establishing a methodologically rigorous framework to probe behavioral data for measuring the history effects. A few recent studies made such efforts. First, a group of researchers[6,9] raised a concern about the conventional practice of inspecting the history effects only in the future direction (prospective history effects in our term) because it does not rule out the possibility that the observed effects are caused by non-systematic slow fluctuations in the decision-making strategy. This group dealt with this concern by subtracting the retrospective history effects from the prospective ones. Another group[43] shared this concern but disagreed about its remedy with the first group by showing that the subtraction method cannot fairly recover diverse systematic updating strategies. As an alternative, they addressed the problem by taking a model-based approach to separate any given updating strategy from random drifts in decision criteria. Although we also share the two groups' concerns and welcome their efforts, we emphasize that BMBU successfully reproduced human history effects in both directions of time without incorporating any unsystematic 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 the almost entire space of PDM episode pairs (Fig. 7f). This suggests that it is an empirical matter whether the decision criterion slowly drifts or not, raising another concern that one might explain away systematic history effects with the slow drifts if the slow drifts do not exist at all or are negligible in magnitude. In this regard, we propose that researchers better treat the retrospective history effects not as a baseline or control condition but as what must be explained by any viable account of history effects, 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[44], and that our approach showcases its one rigorous example.

Methods

Participants

All participants (13 females, 17 males; aged 18-30 years) were recruited from the Seoul National University (SNU) community, gave prior written informed consent in accordance with the guidelines and approval of the Institutional Review Board at Seoul National University, 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, the contrast of which flickered at 20 Hz on a gray luminance background. On each trial, a fixation first appeared for 0.5 s in 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. A total of 5 different ring sizes (radii of 3.84, 3.92, 4.00, 4.08, 4.16 deg in visual angle; denoted by -2,-1,0,+1,+2 in the main text) was randomized within every block of 5 trials.

Task. Participants performed a one-interval two-alternative forced-choice (1I2AFC) perceptual decision-making task. Each participant participated in 5 daily sessions, each of which included 6 runs, each of which consisted of 170 trials, performing a total of 5,100 trials. On each trial, participants made a response during a time window of 1.2 s after stimulus onset by pressing one of the two keys using their index and middle fingers. The key assignment of $small$ and $large$ decisions 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 responses had not been made within the response period, the fixation mark turned into yellow, reminding participants that responses need to be made in time. These late-response trials were 0.5418% of the total trials across participants, and were included in data analysis. Meanwhile, trials where choices were omitted made up 0.0948% of the total trials. For these trials, we excluded them from all analyses and model fitting, resulting in a range of valid trials per participant from [5073, 5100] with an average of 5095.2$\pm$5.81 trials (mean$\pm$ SD). Before each run, we presented participants with the ring stimulus of the median size (radius, 4.00 deg.) 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 ring stimulus is smaller or larger than this reference ring. This procedure was introduced with the purpose of minimizing any possible carryovers of a belief they formed in a previous session of the boundary that splits

stimuli into two different classes. Participants were encouraged to maximize the fraction of correct trials.

Feedback Manipulation. Participants received stochastic feedback. Feedback was stochastically generated by comparing the trial stimulus to a value sampled from a normal distribution, $\mathcal{N}(\mu_{True}, \sigma_{True})$. We set σ_{True} to 1.28 while varying μ_{True} from -0.4 (10 runs), 0 (10 runs), to +0.4 (10 runs), which mimics the real-world situation where references are often unreliable and could be biased. Specifically, μ_{True} was held at 0 for initial 40-50 trials for every run, such that the zero mean was either biased ($\mu_{True} = \pm 0.4$) or stayed unbiased for the rest of the trials.

Data analysis

For analyzing sequential behavior conditional to a given event at trial of interest (*toi*), we calculated the proportion of choice $large$ after the event (*toi+1*) or before the event (*toi-1*) as a function of sensory stimuli x at *toi*.

Psychometric functions (per each event) were estimated (per participant) by fitting cumulative Gaussian distributions with *Psignifit* package[45–47] (https://github.com/wichmann-lab/psignifit) with the following function:

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

where $F(x)$ is a cumulative Gaussian distribution (with mean μ and standard deviation σ). No guess/lapse rates were assumed in quantifying PSE from each event-conditioned psychometric curve in order to constrain any peculiar approximations because some types of events rarely occurred at individual level. The mean μ reflects the point of subjective equality (PSE; the stimulus level with equal probability for a $small$ or $large$ choice), the feature of sequential behavior we are focused on throughout the entire paper.

Value-updating model

In the scenario where values of two options get updated after feedback experience with the degree of stimulus ambiguity, we followed the existing model of stimulus-dependent feedback effects first spotlighted by the previous work[9,10]. It assumed reinforcement learning extended to incorporate the conventional Bayesian understanding of perceptual ambiguity (Fig.1a, Fig.3e), also known as a partially observable Markov decision process. An agent's internal representation of a stimulus S is assumed given a noisy sensory measurement m in the Bayesian view. Since we are dealing with a 2AFC task, it is equal to the probability that the stimulus belongs to '$large$' or class '$small$' denoted by p_L and p_S (hereinafter the p-computation)

where $p_L = \int_0^\infty p(S|m)dS$ and $p_S = 1 - p_L$ referred to as a 'belief-state' in the previous work. Multiplying the p_S and p_L with the learned values of the options of $small$ and $large$, V_S and V_L result in the expected values of the two choices Q_S and Q_L, respectively. The decision rule can either be deterministic using an argmax rule as in the previous work (the higher Q determines the choice C) or be stochastic using a softmax rule, which selects $large$ with probability $\frac{\exp(\beta Q_L + b)}{\exp(\beta Q_S) + \exp(\beta Q_L + b)}$ (the higher Q preferentially selects d) where β is an inverse temperature, controlling the degree of stochasticity, and b represents an agent's inherent action bias. We used the latter version (the value-updating model) to give more room for value-updating scenario to produce behavior as flexible as possible, rather than limiting its behavioral repertoires. This value-updating model with softmax rule was used to represent the value-updating scenario when all the models considered in the study were evaluated (see **Model comparison**).

The predicted value Q_C is defined depending on C:

$$Q_C = \begin{cases} Q_S \; if \; C = small \\ Q_L \; if \; C = large \end{cases}$$

After decision-making, the value of the chosen option gets updated after receiving a feedback by the prediction error δ with learning rate a. No temporal discounting is assumed for simplicity.

$$V_C \leftarrow V_C + a\delta$$

With feedback serving as a reward (*correct*: r =+1, *incorrect*: r =0) and the stimulus exerts its effects through p_C as a modulator on δ, δ is computed:

$$\delta = r - p_C V_C$$

In case of correct feedback, a positive $\boldsymbol{\delta}$ decreases with the strength of the experienced stimulus (that is, being remote from the boundary of the stimulus distribution) which results from higher p_C. In case of incorrect feedback, a negative $\boldsymbol{\delta}$ becomes weak with the strength of the experienced stimulus which results from lower p_C.The model parameters are θ ={$b, \sigma_m, \alpha, \beta, V_{init}$}. Initial values of both choice options were the same as a free parameter.

World-updating model

Within the same picture set out by the existing framework, the model of world-updating shares the core decision process with the value-updating model as follows. First, it has the same process of how the agent's internal representation of a stimulus is modeled except for the fact that the finite limit of the integral to compute p_S and p_L is not fixed to a zero constant anymore, but could vary on a trial-to-trial basis. We assume it goes through the process of Bayesian

inference to figure out a statistic of the stimulus distribution in a physical world. We write $p_L = \int_{\hat{B}}^{\infty} p(S|m)dS$ where $\hat{B}$ is a currently held belief of the world state as a trial estimate of the boundary of splitting stimuli into two classes. Second, it has the same process of how the basic choice computations are modeled except for the fact that values have no role (virtually equivalent to V_S = V_L=1 from the value-updating perspective). Based on the statistical structure of a binary classification task, we built an observer model from a generative model where causal relationships are specified with the solid arrows in the diagram (Fig.1a, Fig.3d). For the task, the observer needs (1) to learn (or estimate) the true boundary μ_{True} based on evidence given by past stimuli and past feedback, (2) to infer a presented stimulus in a given trial, and (3) finally to commit to a decision by comparing the inferred stimulus with the learned boundary. The decision $large$ is determined if $p_L > 0.5$ (the higher p determines the choice). After decision-making, the belief of the world-state (the estimate of the boundary) gets updated after receiving a feedback. Given a noisy sensory measurement in memory m'_t and feedback information F_t (Note that F used throughout the section **Methods** for simplicity is equivalent to CL in the main text, which stands for the class variable with the two states, $small$ and $large$. To be specific, CL is jointly determined by choice and correctness), the inference is defined as the posterior probability distribution $p(B_t|m'_t, F_t) \sim N(\hat{B}_t^{post}, \sigma^2{}_t^{post})$ according to Bayes rule:

$$p(B_t|m'_t, F_t) \propto p(m'_t, F_t|B_t)p(B_t) \quad .$$

where $p(B_t)$ is the prior, the updated belief from experiencing the previous trial.

The likelihood $p(m'_t, F_t|B_t)$ can be expressed using the Bayesian network from the causal graph.

$$\begin{aligned} p(m'_t, F_t|B_t) &= \int p(m'_t, F_t, S_t|B_t)dS_t \\ &= \int [\int p(m'_t|m_t)p(m_t|S_t)dm_t]p(S_t|B_t)p(F_t|S_t, B_t)dS_t \\ &= \int N\left(m'_t, \sigma_{m'}^2 + \sigma_m^2\right)N(S_t, \sigma_S^2)p(F_t|S_t, B_t)dS_t \end{aligned}$$

The integral above can be decomposed into the two regimes that differ in the inequality between S_t and B_t.

$$\int_{S_t=-\infty}^{S_t<B_t} N\left(m'_t, \sigma_{m'}^2 + \sigma_m^2\right)N(S_t, \sigma_S^2)p(F_t|S_t, B_t)dS_t$$
$$+ \int_{S_t>B_t}^{S_t=+\infty} N\left(m'_t, \sigma_{m'}^2 + \sigma_m^2\right)N(S_t, \sigma_S^2)p(F_t|S_t, B_t)dS_t$$

This decomposition allows the likelihood term that involves the feedback to keep only one of the two decomposed integrals over S_t depending on feedback $small$ or $large$.

When $F_t = small$, the $p(F_t|S_t, B_t)$ is always 1 for the regime $S_t < B_t$ if one simply takes F_t as the given inequality between S_t and B_t, and 0 for the regime $S_t > B_t$. Therefore, the first decomposed integral does remain as below.

$$= \int_{S_t=-\infty}^{S_t<B_t} N\left(m_t', \sigma_{m'}^2 + \sigma_m^2\right) N(S_t, \sigma_S^2) dS_t \propto \int_{S_t=-\infty}^{S_t<B_t} e^{-\frac{\left(S_t - \frac{B_t\sigma_M^2 + m_t'\sigma_S^2}{\sigma_M^2+\sigma_S^2}\right)^2}{\frac{\sigma_M^2\sigma_S^2}{2(\sigma_M^2+\sigma_S^2)}}} dS_t \times e^{-\frac{(B_t - m_t')^2}{2(\sigma_M^2+\sigma_S^2)}}$$

where $\sigma_{m'}^2 + \sigma_m^2$ is replaced by σ_M^2 for simplicity.

When $F_t = large$,

$$= \int_{S_t>B_t}^{S_t=+\infty} N\left(m_t', \sigma_{m'}^2 + \sigma_m^2\right) N(S_t, \sigma_S^2) dS_t \propto \int_{S_t>Z_t}^{S_t=+\infty} e^{-\frac{\left(S_t - \frac{B_t\sigma_M^2 + m_t'\sigma_S^2}{\sigma_M^2+\sigma_S^2}\right)^2}{\frac{\sigma_M^2\sigma_S^2}{2(\sigma_M^2+\sigma_S^2)}}} dS_t \times e^{-\frac{(B_t - m_t')^2}{2(\sigma_M^2+\sigma_S^2)}}$$

To sum, the likelihood term boils down to a form of the product of a cumulative distribution function of $N\left(\frac{B_t\sigma_M^2 + m_t'\sigma_S^2}{\sigma_M^2+\sigma_S^2}, \frac{\sigma_M^2\sigma_S^2}{4(\sigma_M^2+\sigma_S^2)}\right)$ and a probability density function of $N(m_t', \sigma_M^2 + \sigma_S^2)$. Analogous to inherent action bias in the value-updating scenario, we assumed the updated distribution $\hat{B}_t^{post}$ inclined to a long-term prior distribution $N(\mu_0, \sigma_0^2)$ for each individual. This decay occurs while going through an inter-trial interval until confronting next decision-making. Upon decision-making at trial $t+1$, the agent recalls the belief about class boundary as prior $p(B_{t+1}) = N(\hat{B}_{t+1}, \sigma_{B_{t+1}}^2)$ where:

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

$$\sigma_{B_{t+1}}^2 = \lambda \sigma_t^{2\,post} + \sigma_{diffusion}^2$$

A decay parameter $\lambda = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_t^{2\,post}}$ influences the belief distribution in terms of its width as well as final location. Another additive term $\sigma_{diffusion} > 0$ helps to maintain the width of the distribution over multiple trials, thus avoiding sharpening and stopping the updating process. The model parameters are $\theta = \{\mu_0, \sigma_m, \sigma_s, \sigma_0, \sigma_{m'}, \sigma_{diffusion}\}$. Similarly, the world-updating model with these accessory assumptions that could reflect the potential idiosyncrasy in individuals was evaluated for fair comparison between models (see **Model comparison**).

Baseline models

For the sake of setting up the baselines for the alternative model we opted to test, we tested models still not requiring the extra assumption of values in modeling PDM. To do so, we also minimized accessory assumptions in the model 'World' we used to capture diverse behavioral pattern in individual data. We (1) assumed a fixed unbiased boundary perception model without any RL mechanism ('Base'; θ ={σ_m}), and (2) replaced the finite limit of the integral to compute p_S and p_L to be a constant value μ_0, representing inherent bias in the belief ('Fixed': θ={μ_0, σ_m}).

Model fitting

For each participant, we fitted different model variants to choices over valid N trials (N≤170) for each experimental run (M=10), for each condition (K=3). The likelihood of a set of parameters θ for a particular model is the probability of the data given θ $p(data|\theta, model)$. For a given model, we define the log-likelihood as:

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

where C_i denotes the participant's response ($large$ or $small$) in the i-th trial and $p(C_i|\theta, model) = pL_i$ for $C_i = large$ and $p(C_i|\theta, model) = 1 - pL_i$ for $C_i = small$. We estimated the log-likelihood using inverse binomial sampling (IBS[48]), an efficient way of generating unbiased estimates via simulation. We obtained maximum-likelihood estimates of the model parameters using Bayesian Adaptive Direct Search (BADS[49]) method. It is 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 finding local optima, we performed 20 independent restarts of the fitting procedure via randomly choosing the starting positions using Latin hypercube sampling (*lhsdesign_modifed.m* by Nassim Khlaled, https://www.mathworks.com/matlabcentral/fileexchange/45793-latin-hypercube), then picked the highest log likelihood (i.e. lowest -LL) value for each model. We excluded trials where choices were not made within the response window. To avoid infinite loops from using IBS, we implemented individual lapse rates not arbitrarily, instead by taking the average of lapse rate and guess rate from a cumulative Gaussian fitting to each individual's grand mean psychometric curve (total valid trials used, not conditional to any episode). With individual lapse probability (mean rate of 0.05, which ranged from $[0.0051, 0.1714]$), trials were randomly selected as lapse trials where random responses were generated.

Model comparison

To evaluate model performance, we carried out comparison in two ways (1) predictive power, and (2) generative power. For predictive performance, we firstly compared the goodness of fit of the model agents using corrected Akaike information criterion (AICc) and Bayesian Information Criteria (BIC) based on maximum-likelihood estimation fitting.

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

$$BIC = -2 \cdot \mathrm{LL}(\theta^*) + p \cdot log(N \text{ x } M \text{ x } K)$$

where p is the number of parameters of the model and the total number of trials in the dataset is $N \text{ x } M \text{ x } K$. Log model evidence was obtained[33] for each participant by multiplying AICc or BIC by -1/2. Furthermore, we take a hierarchical Bayesian Model Selection approach (BMS; *spm_BMS.m* in SPM12, http://www.fil.ion.ucl.ac.uk/spm/) 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 report protected exceedance probabilities for each model and Bayesian omnibus risk (BOR) (see Fig. 6e,f). To evaluate generative performance, we ran *ex-post* simulations with their individually-recovered best-fitting parameter set (see ***Ex post* model simulations** for details).

Ex ante model simulations

Model simulations were conducted using actual sequence runs that were exposed to the participants during actual experiments. The purpose of model simulations *ex ante* was to compare the world-updating and the value-updating model in terms of how they could be distinguished from each other's behavior in general when the sole parameter σ_m commonly shared across the models was assumed only to change (25 levels, uniformly-spaced $[0.15, 3.27]$). We repeated 100 times of the whole dataset (100 x N x M x K = 507,300~510,000 trials per participant). For *simulations ex ante* where the pure model behavior was supposed to be evaluated, we assumed all trials to be performed with no lapse and assumed an individual's intrinsic bias to be zero throughout trials. Model parameters used to generate Fig.4 and Fig.5 are summarized in the Supplementary Table 1. The other parameters specific to each model were selected to achieve human-level task performances (see Supplementary Fig.1 for details and statistical results of not only unpaired two-tailed *t*-tests which compared performances of humans versus each model's simulations generated using 25 different sets of parameters, but also the test which compared simulations made by the world-updating model versus the value-updating model).

Ex post model simulations

For simulations *ex post* (after recovering individual model parameter sets from **Model fitting**), the same principle was applied as described in ***Ex ante* simulations** except that we used the parameters recovered from model fitting. We also allowed lapse trials as done for model fitting.

Statistical tests

Unless otherwise mentioned, the statistical comparisons were performed using paired *t*-tests (two-tailed, N=30). To test the reversed feedback effects at extreme stimuli conditions, we applied one-sample *t*-tests (one-tailed, N=27 for Supplementary Fig.4a, N=8 for Supplementary Fig.4b). Repeated *t*-tests on PSEs between data and model (Fig.7) were performed and p-values (Supplementary Table 3) were reported after Bonferroni correction (two-tailed, N=30; *: $P<0.00083$, **: $P<0.000167$, ***: $P<0.0000167$). Same tests using Wilcoxon signed rank tests were also provided after correction. Repeated *t*-tests on each cell of episode frequency maps between data and model were performed and p-values (Fig.7e,f) were reported after 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, Supplementary Fig.1).

References

1. Gold, J. I., Law, C.-T., Connolly, P. & Bennur, S. The relative influences of priors and sensory evidence on an oculomotor decision variable during perceptual learning. *J Neurophysiol* 100, 2653–2668 (2008).

2. Hwang, E. J., Dahlen, J. E., Mukundan, M. & Komiyama, T. History-based action selection bias in posterior parietal cortex. *Nat Commun* 8, 1242 (2017).

3. Abrahamyan, A., Silva, L. L., Dakin, S. C., Carandini, M. & Gardner, J. L. Adaptable history biases in human perceptual decisions. *Proc National Acad Sci* 113, E3548–E3557 (2016).

4. Busse, L. *et al.* The detection of visual contrast in the behaving mouse. *J Neurosci* 31, 11351–11361 (2011).

5. Scott, B. B., Constantinople, C. M., Erlich, J. C., Tank, D. W. & Brody, C. D. Sources of noise during accumulation of evidence in unrestrained and voluntarily head-restrained rats. *eLife* 4, e11308 (2015).

6. Mendonça, A. G. *et al.* The impact of learning on perceptual decisions and its implication for speed-accuracy tradeoffs. *Nat Commun* 11, 2757 (2020).

7. Fernberger, S. W. Interdependence of judgments within the series for the method of constant stimuli. *J Exp Psychol* 3, 126–150 (1920).

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* 27, 821–832 (2017).

9. Lak, A. *et al.* Reinforcement biases subsequent perceptual decisions when confidence is low, a widespread behavioral phenomenon. *eLife* 9, e49834 (2020).

10. Lak, A. *et al.* Dopaminergic and prefrontal basis of learning from sensory confidence and reward value. *Neuron* 105, 700-711.e6 (2020).

11. Lee, D., Seo, H. & Jung, M. W. Neural basis of reinforcement learning and decision making. *Annu Rev Neurosci* 35, 287–308 (2012).

12. Daw, N. D. & Doya, K. The computational neurobiology of learning and reward. *Curr Opin Neurobiol* 16, 199–204 (2006).

13. Sutton, R. & Barto, A. *Reinforcement Learning: An Introduction*. (MIT Press, 1998).

14. Corrado, G. S., Sugrue, L. P., Seung, H. S. & Newsome, W. T. Linear-nonlinear-poisson models of primate choice dynamics. *J Exp Anal Behav* 84, 581–617 (2005).

15. Lau, B. & Glimcher, P. W. Dynamic response-by-response models of matching behavior in rhesus monkeys. *J Exp Anal Behav* 84, 555–579 (2005).

16. Odoemene, O., Pisupati, S., Nguyen, H. & Churchland, A. K. Visual evidence accumulation guides decision-making in unrestrained mice. *J Neurosci* 38, 10143–10155 (2018).

17. Hermoso-Mendizabal, A. *et al.* Response outcomes gate the impact of expectations on perceptual decisions. *Nat Commun* 11, 1057 (2020).

18. Akrami, A., Kopec, C. D., Diamond, M. E. & Brody, C. D. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature* 554, 368–372 (2018).

19. Summerfield, C. & Tsetsos, K. Building bridges between perceptual and economic decision-making: neural and computational mechanisms. *Front Neurosci* 6, 70 (2012).

20. Kahnt, T., Grueschow, M., Speck, O. & Haynes, J.-D. Perceptual learning and decision-making in human medial frontal cortex. *Neuron* 70, 549–559 (2011).

21. Polanía, R., Krajbich, I., Grueschow, M. & Ruff, C. C. Neural oscillations and synchronization differentially support evidence accumulation in perceptual and value-based decision making. *Neuron* 82, 709–720 (2014).

22. Körding, K. P. & Wolpert, D. M. Bayesian decision theory in sensorimotor control. *Trends Cogn Sci* 10, 319–326 (2006).

23. Trommershäuser, J., Maloney, L. T. & Landy, M. S. Decision making, movement planning and statistical decision theory. *Trends Cogn Sci* 12, 291–297 (2008).

24. Griffiths, T. L. & Tenenbaum, J. B. Optimal predictions in everyday cognition. *Psychol Sci* 17, 767–73 (2006).

25. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285 (2011).

26. Glaze, C. M., Filipowicz, A. L. S., Kable, J. W., Balasubramanian, V. & Gold, J. I. A bias–variance trade-off governs individual differences in on-line learning in an unpredictable environment. *Nat Hum Behav* 2, 213–224 (2018).

27. Ma, W. J. Bayesian decision models: A primer. *Neuron* 104, 164–175 (2019).

28. Hachen, I., Reinartz, S., Brasselet, R., Stroligo, A. & Diamond, M. E. Dynamics of history-dependent perceptual judgment. *Nat Commun* 12, 6036 (2021).

29. Burnham & Anderson. *Model selection and inference, A practical information-heuristic approach*. (Springer-Verlag, 2002).

30. Schwarz, G. Estimating the dimension of a model. *Ann Statistics* 6(2), 461–464 (1978).

31. Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *Neuroimage* 46, 1004–1017 (2009).

32. Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. Bayesian model selection for group studies — Revisited. *Neuroimage* 84, 971–985 (2014).

33. Cao, Y., Summerfield, C., Park, H., Giordano, B. L. & Kayser, C. Causal inference in the multisensory brain. *Neuron* 102, 1076-1087.e8 (2019).

34. Urai, A. E., Braun, A. & Donner, T. H. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nat Commun* 8, 14637 (2017).

35. Zariwala, H. A., Kepecs, A., Uchida, N., Hirokawa, J. & Mainen, Z. F. The limits of deliberation in a perceptual decision task. *Neuron* 78, 339–351 (2013).

36. Renart, A. & Machens, C. K. Variability in neural activity and behavior. *Curr Opin Neurobiol* 25, 211–220 (2014).

37. Qamar, A. T. *et al.* Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proc National Acad Sci* 110, 20332–20337 (2013).

38. Summerfield, C., Behrens, T. E. & Koechlin, E. Perceptual classification in a rapidly changing environment. *Neuron* 71, 725–736 (2011).

39. Norton, E. H., Fleming, S. M., Daw, N. D. & Landy, M. S. Suboptimal criterion learning in static and dynamic environments. *Plos Comput Biol* 13, e1005304 (2017).

40. Treisman, M. & Williams, T. C. A theory of criterion setting with an application to sequential dependencies. *Psychol Rev* 91, 68–111 (1984).

41. Lages, M. & Treisman, M. A criterion setting theory of discrimination learning that accounts for anisotropies and context effects. *Seeing Perceiving* 23, 401–434 (2010).

42. Drugowitsch, J., Mendonça, A. G., Mainen, Z. F. & Pouget, A. Learning optimal decisions with confidence. *Proc National Acad Sci* 116, 24872–24880 (2019).

43. Gupta, D. & Brody, C. D. Limitations of a proposed correction for slow drifts in decision criterion. *Arxiv* (2022) doi:10.48550/arxiv.2205.10912.

44. Carr, E. H. *What is History?* (University of Cambridge & Penguin Books, 1961).

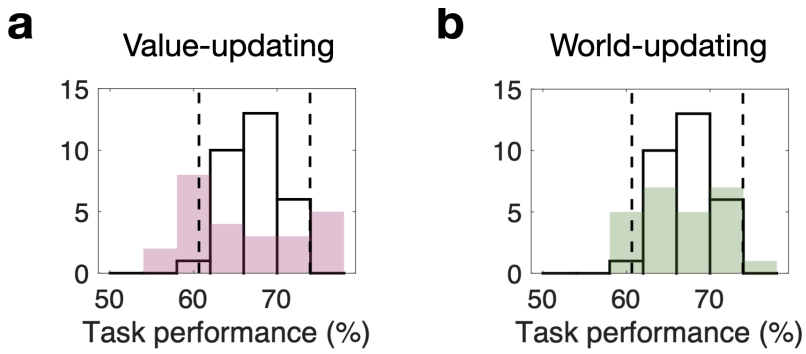
45. Wichmann, F. A. & Hill, N. J. The psychometric function: I. Fitting, sampling, and goodness of fit. *Percept Psychophys* 63, 1293–1313 (2001).

46. Wichmann, F. A. & Hill, N. J. The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Percept Psychophys* 63, 1314–1329 (2001).

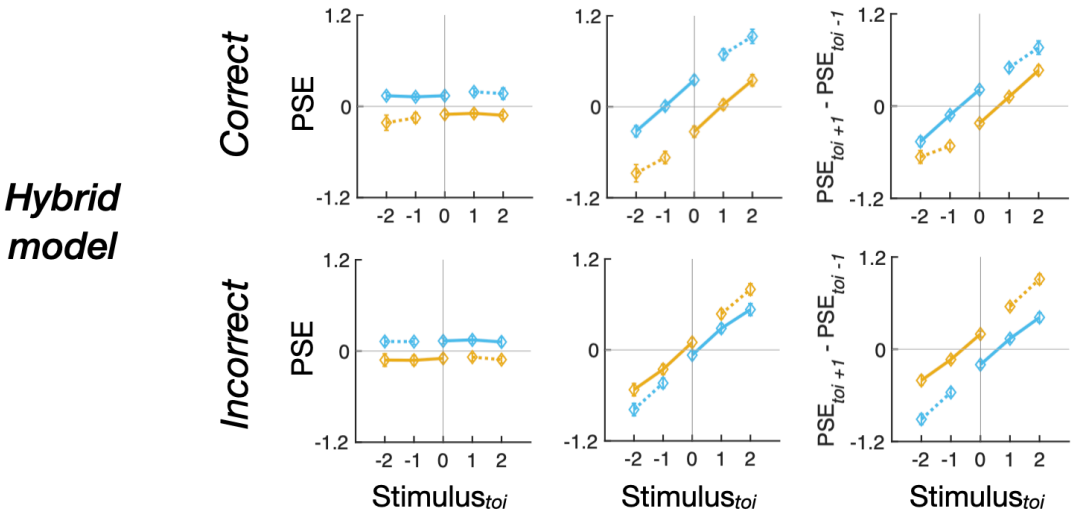
47. Schütt, H. H., Harmeling, S., Macke, J. H. & Wichmann, F. A. Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Res* 122, 105–123 (2016).

48. Opheusden, B. van, Acerbi, L. & Ma, W. J. Unbiased and efficient log-likelihood estimation with inverse binomial sampling. *Plos Comput Biol* 16, e1008483 (2020).

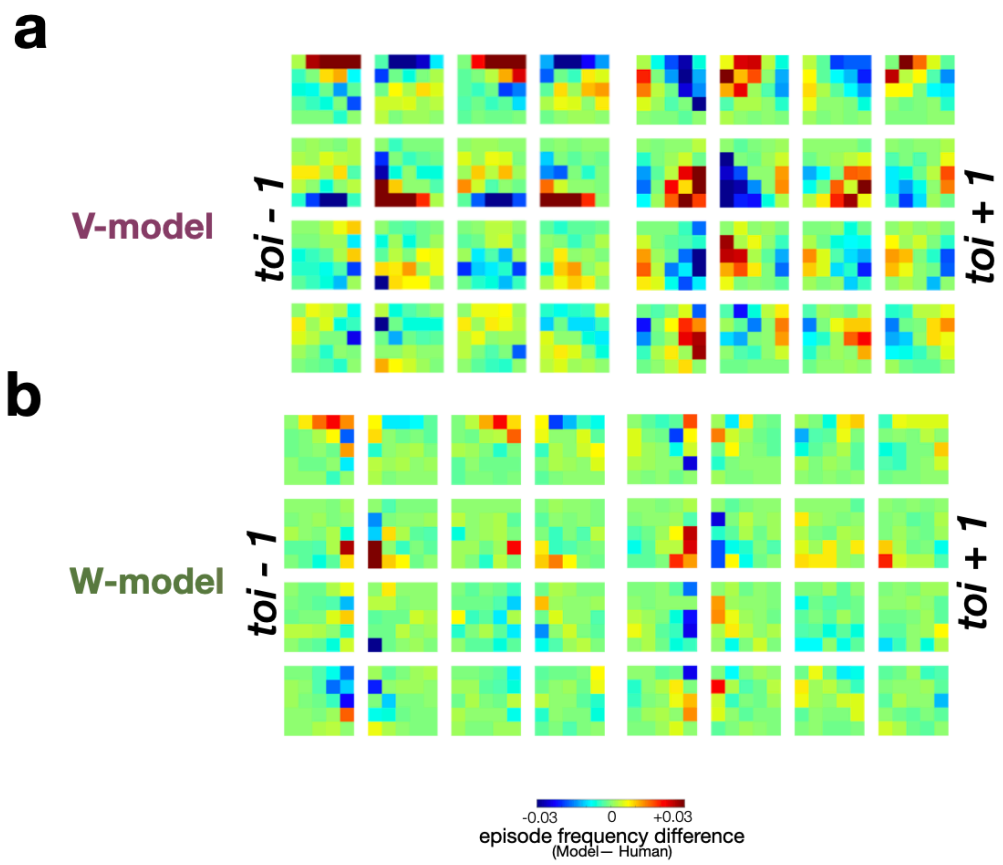
49. Acerbi, L. & Ma, W. J. Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. *Advances in Neural Information Processing Systems.* 1836–1846 (2017).



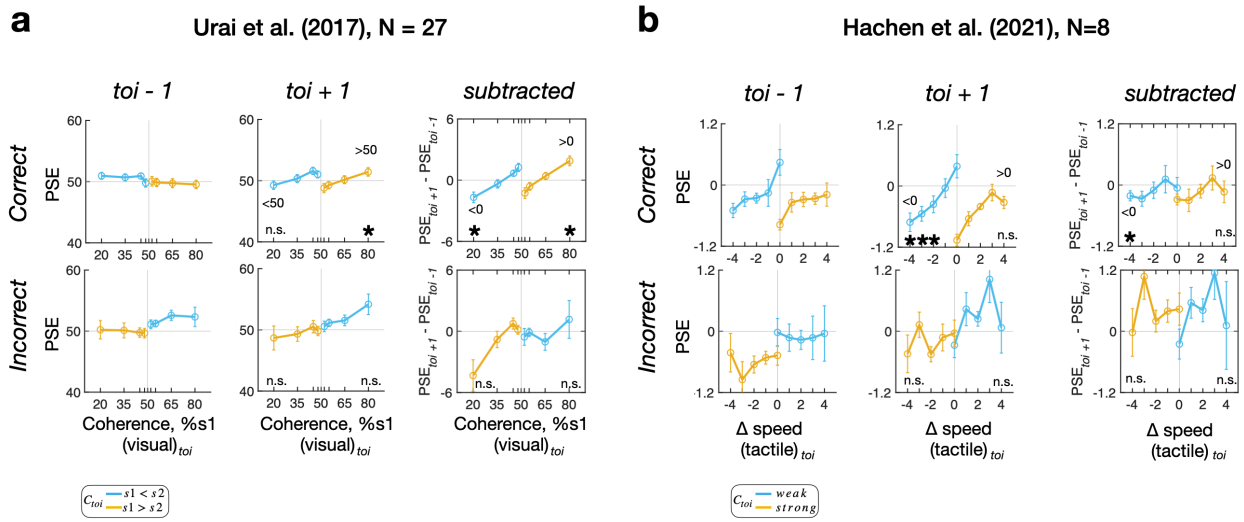
Supplementary Fig. 1. Parameters used in *ex ante* simulations were selected to achieve human-level task performance. a,b, Task performances under the range of model parameters that were used to generate those *ex ante* simulation results in Fig.4 and Fig.5. A model's task performance for each set of parameters is quantified as the percentage of trials where its simulated choices matched feedback had been assigned in the actual experiments. Histogram of human performances are superimposed (Dotted lines mark their [min, max]=[60.65%, 73.94%], Human average: 67.85%). **a,** Performances achieved by value-updating model were no different than observed performances (P=0.1549; Null hypothesis: model's performance vector and humans' performance vector come from populations with equal means, unpaired two-sided *t*-test). **b,** Performances achieved by world-updating model were no different than observed performances (P=0.3361, unpaired two-sided *t*-test), and were no different than model Value (P=0.5691, unpaired two-sided *t*-test; value-updating model average: 65.88%, world-updating model average: 66.80%).



Supplementary Fig. 2. Sequential PSE features reproduced by the Hybrid model's *ex post* simulations. The Hybrid model's sequential PSE features from simulations *ex post* when conditioned to correct feedback (top row) and incorrect feedback (bottom row) given at *toi*. Retrospective (left column), prospective (middle column), and subtracted PSE features (right column). Colors label choices (blue: $small$ and yellow: $large$).



Supplementary Fig. 3. Difference in episode frequency maps between data and model simulations *ex post*. a,b, Difference maps (data vs. model) of probabilities of occurrence of all possible events that happened at *toi-1* (or, *toi+1*) and *toi*. We call them episode frequency maps, each element of which represents a pair out of 20-by-20 combinations of episodes. Colors represent how much the model's episode frequency is deviated from the observed frequency. Statistical results are summarized in Fig.7e,f. **a,** Episode frequency difference map of value-updating model. **b,** Episode frequency difference map of world-updating model.



Supplementary Fig.4. Reversed feedback effects were also found in other datasets. a,b, Sequential features indicative of reversed feedback effects were also found in other previously published datasets. Reversal of the signs of feedback effects at extreme stimuli were observed (their statistical significance limited to correct cases), which was similar to the finding in the study. They were plotted in the format of Fig.7a, the same way that we analyzed human observers in our work (N=30, mean±s.e.m). **a,** Sequential features of human observers (N=27) analyzed in our way from human dataset that once had been published[1], then analyzed in the previous study[2]. They performed a two-interval forced choice motion coherence discrimination task (a first-interval reference stimulus of 70% coherence vs. a second-interval test stimulus). Choices were made to indicate the difference in coherence ('s1<s2' vs. 's1>s2') between two successively presented random dot kinematograms. 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]). For convenience, we merged trials of 0.625 and trials of 1.25 into a single stimulus bin (similarly, not only trials of [2.5, 5] merged into one but also those of [10, 20] merged into one). 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) followed Figure 8 of the previous study[2] 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* (*P=0.0272* for stimulus 80%) and on the panel *subtracted* (*P=0.0018 and 0.000318* for stimulus 20% and 80%). **b,** Sequential features of human observers (N=8) published in another previous study[3]. They performed a one-interval forced choice vibrotactile discrimination task. Choices were made to indicate the speed of the presented vibration ('low-speed(weak)' vs. 'high-speed(strong)'). Note that the nine-level stimuli of [-4,-3,-2,-1,0,1,2,3,4] on the x-axis followed the original coding in the dataset[3]. 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* (*P=0.004, 0.0048, and 0.04* for stimulus -4, -3, and -2) and on the panel *subtracted* (*P=0.034* for stimulus -4).

Supplementary Table 1. 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	b	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		

Supplementary Table 2-1. Parameters recovered from fitting of the main models, world-updating model and value-updating model, 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)	b	0.02 (0.43)
σ_0	4.27 (1.22)	α	0.17 (0.20)
σ_s	0.84 (0.92)	β	7.43 (7.01)
$\sigma_{m'}$	3.70 (1.00)	V_{init}	0.92 (0.16)
$\sigma_{diffusion}$	3.20 (1.57)		

Supplementary Table 2-2. Parameters recovered from fitting of 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	1.52 (0.48)	σ_m	1.90 (0.50)
μ_0	0.01 (0.61)	μ_0	0.06 (0.40)		
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)				

Supplementary Table 3. 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)				*ALL* (60 conditions)	
	n.s.		significance		n.s.		significance		n.s.		significance		n.s.	
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)

Paired *t*-test | Wilcoxon signed rank

* : p<0.05
** : p<0.01
*** : p<0.001
, Bonferroni corrected for 60 tests

References

1. Urai, A. E., Braun, A. & Donner, T. H. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nat Commun* 8, 14637 (2017).

2. Lak, A. *et al.* Reinforcement biases subsequent perceptual decisions when confidence is low, a widespread behavioral phenomenon. *eLife* 9, e49834 (2020).

3. Hachen, I., Reinartz, S., Brasselet, R., Stroligo, A. & Diamond, M. E. Dynamics of history-dependent perceptual judgment. *Nat Commun* 12, 6036 (2021).