

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

3 Short title: Corrective feedback on perceptual decisions

4 Hyang-Jung Lee<sup>1</sup>, Heeseung Lee<sup>1</sup>, Chae Young Lim<sup>2</sup>, Issac Rhim<sup>3</sup>, Sang-Hun Lee<sup>1\*</sup>

5

6

7 <sup>1</sup>Department of Brain and Cognitive Sciences, Seoul National University, Seoul, South Korea

8 <sup>2</sup>Department of Statistics, Seoul National University, Seoul, South Korea

9 <sup>3</sup>Institute of Neuroscience, University of Oregon, Eugene, OR, USA

10

11

12 \*Corresponding author

13 E-mail: [visionsl@snu.ac.kr](mailto:visionsl@snu.ac.kr)

14

15 **Abbreviations**

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

20 **Keywords**

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

## 23 **Abstract**

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

39

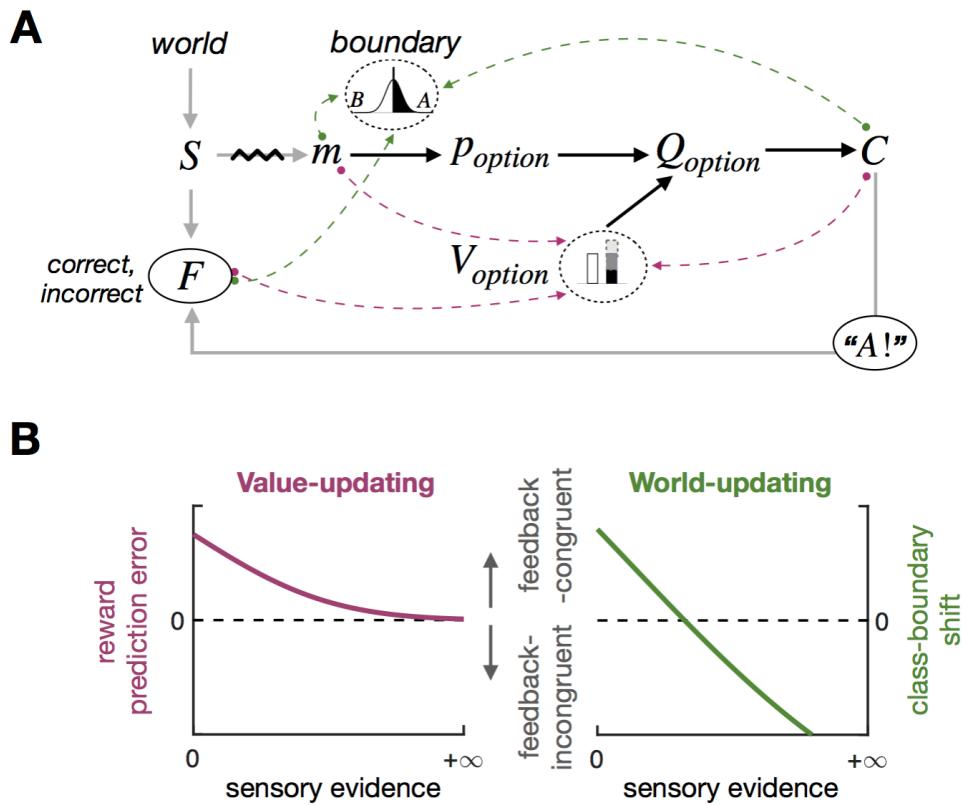
## 40 **Introduction**

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

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

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

71 prediction error compared to the latter (as illustrated in the left panel of Fig 1B). A recent study  
72 [9] referred to this sensory evidence-dependent impact of feedback as “confidence-guided  
73 choice updating” based on the tight linkage between decision confidence and sensory evidence.  
74 This RL-grafting account, referred to as the *value-updating scenario* hereinafter, appears natural  
75 given that corrective feedback is typically provided as physical rewards such as juice or water in  
76 animal PDM experiments [4,5,8–10,19–21]. The value-updating scenario seems plausible from  
77 the perspective that PDM and VDM might share common mechanisms [22], as suggested by  
78 some empirical studies [23,24].



79  
80

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

90 world-updating scenarios. According to the value-updating scenario (left), as sensory evidence becomes  
91 stronger,  $p_{option}$  increases, and accordingly, so does  $Q_{option}$ . As a result, reward prediction errors become  
92 smaller but remain in the direction congruent with feedback, which predicts that feedback effects on  
93 subsequent trials diminish asymptotically as a function of the strength of sensory evidence. According to  
94 the world-updating scenario (right), as sensory evidence becomes stronger, the stimulus distribution, and  
95 accordingly  $B$  too, becomes shifted farther towards the stimulus in the direction counteracting the  
96 influence of feedback. As a result, the direction of feedback effects is the same as that predicted by the  
97 value-updating scenario for weak sensory evidence but eventually reverses to the direction incongruent  
98 with feedback as sensory evidence becomes stronger.  
99

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

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

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

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

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

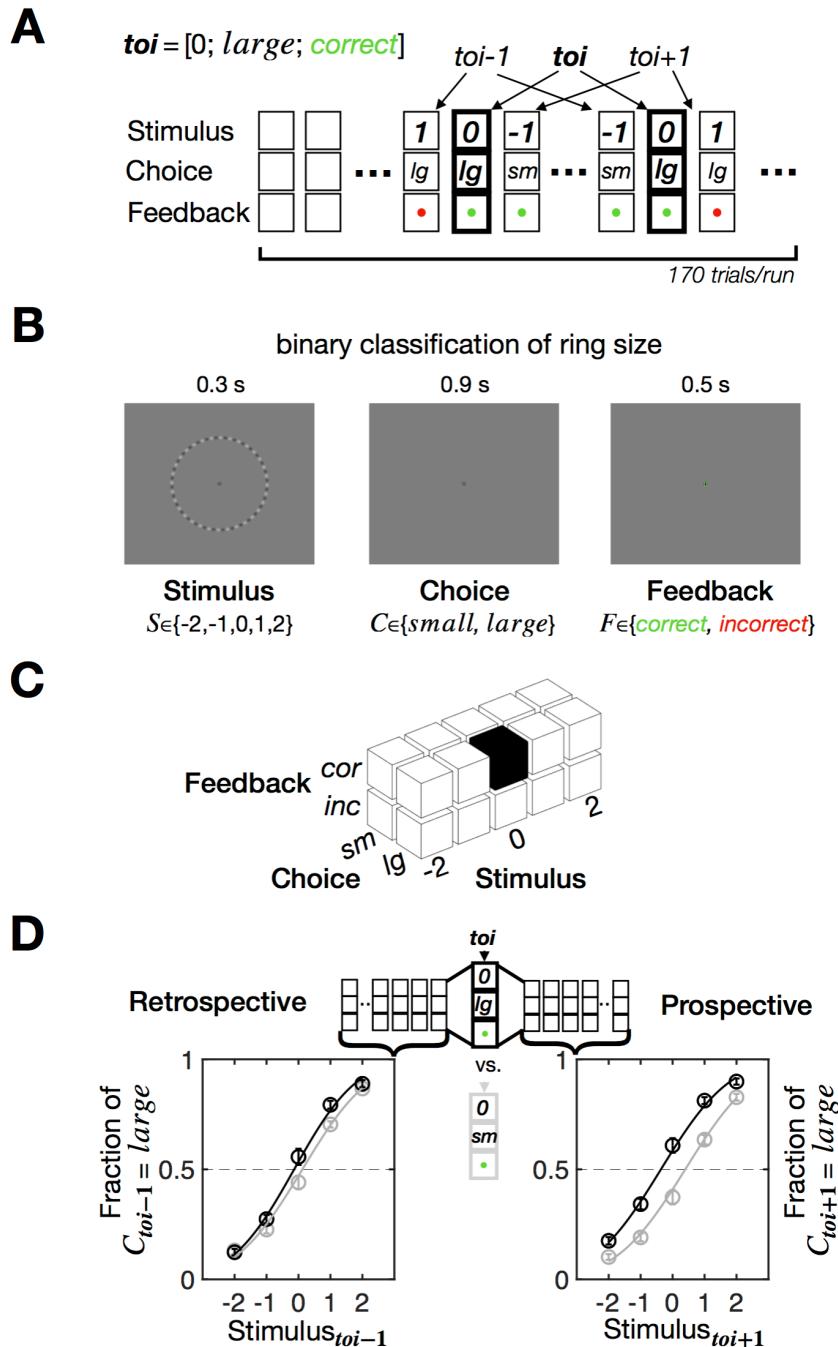
149 We found the world-updating model superior to the value-updating model in explaining  
150 human history effects of corrective feedback on PDM. Critically, the value-updating model fails  
151 to account for the observed stimulus-dependent feedback effects. Our findings suggest that  
152 humans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

## 156 **Results**

### 157 **Quantifying the retrospective and prospective history effects 158 of feedback on binary classification**

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

168 was occasionally given to incorrect and correct choices, respectively, to cover the entire 3D  
 169 space of decision-making episodes defined orthogonally over 'stimulus', 'choice', and 'feedback'  
 170 ( $5 \times 2 \times 2 = 20$  episodes; Fig 2C; Materials and methods).



171

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

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

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

## 199 **Decision-making processes for binary classification**

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

206 
$$p(CL = \text{large}) = p(S > B|m) = \int_B^{\infty} p(S|m)dS;$$

207 
$$p(CL = \text{small}) = 1 - p(CL = \text{large}),$$

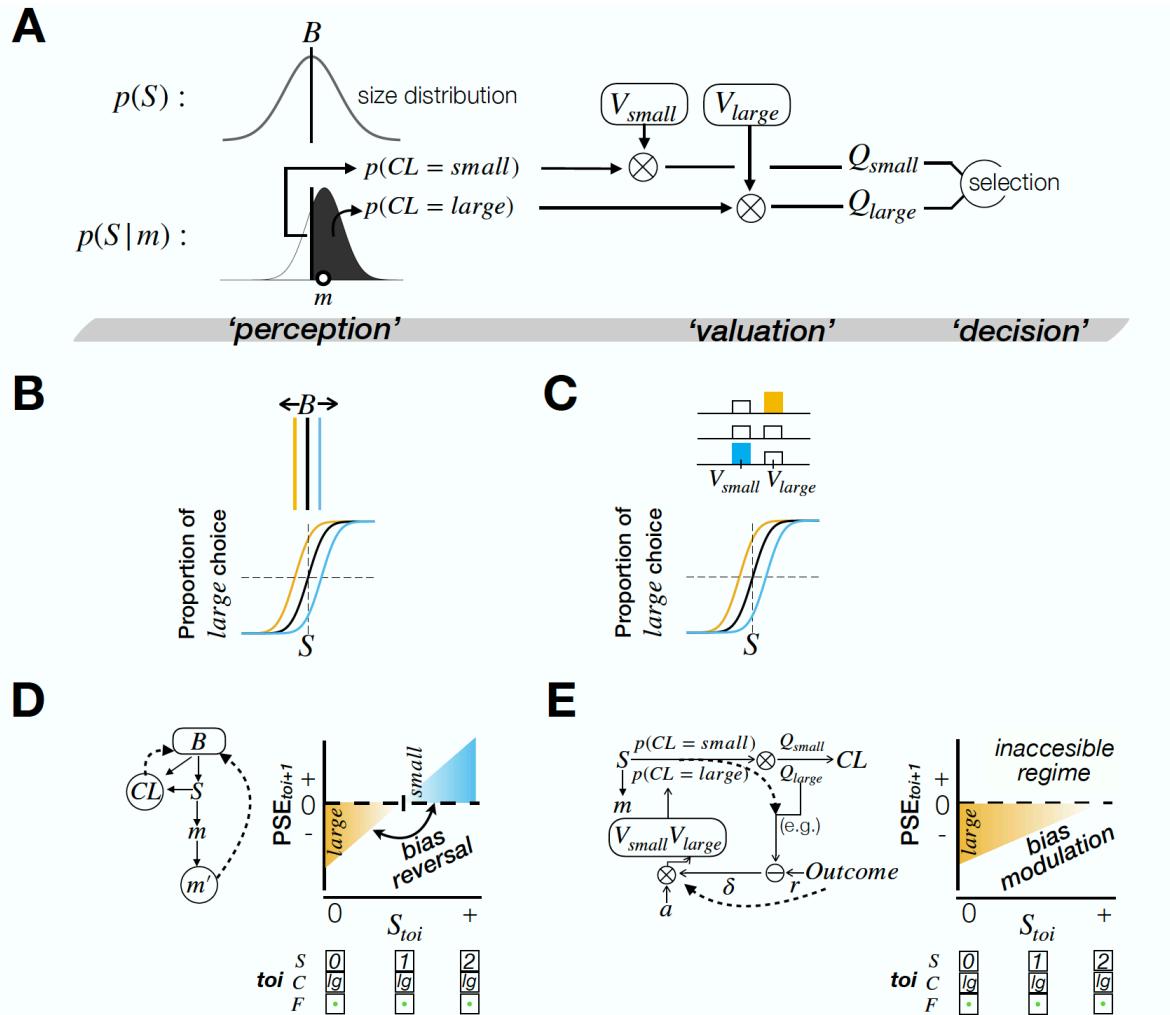
208 where  $CL$  stands for the class variable with the two ( $\text{small}$  and  $\text{large}$ ) states.

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

212 
$$Q_{\text{large}} = p(CL = \text{large}) \times V_{\text{large}};$$

213 
$$Q_{\text{small}} = p(CL = \text{small}) \times V_{\text{small}}.$$

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



223

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

240 **The belief-based RL model**

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

245 
$$V_{large(small)} \leftarrow V_{large(small)} + \alpha\delta;$$

246 
$$\delta = r - Q_{large(small)} = r - p(CL = large(small)) \times V_{large(small)},$$

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

## 258 **The Bayesian model of boundary-updating (BMBU)**

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

265 inverting its learned generative model of how  $m'$  and  $CL$  are generated (Fig 3D, left; Equations  
266 3-6 in Materials and methods for the detailed formalisms for the learned generative model), as  
267 follows:

268 
$$p(B|m', CL) \propto p(m', CL|B)p(B) \equiv p(m', C, F|B)p(B).$$

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

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

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

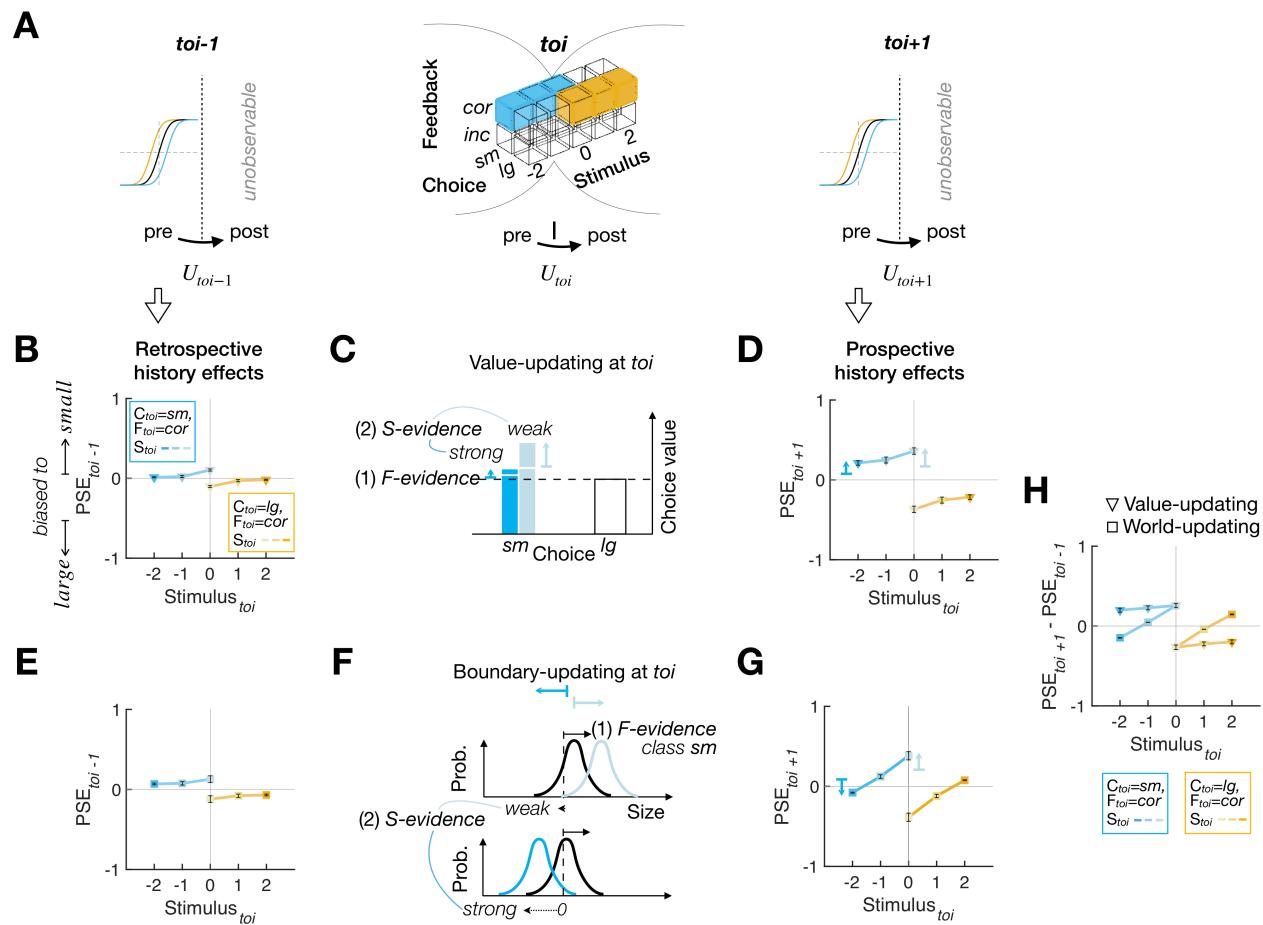
296 Next, to appreciate the stimulus-dependent nature of feedback effects in the world-  
297 updating scenario, let's consider another case where the decision-maker experiences a PDM  
298 episode with an unambiguous stimulus: the ring with size 2 is presented and produces a  
299 sensory measurement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

314 We stress that this ‘stimulus-dependent bias reversal’ is a hallmark of the world-updating  
315 scenario’s prediction of the history effects in PDM. Specifically, the direction of bias reversal is  
316 always from *small* to *large* as long as the feedback in conjunction with the choice indicates  
317  $CL = \text{small}$  (e.g.,  $\{S = 0 \rightarrow -1 \rightarrow -2, C = \text{small}, F = \text{correct}\}$  or  $\{S = 0 \rightarrow -1 \rightarrow -2, C =$   
318  $\text{large}, F = \text{incorrect}\}$ ) and always from *large* to *small* as long as the feedback in conjunction  
319 with the choice indicates  $CL = \text{large}$  (e.g.,  $\{S = 0 \rightarrow 1 \rightarrow 2, C = \text{large}, F = \text{correct}\}$  or  $\{S = 0 \rightarrow$   
320  $1 \rightarrow 2, C = \text{small}, F = \text{incorrect}\}$ ). Critically, the value-updating scenario does not predict the  
321 bias reversal (Fig 3E, right). It predicts that the feedback effects only asymptotically decrease as  
322 a function of sensory evidence but never switch to the other direction. This is because the  
323 decision confidence,  $p(CL = \text{large}(\text{small}))$ , only modulates the amount of value-updating but  
324 never changes the direction of value-updating.

325 ***Ex ante simulation of the feedback effects under the two***  
326 ***scenarios***

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

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



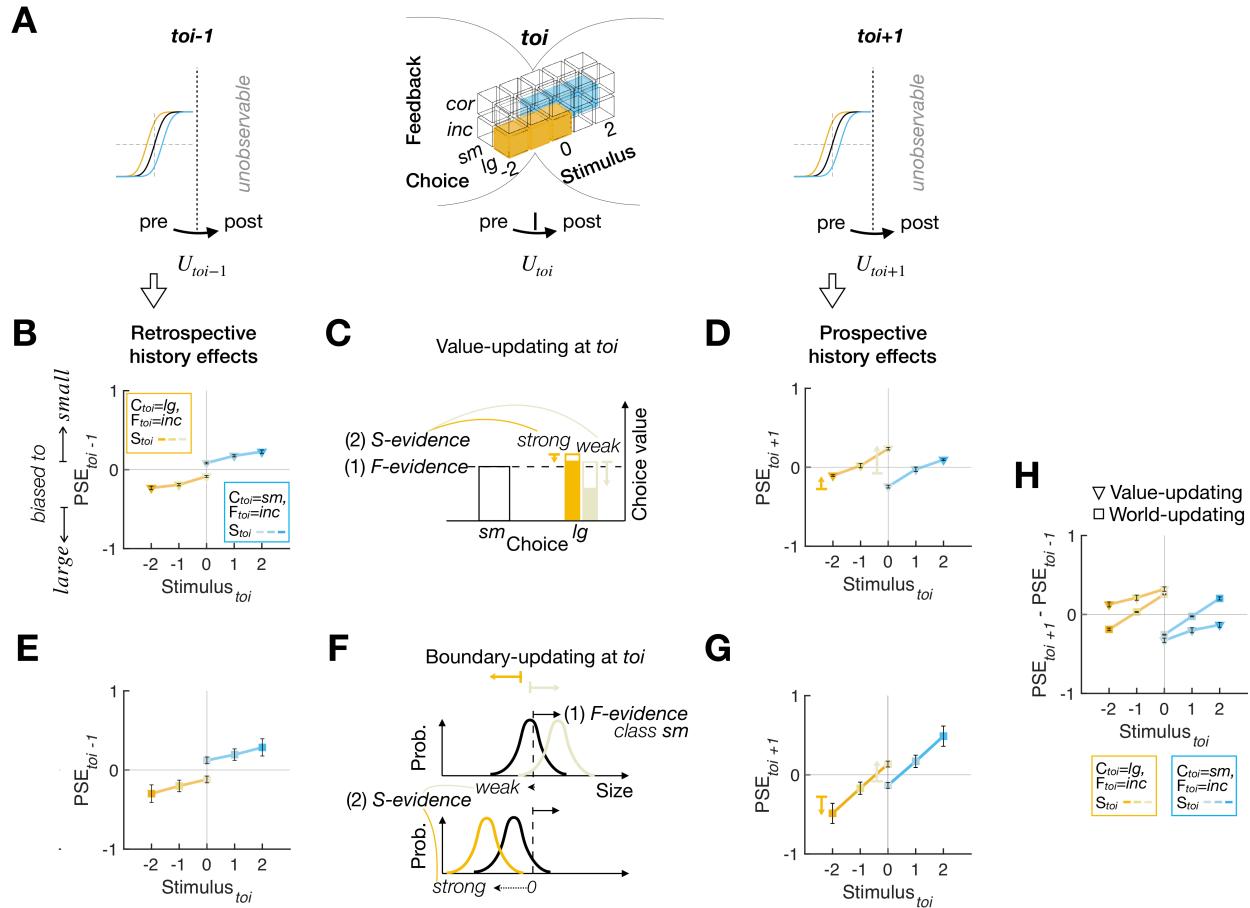
348  
 349 **Fig 4. Ex ante simulation results for the PDM episodes with *correct* feedback.** (A) Illustration of how  
 350 the retrospective (left) and prospective (right) history effects relate to the value updates and boundary

351 updates (bottom) occurring over the trials overarching the trial of interest. While the updating occurs  
352 latently at every trial (as indicated by  $U_{toi-1}$ ,  $U_{toi}$ ,  $U_{toi+1}$ ), its behavioral consequences are observable  
353 only at the pre-updating phase at  $toi-1$  and  $toi+1$ . **(B-D)** The observable retrospective **(B)** and prospective  
354 **(D)** history effects and latent value-updating processes **(C)** for the value-updating model agent. **(C)** Since  
355 *correct* feedback is treated as a positive reward, the chosen value is updated positively while the amount  
356 of value-updating varies depending on the strength of sensory evidence, as indicated by the length of the  
357 vertical arrows in different colors (weak sensory evidence, pale blue; strong sensory evidence, dark blue).  
358 The short horizontal bars and arrow heads of the colored arrows indicate the chosen values before and  
359 after  $U_{toi}$ , respectively. **(E-G)** The observable retrospective **(E)** and prospective **(G)** history effects and  
360 latent boundary-updating processes **(F)** for the world-updating model agent. **(F)** Since *correct* feedback is  
361 treated as a logical indicator of the true state of the class variable (i.e., the true inequality between the  
362 class boundary and the stimulus), the class boundary shifts as a joint function of feedback and sensory  
363 evidence, where the boundary shift due to sensory evidence (solid black arrows) counteracts that due to  
364 feedback (dotted black arrows), as indicated by the arrows in different colors (weak sensory evidence,  
365 pale blue; strong sensory evidence, dark blue). The short vertical bars and arrow heads of the colored  
366 arrows at the top indicate the class boundary before and after  $U_{toi}$ , respectively. **(H)** Juxtaposition of the  
367 differences between the retrospective and prospective history effects displayed by the two model agents.  
368 **(C, F)** The contributions of both sensory and feedback evidence are indicated by *S-evidence* and *F-*  
369 *evidence*, respectively. **(B, D, E, G)** Data points are the means and s.e.m.s across the parameter sets  
370 used in *ex ante* simulations (see Materials and methods). The data underlying this figure **(B, D, E, G, H)**  
371 can be found in S1 Data.  
372

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

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

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



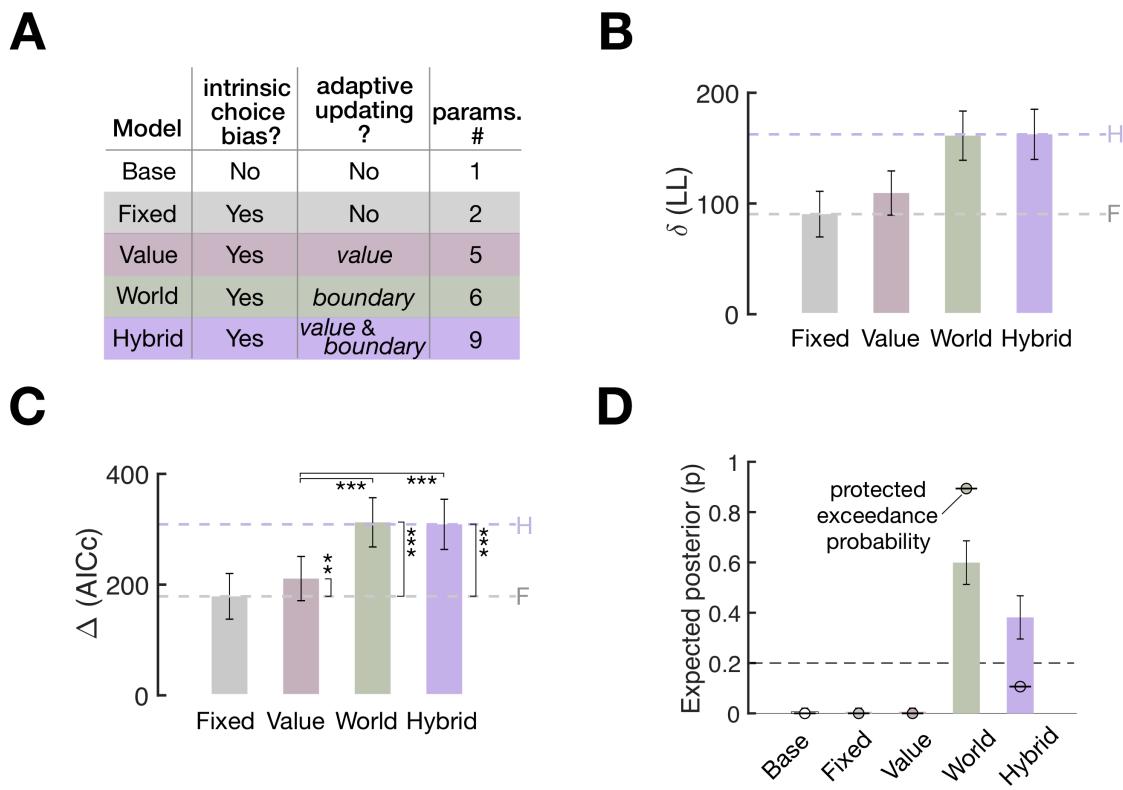
407

408 **Fig 5. Ex ante simulation results for the PDM episodes with *incorrect* feedback.** The format is  
 409 identical to that in Fig 4. The data underlying this figure (B, D, E, G, H) can be found in S1 Data.  
 410

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

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

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



428

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

432 participant average differences from the goodness-of-fit measures of the Base model (N=30,  
433 mean $\pm$ s.e.m.). Both difference measures indicate a better fit for higher values. Dashed lines in purple  
434 (Hybrid model) and gray (Fixed model) provide the reference points for evaluating the value-updating and  
435 world-updating models' accountability of the trial-to-trial choice variability (see Main text for their exact  
436 meanings). Pair-wise model comparisons were performed using paired one-tailed *t*-tests (asterisks  
437 indicate significance: \*,  $P<0.05$ ; \*\*,  $P<0.005$ ; \*\*\*,  $P<10^{-8}$ ) (D) Model comparisons in the hierarchical  
438 Bayesian Model Selection measures. Height of bars, expected posterior probabilities; error bars, standard  
439 deviation of posterior probabilities. Dots marked with short dashes, protected exceedance probability.  
440 Dashed lines, chance level ( $p = 0.2$ ), indicating the probability that a model is favored over others in  
441 describing the data by random chance. Bayesian omnibus risk (BOR), the estimated probability that  
442 observed differences in model frequencies may be due to chance, is reported ( $\text{BOR} = 1.7636 \times 10^{-10}$ ). The  
443 data underlying this figure (B, C, D) can be found in S1 Data.  
444

445 The Fixed model's performance relative to the Base model's (gray dashed lines in Fig  
446 6B, 6C) reflects the fraction of choice variability that is attributed to arbitrary choice preference.  
447 On the other hand, the Hybrid model's performance relative to the Base model's (purple dashed  
448 lines in Fig 6B, 6C) reflects the maximum fraction of choice variability that can be potentially  
449 explained by either the value-updating model, the world-updating model, or both. Thus, the  
450 difference in performance between the Hybrid and Fixed models (the space spanned between  
451 the gray and purple dashed lines in Fig 6B, 6C) quantifies the meaningful fraction of choice  
452 variability that the two competing models of interest are expected to capture. Prior to model  
453 evaluation, we confirmed that the two competing models (the value-updating and world-updating  
454 models) and two reference models (the Base and Hybrid models) are empirically distinguishable  
455 by carrying out a model recovery test (S3 Fig).

456 With this target fraction of choice variability to be explained, we evaluated the two  
457 competing models by comparing them against the Fixed and Hybrid models' performances while  
458 taking into account model complexity with AICc. The value-updating model was moderately  
459 better than the Fixed model (paired one-tailed *t*-test,  $t(29) = -2.8540$ ,  $P = 0.0039$ ) and  
460 substantially worse than the Hybrid model (paired one-tailed *t*-test,  $t(29) = 7.6996$ ,  $P =$   
461  $8.6170 \times 10^{-9}$ ) and the world-updating model (paired one-tailed *t*-test,  $t(29) = 8.3201$ ,  $P =$   
462  $1.7943 \times 10^{-9}$ ). By contrast, the world-updating model was substantially better than the Fixed  
463 model (paired one-tailed *t*-test,  $t(29) = -10.3069$ ,  $P = 1.6547 \times 10^{-11}$ ) but not significantly better

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

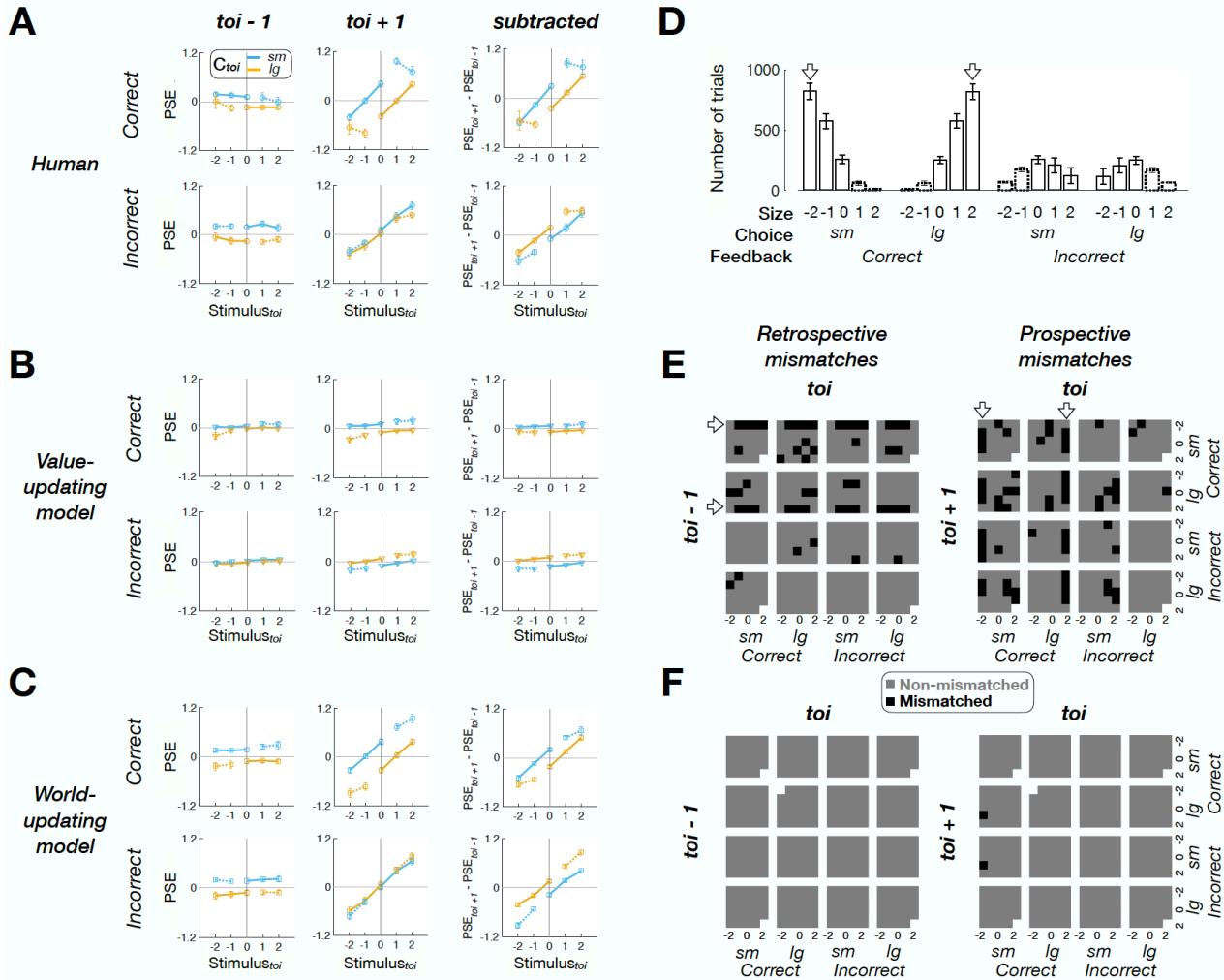
468 To complement the above pair-wise comparisons, we took the hierarchical Bayesian  
469 Model Selection approach [33–35] using AICc model evidence, to assess how probable it is that  
470 each of the five models prevails in the population (expected posterior probability; vertical bars in  
471 Fig 6D) and how likely it is that any given model is more frequent than the other models  
472 (protected exceedance probability; dots with horizontal bars in Fig 6D). Both measures  
473 corroborated the outcomes of the pair-wise comparisons: the world-updating model  
474 predominated in expected posterior probability (0.5992) and protected exceedance probability  
475 (0.8938).

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

## 478 ***Ex post simulation of the feedback effects under the two 479 scenarios***

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

487 The *ex post* simulation was identical to the *ex ante* simulation except that each decision-  
488 maker's best-fit model parameters were used (Table B in S1 Appendix; Materials and  
489 methods). We assessed how well the models reproduce the human history effects of feedback  
490 in two different ways. First, we compared the models and the humans similarly to the *ex ante*  
491 simulation (Fig 7A-7C). We included the PDM episodes with non-veridical feedback (symbols  
492 with dotted lines in Fig 7A-7C) though those episodes infrequently occurred ( $12.09 \pm 0.02\%$   
493 (mean  $\pm$  s.e.m.) out of total *toi* episode trials; bars with dotted outlines in Fig 7D). As a result, we  
494 inspected the retrospective and prospective history effects, and their differences, for all the  
495 possible combinations of 'stimulus,' 'choice,' and 'feedback' (20 PDM episodes in total), which  
496 resulted in a total of 60 PSE pairs to compare. The PSEs simulated by the world-update model  
497 closely matched the human PSEs, in both pattern and magnitude (Fig 7A and 7C), whereas  
498 those by the value-update model substantively deviated from the human PSEs (Fig 7A and 7B).  
499 The statistical comparison (paired two-tailed *t*-tests with Bonferroni correction) indicates that the  
500 value-updating model's PSEs significantly deviated from the corresponding human PSEs for  
501 almost half of the entire pairs (29 out of 60 pairs) whereas none of the world-updating model's  
502 PSEs significantly differed from the human PSEs (0 out of 60 pairs). Notably, most mismatches  
503 occurred because the value-updating model does not reverse the direction of feedback effects  
504 as sensory evidence becomes stronger while humans do so (compare the third columns of Fig  
505 7A and 7B).



506

507 **Fig 7. Ex post simulation results.** (A-C) Retrospective (left columns), prospective (middle columns), and  
508 subtractive (right columns) history effects in PSE for the human (A), value-updating (B), and world-  
509 updating (C) decision-makers. Top and bottom rows in each panel show the PSEs associated with the *toi*  
510 episodes involving *correct* and *incorrect* feedback. Symbols with error bars, mean $\pm$ s.e.m. across 30  
511 decision-makers. See S5 Fig for the results from the Hybrid model decision-makers. (D) Frequency of  
512 PDM episodes in the human data (mean and SD across participants). (E, F) Maps of significant  
513 deviations of the value-updating (E) and world-updating (F) model agents from the human decision-  
514 makers in the retrospective (left) and prospective (right) history effects. Gray and black cells of the maps  
515 mark the insignificant and significant deviations (paired two-tailed *t*-tests with the Bonferroni correction for  
516 multiple comparisons). Empty cells are data points with NaN values due to insufficient trials. The data  
517 underlying this figure (A, B, C, D, E, F) can be found in S1 Data.  
518

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

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

527 We carried out repeated  $t$ -tests with Bonferroni correction to see where the model-  
528 human mismatches occur (data were missing for a few cells—mostly those including non-  
529 veridical-feedback episodes, as indicated by the empty cells in Fig 7E and 7F, because those  
530 episodes were too rare (Fig 7D) to occur for all participants). For the remaining cells, the world-  
531 updating model showed a remarkable level of correspondence with the humans, deviating from  
532 the humans at only two cells (out of 790 cells, 0.25%; Fig 7F). By contrast, the value-updating  
533 model failed to match the humans for 94 cells (out of 792 cells, 11.87%; Fig 7E). Here, the  
534 mismatches occurred systematically: they were frequent when the preceding episode defining  
535 any given cell (i.e., episodes at  $toi-1$  for the retrospective cells or episodes at  $toi$  for the  
536 prospective cells) was featured with strong sensory evidence (as indicated by the arrows in Fig  
537 7E). This systematic deviation precisely reflects the incapability of the value-updating model to  
538 reverse the direction of feedback effects as sensory evidence strengthens.

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

541

## 542 **Discussion**

543 Here, we explored the two possible scenarios for what humans learn from corrective feedback  
544 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

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

578 To be sure, the idea of shifting the decision or class boundary toward past stimuli *per se*  
579 is not new and has been previously hypothesized [38,39] or implemented into various models  
580 [40–44]. However, BMBU goes beyond these efforts by offering a normative formalism of  
581 incorporating *correct* and *incorrect* feedback as evidence for the class boundary such that it  
582 has an equal footing as sensory evidence in PDM tasks. This integration of feedback and  
583 sensory evidence within the framework of BDT advances the current computational account of  
584 the history effects because it addresses the history factors in the complete dimensions of PDM  
585 ('stimulus', 'choice', and 'feedback'), which is important given the multiplexed nature of history  
586 effects emphasized by prior studies [8–11,31,45]. Our modeling work joins recent computational  
587 and empirical efforts of incorporating feedback in the normative evidence accumulation model  
588 [6,46], a framework commonly employed in various classic PDM tasks, such as a random-dot  
589 motion task. Furthermore, a study on rats' binary classification behavior has shown that rats can  
590 use information about the correct class state (referred to as "second-order prior" by the authors)  
591 by integrating their own choices with feedback (reward outcome) and that the population neural  
592 activity in the orbitofrontal cortex represents this information [11]. Together with these studies,  
593 our work supports a general view that decision-makers use corrective feedback as evidence for  
594 updating their world knowledge pertinent to the PDM task engaging them. Having mentioned the

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

597 Previously, the so-called “Anna Karenina” account was presented to describe the  
598 seemingly idiosyncratic *incorrect* feedback effects [9]. The Anna-Karenina account leaves the  
599 crucial aspect of feedback effects—the different consequences of *correct* vs *incorrect*  
600 feedback—unexplained. Since the belief-based RL model predicts the specific pattern of  
601 feedback effects for incorrect trials, as shown via *ex ante* simulation, endorsing the Anna-  
602 Karenina account admits that the belief-based RL model fails to account for the effects of  
603 *incorrect* feedback observed in animals. For that matter, past studies on the history effects in  
604 PDM paid little attention to incorrect trials because they are, owing to their infrequency,  
605 considered too noisy and unreliable to be properly analyzed. By contrast, BMBU accounts for  
606 the effects of feedback in a principled way, regardless of whether the feedback is *correct* or  
607 *incorrect*. Furthermore, BMBU explains why the feedback effects appear different between the  
608 correct and incorrect trials on the surface (compare the prospective history effects between Fig  
609 4 and Fig 5): the correct and incorrect trials share the same deterministic boundary-updating  
610 process but had different histories of their own stochastic events, which led to correct versus  
611 incorrect choices, respectively.

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

620 slow fluctuations in the decision-making strategy may cause the observed effects. This group  
621 dealt with this concern by subtracting the retrospective history effects from the prospective  
622 ones. A more recent study [48] shared this concern but disagreed about its remedy by showing  
623 that the subtraction method cannot fairly recover diverse systematic updating strategies.  
624 Alternatively, they took a model-based approach to separate any given updating strategy from  
625 random drifts in decision criteria. We acknowledge the importance of the efforts by these studies  
626 and share the same concern. But we emphasize that BMBU successfully reproduced human  
627 history effects in both directions of time without incorporating any non-systematic components  
628 arising from random drifts. BMBU's concurrent reproduction of the retrospective and prospective  
629 history effects was confirmed not just for the summary statistics (the PSEs in Fig 7C) but also  
630 for the individual data points spanning almost the entire space of PDM episode pairs (Fig 7F).  
631 This suggests that it is an empirical matter of whether the decision criterion slowly drifts or not,  
632 raising another concern that systematic history effects might be explained away as non-existing  
633 slow drifts. In this sense, we propose that researchers should treat the retrospective history  
634 effects not as a baseline or control condition but as what must be explained, the phenomenon  
635 equally important as the prospective history effects, before resorting to any non-systematic  
636 sources. We believe that such a treatment is the way historians treat historical events [49], and  
637 that our approach showcases its one rigorous example.  
638

## 639 **Materials and methods**

### 640 **Ethics statement**

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

645

### 646 **Participants**

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

649

### 650 **Procedure**

651 **Stimuli.** The stimulus was a thin (.07 deg in visual angle), Gaussian-noise filtered, black-and-  
652 white ring flickering at 20 Hz on a gray luminance background. On each trial, a fixation first  
653 appeared for 0.5 s on average (fixation duration uniformly jittered from 0.3 s to 0.7 s on a trial-  
654 to-trial basis) before the onset of a ring stimulus. Five different ring sizes (radii of 3.84, 3.92,  
655 4.00, 4.08, 4.16 deg in visual angle (d.v.a.), denoted by  $-2, -1, 0, 1, 2$ , respectively, in the main  
656 text) were randomized within every block of 5 trials.

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

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

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

684

685 **Data analysis**

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

691 
$$\psi(x; \mu, \sigma) = F(x; \mu, \sigma),$$
  
692 where  $\mu$  and  $\sigma$  are the mean and standard deviation of  $F$ . By finding the best-fitting value of  $\mu$ ,  
693 we defined the point of subjective equality (PSE; the stimulus level with equal probability for a  
694 *small* or *large* choice), which was used as the summary statistics that quantifies the history  
695 effects associated with a given PDM episode. To ensure reliable PSE estimates, we acquired  
696 bootstrap samples (N=5,000) of psychometric curves based on the binomial random process  
697 and took their average as the final estimate for each PDM episode. In our main data analysis,  
698 the results of which are displayed in Fig 7, we chose not to include the parameters for guess or  
699 lapse rates in estimating PSEs. This was done to prevent unfair overfitting problems from  
700 occurring in infrequent episode types with small numbers of trials available for fitting. On the  
701 other hand, to preclude any potential confounding problem related to the task difficulty  
702 associated with PDM episode types, we also repeated the above PSE estimation procedure  
703 with guess ( $\gamma$ ) and lapse ( $\lambda$ ) rates included as free parameters:  $\psi(x; \mu, \sigma, \gamma, \lambda) = \gamma + (1 - \gamma -$   
704  $\lambda)F(x; \mu, \sigma)$ . The results did not differ between the original estimation procedure without the  
705 lapse and guess rates and the procedure with the lapse and guess rates (Bonferroni-corrected  
706  $P = 0.2023 \sim 1.000$ ; paired two-tailed *t*-tests; see S2 Data for detailed statistical information).

707

## 708 **Value-updating model**

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

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

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

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

733 
$$V_C \leftarrow V_C + \alpha \delta. \quad (1)$$

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

737        
$$\delta = r - Q_C = r - p_C V_C. \quad (2)$$

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

747

## 748        **World-updating model**

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

754        **“Learned” generative model.** In BDT, the learned generative model refers to the decision-  
755        maker's subjective internal model that relates task-relevant variables ( $m$ ,  $m'$ , and  $B$  in the left  
756        panel of Fig 3D) to external stimuli and behavioral choices ( $S$  and  $CL$ , respectively, in the left  
757        panel of Fig 3D). As previously known [53,54], the decision-maker's internal model is likely to  
758        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  $\sigma_S^2$  (as depicted by  $B \rightarrow S$  in Fig 3D):

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

771 which defines the probability distribution of stimuli conditioned on the class boundary, where  $\sigma_S^2$   
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):

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

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  $\sigma_m^2$  (as depicted by  $S \rightarrow m$  in Fig 3D):

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

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

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

$$785 \quad p(m'|m) = N(m'; m, \sigma_{m'}^2), \quad (6)$$

786 which defines the probability distribution of mnemonic measurements conditioned on the  
787 sensory measurement, where  $\sigma_{m'}^2$ , corresponds to the extent to which the decision-maker's  
788 working memory system is noisy. This generative process ( $m \rightarrow m'$ ) is required because the  
789 sensory evidence of the stimulus is no longer available in the sensory system—due to a brief  
790 (0.3 sec; Fig 2B) stimulus duration—at the moment of updating the state of the class boundary  
791 (as will be shown below in the subsection titled “Boundary-updating”) and instead must be  
792 retrieved from the working memory system. The mnemonic recall of the stimulus is known to be  
793 noisy, becoming quickly deteriorated right away after stimulus offset, especially for continuous  
794 visual evidence such as color and orientation [56,57]. The generative process relating  $m$  to  $m'$   
795 has been adopted for the same reason by recent studies [58,59], including our group [55], and  
796 is consistent with the non-zero levels of memory noise in the model-fit results ( $\sigma_{m'}^2 = [1.567,$   
797  $5.606]$ ). The substantial across-individual variability of the fitted levels of  $\sigma_{m'}^2$ , is also consistent  
798 with the previous studies [55,58,59].

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

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

807

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

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

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

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

831             $CL_t = \text{large}$  if  $C_t = \text{large}$  and  $F_t = \text{correct}$  or if  $C_t = \text{small}$  and  $F_t = \text{incorrect}$ ;  
 832             $CL_t = \text{small}$  otherwise,  
 833            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  
 834            likelihood of the class boundary is computed by propagating  $m'_t$  and  $CL_t$  inversely over the  
 835            learned generative model (as defined by Equations 3-6):

836            
$$p(m'_t, CL_t | B_t) = \int p(m'_t, CL_t, S_t | B_t) dS_t = \int p(m'_t | S) p(CL_t | S_t, B_t) p(S_t | B_t) dS_t, \quad (8)$$

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

842            
$$\int p(m'_t | S) p(CL_t | S_t, B_t) p(S_t | B_t) dS_t =$$
  
 843            
$$= \int_{-\infty}^{B_t} p(m'_t | S_t) p(CL_t | S_t, B_t) p(S_t | B_t) dS_t + \int_{B_t}^{+\infty} p(m'_t | S_t) p(CL_t | S_t, B_t) p(S_t | B_t) dS_t, \quad (9)$$

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

852            
$$p(m'_t, CL_t = \text{small} | B_t) = \int_{-\infty}^{B_t} p(m'_t | S_t) p(S_t | B_t) dS_t;$$
  
 853            
$$p(m'_t, CL_t = \text{large} | B_t) = \int_{B_t}^{+\infty} p(m'_t | S_t) p(S_t | B_t) dS_t \quad (10)$$

854 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  
 855 generative model ( $S \rightarrow m \rightarrow m'$  in the left panel of Fig 3D; Equation 5-6; see Equation S2 for  
 856 derivations in Text in S1 Appendix). Equation 10 indicates that BMBU calculates how likely  
 857 hypothetical boundary states bring about the mnemonic measurement ( $B \rightarrow S \rightarrow m \rightarrow m'$ ) while  
 858 taking into account the informed state of the class variable ( $B \rightarrow CL \leftarrow S$ ), by constraining the  
 859 possible range of the stimulus states. To help readers intuitively appreciate these respective  
 860 contributions of the mnemonic measurement and the informed state of the class variable  
 861 (feedback) to the boundary likelihood, we further elaborated on how Equation 9 is reduced to  
 862 Equation 10 depending on the informed state of  $CL_t$  (see Text in SI Appendix and S1 Fig).

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

$$866 p(m'_t, CL_t = \text{small}|B_t) = \frac{1}{\sqrt{2\pi \left( \frac{\sigma_M^2 \sigma_S^2}{(\sigma_M^2 + \sigma_S^2)} \right)}} \int_{-\infty}^{B_t} e^{-\frac{\left( S_t - \frac{B_t \sigma_M^2 + m'_t \sigma_S^2}{\sigma_M^2 + \sigma_S^2} \right)^2}{2(\sigma_M^2 + \sigma_S^2)}} dS_t \times \frac{1}{\sqrt{2\pi(\sigma_M^2 + \sigma_S^2)}} e^{-\frac{(B_t - m'_t)^2}{2(\sigma_M^2 + \sigma_S^2)}}. \quad (11)$$

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

$$870 p(m'_t, CL_t = \text{large}|B_t) = \frac{1}{\sqrt{2\pi \left( \frac{\sigma_M^2 \sigma_S^2}{(\sigma_M^2 + \sigma_S^2)} \right)}} \int_{B_t}^{\infty} e^{-\frac{\left( S_t - \frac{B_t \sigma_M^2 + m'_t \sigma_S^2}{\sigma_M^2 + \sigma_S^2} \right)^2}{2(\sigma_M^2 + \sigma_S^2)}} dS_t \times \frac{1}{\sqrt{2\pi(\sigma_M^2 + \sigma_S^2)}} e^{-\frac{(B_t - m'_t)^2}{2(\sigma_M^2 + \sigma_S^2)}}. \quad (12)$$

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

875 
$$p(B_t|m'_t, CL_t) \propto p(m'_t, CL_t|B_t)p(B_t). \quad (13)$$

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

881 
$$\hat{B}_t = \lambda \hat{B}_{t-1}^{post} + (1 - \lambda) \mu_0;$$
  
882 
$$\sigma_{B_t}^2 = \lambda \sigma_{t-1}^{2post} + \sigma_{diffusion}^2, \quad (14)$$

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

884 Note that the decay parameter  $\lambda = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_{t-1}^{2post}}$  influences the width and location of the  
885 belief distribution, and that the diffusive noise of  $\sigma_{diffusion} > 0$  helps to keep the width of the  
886 distribution over multiple trials, thus avoiding sharpening and stopping the updating process  
887 [60]. In this way,  $\lambda$  and  $\sigma_{diffusion}$  allows BMBU to address the idiosyncratic choice bias and  
888 noise, as we allow the belief-based RL model to do with  $\mu_0$  and the softmax rule.

889 In sum, BMBU posits that human individuals carry out a sequence of binary classification  
890 trials with their learned generative model while continually updating their belief about the  
891 location of the class boundary in that generative model. BMBU describes these decision-making  
892 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.

894

895 **Reference models**

896 As the references for evaluating the belief-based RL model and BMBU in predicting the  
897 variability of human choices, we created three reference models. The 'Base' model captures the  
898 choice variability that can be explained by the  $p$ -computation with the class boundary fixed at 0  
899 unanimously for all participants and without any value-updating process. Thus, it has only a  
900 single free parameter representing the variability of the sensory measurement ( $\theta = \{\sigma_m\}$ ). The  
901 'Fixed' model captures the choice variability that can be explained by the  $p$ -computation with the  
902 class boundary set free to a fixed constant  $\mu_0$  for each participant and without any value-  
903 updating process. Thus, it has two free parameters ( $\theta = \{\mu_0, \sigma_m\}$ ). The 'Hybrid' model captures  
904 the choice variability that can be explained both by the  $p$ -computation with the inferred class  
905 boundary by BMBU and by the value-updating process implemented by the belief-based RL  
906 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  
907 6B-6D, the differential goodness-of-fit measures on the y-axis indicate the subtractions of the  
908 performance of the 'Base' model from those of the remaining models.

909

## 910 **Model fitting**

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

$$915 \quad LL(\theta, model) = \log p(data | \theta, model) = \sum_{k=1}^{K_{cond}} \sum_{j=1}^{M_{runs}} \sum_{i=1}^{N_{trials}} \log p(C_{i,j,k} | \theta, model),$$

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

920 the model parameters was obtained with Bayesian Adaptive Direct Search (BADS) [62], a  
921 hybrid Bayesian optimization to find the parameter vector  $\theta^*$  that maximizes the log likelihood,  
922 which works well with stochastic target functions. To reduce the risk of being stuck at local  
923 optima, we repeated 20 independent fittings by setting the starting positions randomly using  
924 Latin hypercube sampling (*lhsdesign\_modified.m* by Nassim Khlaled,  
925 <https://www.mathworks.com/matlabcentral/fileexchange/45793-latin-hypercube>), then picked the  
926 fitting with the highest log likelihood. To avoid infinite loops from using IBS, we did not impose  
927 individual lapse rates in an arbitrary manner. Instead, we calculated the average of the lapse  
928 rate and guess rate from the cumulative Gaussian fit to a given individual's grand mean (based  
929 on the entire trials) psychometric curve. With these individual lapse probabilities (mean rate of  
930 0.05, which ranged [0.0051, 0.1714]), trials were randomly designated as lapse trials, in which  
931 the choice was randomly determined to be either *small* or *large*.

932

### 933 **Model comparison in goodness-of-fit**

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

$$936 \quad AICc = -2 \cdot LL(\theta^*) + 2p + \frac{2p(p+1)}{(N \times M \times K) - p - 1},$$

937 where  $p$  is the number of parameters of the model and the total number of trials in the dataset is  
938  $N \times M \times K$ . Log model evidence was obtained for each participant by multiplying AICc by  $-1/2$   
939 [35]. Furthermore, we took a hierarchical Bayesian Model Selection approach that infers the  
940 posterior over model frequencies in the population based on log model evidence values in each  
941 subject. To conclude whether a given model is the most likely model above and beyond chance,  
942 we also reported protected exceedance probabilities for each model (see Fig 6E and 6F). The

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

945

## 946 **Model recovery analysis**

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

957

## 958 ***Ex ante* and *Ex post* model simulations**

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

968 507,300~510,000 trials per participant. For simplicity, we assumed neither lapse trials nor any  
969 arbitrary choice bias.

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

972

## 973 **Statistical tests**

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

987

## 988 **Acknowledgments**

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

991

## 992 References

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

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

998 3. Abrahamyan A, Silva LL, Dakin SC, Carandini M, Gardner JL. Adaptable history  
999 biases in human perceptual decisions. *Proc National Acad Sci.* 2016;113(25):E3548–  
1000 57.

1001 4. Busse L, Ayaz A, Dhruv NT, Katzner S, Saleem AB, Scholvinck ML, Zaharia AD,  
1002 Carandini M. The detection of visual contrast in the behaving mouse. *J Neurosci.*  
1003 2011;31(31):11351–61.

1004 5. Scott BB, Constantinople CM, Erlich JC, Tank DW, Brody CD. Sources of noise  
1005 during accumulation of evidence in unrestrained and voluntarily head-restrained rats.  
1006 *eLife.* 2015;4:e11308.

1007 6. Mendonça AG, Drugowitsch J, Vicente MI, DeWitt EEJ, Pouget A, Mainen ZF. The  
1008 impact of learning on perceptual decisions and its implication for speed-accuracy  
1009 tradeoffs. *Nat Commun.* 2020;11(1):2757.

1010 7. Fernberger SW. Interdependence of judgments within the series for the method of  
1011 constant stimuli. *J Exp Psychol.* 1920;3(2):126–50.

1012 8. Lak A, Nomoto K, Keramati M, Sakagami M, Kepecs A. Midbrain dopamine neurons  
1013 signal belief in choice accuracy during a perceptual decision. *Curr Biol.* 2017;27(6):821–  
1014 32.

1015 9. Lak A, Hueske E, Hirokawa J, Masset P, Ott T, Urai AE, Donner TH, Carandini M,  
1016 Tonegawa S, Uchida N, Kepecs A. Reinforcement biases subsequent perceptual  
1017 decisions when confidence is low, a widespread behavioral phenomenon. *eLife.*  
1018 2020;9:e49834.

1019 10. Lak A, Okun M, Moss MM, Gurnani H, Farrell K, Wells MJ, Reddy CB, Kepecs A,  
1020 Harris KD, Carandini M. Dopaminergic and prefrontal basis of learning from sensory  
1021 confidence and reward value. *Neuron.* 2020;105(4):700–711.e6.

1022 11. Nogueira R, Abolafia JM, Drugowitsch J, Balaguer-Ballester E, Sanchez-Vives MV,  
1023 Moreno-Bote R. Lateral orbitofrontal cortex anticipates choices and integrates prior with  
1024 current information. *Nat Commun.* 2017;8(1):14823.

1025 12. Lee D, Seo H, Jung MW. Neural basis of reinforcement learning and decision  
1026 making. *Annu Rev Neurosci.* 2012;35(1):287–308.

1027 13. Daw ND, Doya K. The computational neurobiology of learning and reward. *Curr*  
1028 *Opin Neurobiol.* 2006;16(2):199–204.

1029 14. Sutton R, Barto A. *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT  
1030 Press; 1998.

1031 15. Corrado GS, Sugrue LP, Seung HS, Newsome WT. Linear-nonlinear-poisson  
1032 models of primate choice dynamics. *J Exp Anal Behav.* 2005;84(3):581–617.

1033 16. Lau B, Glimcher PW. Dynamic response-by-response models of matching behavior  
1034 in rhesus monkeys. *J Exp Anal Behav.* 2005;84(3):555–79.

1035 17. Gläscher J, Daw N, Dayan P, O'Doherty JP. States versus rewards: dissociable  
1036 neural prediction error signals underlying model-based and model-free reinforcement  
1037 learning. *Neuron.* 2010;66(4):585–95.

1038 18. Lee SW, Shimojo S, O'Doherty JP. Neural computations underlying arbitration  
1039 between model-based and model-free learning. *Neuron.* 2014;81(3):687–99.

1040 19. Odoemene O, Pisupati S, Nguyen H, Churchland AK. Visual evidence accumulation  
1041 guides decision-making in unrestrained mice. *J Neurosci.* 2018;38(47):10143–55.

1042 20. Hermoso-Mendizabal A, Hyafil A, Rueda-Orozco PE, Jaramillo S, Robbe D, Rocha  
1043 J de la. Response outcomes gate the impact of expectations on perceptual decisions.  
1044 *Nat Commun.* 2020;11(1):1057.

1045 21. Akrami A, Kopec CD, Diamond ME, Brody CD. Posterior parietal cortex represents  
1046 sensory history and mediates its effects on behaviour. *Nature.* 2018;554(7692):368–72.

1047 22. Summerfield C, Tsetsos K. Building bridges between perceptual and economic  
1048 decision-making: neural and computational mechanisms. *Front Neurosci.* 2012;6:70.

1049 23. Kahnt T, Grueschow M, Speck O, Haynes JD. Perceptual learning and decision-  
1050 making in human medial frontal cortex. *Neuron.* 2011;70(3):549–59.

1051 24. Polanía R, Krajbich I, Grueschow M, Ruff CC. Neural oscillations and  
1052 synchronization differentially support evidence accumulation in perceptual and value-  
1053 based decision making. *Neuron.* 2014;82(3):709–20.

1054 25. Körding KP, Wolpert DM. Bayesian decision theory in sensorimotor control. *Trends*  
1055 *Cogn Sci.* 2006;10(7):319–26.

1056 26. Trommershäuser J, Maloney LT, Landy MS. Decision making, movement planning  
1057 and statistical decision theory. *Trends Cogn Sci.* 2008;12(8):291–7.

1058 27. Griffiths TL, Tenenbaum JB. Optimal predictions in everyday cognition. *Psychol Sci.*  
1059 2006;17(9):767–73.

1060 28. Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND. How to grow a mind: statistics,  
1061 structure, and abstraction. *Science.* 2011;331(6022):1279–85.

1062 29. Glaze CM, Filipowicz ALS, Kable JW, Balasubramanian V, Gold JI. A bias–variance  
1063 trade-off governs individual differences in on-line learning in an unpredictable  
1064 environment. *Nat Hum Behav.* 2018;2(3):213–24.

1065 30. Ma WJ. Bayesian decision models: A primer. *Neuron.* 2019;104(1):164–75.

1066 31. Hachen I, Reinartz S, Brasselet R, Stroligo A, Diamond ME. Dynamics of history-  
1067 dependent perceptual judgment. *Nat Commun.* 2021;12(1):6036.

1068 32. Burnham KP, Anderson DR. *Model Selection and Inference, A Practical Information-  
1069 Heuristic Approach.* New York: Springer-Verlag; 2002.

1070 33. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model  
1071 selection for group studies. *Neuroimage.* 2009;46(4):1004–17.

1072 34. Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group  
1073 studies — Revisited. *Neuroimage.* 2014;84:971–85.

1074 35. Cao Y, Summerfield C, Park H, Giordano BL, Kayser C. Causal inference in the  
1075 multisensory brain. *Neuron.* 2019;102(5):1076-1087.e8.

1076 36. Palminteri S, Wyart V, Koechlin E. The importance of falsification in computational  
1077 cognitive modeling. *Trends Cogn Sci.* 2017;21(6):425–33.

1078 37. Urai AE, Braun A, Donner TH. Pupil-linked arousal is driven by decision uncertainty  
1079 and alters serial choice bias. *Nat Commun.* 2017;8(1):14637.

1080 38. Zariwala HA, Kepecs A, Uchida N, Hirokawa J, Mainen ZF. The limits of deliberation  
1081 in a perceptual decision task. *Neuron.* 2013;78(2):339–51.

1082 39. Renart A, Machens CK. Variability in neural activity and behavior. *Curr Opin  
1083 Neurobiol.* 2014;25:211–20.

1084 40. Qamar AT, Cotton RJ, George RG, Beck JM, Prezhdo E, Laudano A, Tolias AS, Ma  
1085 WJ. Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual  
1086 categorization. *Proc National Acad Sci.* 2013;110(50):20332–7.

1087 41. Summerfield C, Behrens TE, Koechlin E. Perceptual classification in a rapidly  
1088 changing environment. *Neuron*. 2011;71(4):725–36.

1089 42. Norton EH, Fleming SM, Daw ND, Landy MS. Suboptimal criterion learning in static  
1090 and dynamic environments. *PLoS Comput Biol*. 2017;13(1):e1005304.

1091 43. Treisman M, Williams TC. A theory of criterion setting with an application to  
1092 sequential dependencies. *Psychol Rev*. 1984;91(1):68–111.

1093 44. Lages M, Treisman M. A criterion setting theory of discrimination learning that  
1094 accounts for anisotropies and context effects. *Seeing Perceiving*. 2010;23(5):401–34.

1095 45. Fritsche M, Mostert P, Lange FP de. Opposite effects of recent history on perception  
1096 and decision. *Curr Biol*. 2017;27(4):590–5.

1097 46. Drugowitsch J, Mendonça AG, Mainen ZF, Pouget A. Learning optimal decisions  
1098 with confidence. *Proc National Acad Sci*. 2019;116(49):24872–80.

1099 47. Mochol G, Kiani R, Moreno-Bote R. Prefrontal cortex represents heuristics that  
1100 shape choice bias and its integration into future behavior. *Curr Biol*. 2021;31(6):1234–  
1101 1244.e6.

1102 48. Gupta D, Brody CD. Limitations of a proposed correction for slow drifts in decision  
1103 criterion. *arXiv preprint arXiv:220510912*. 2022.

1104 49. Carr EH. *What is History?* London: University of Cambridge & Penguin Books; 1961.

1105 50. Wichmann FA, Hill NJ. The psychometric function: I. Fitting, sampling, and  
1106 goodness of fit. *Percept Psychophys*. 2001;63(8):1293–313.

1107 51. Wichmann FA, Hill NJ. The psychometric function: II. Bootstrap-based confidence  
1108 intervals and sampling. *Percept Psychophys*. 2001;63(8):1314–29.

1109 52. Schütt HH, Harmeling S, Macke JH, Wichmann FA. Painfree and accurate Bayesian  
1110 estimation of psychometric functions for (potentially) overdispersed data. *Vision Res*.  
1111 2016;122:105–23.

1112 53. Ma WJ. Organizing probabilistic models of perception. *Trends Cogn Sci*.  
1113 2012;16(10):511–8.

1114 54. Haefner RM, Berkes P, Fiser J. Perceptual decision-making as probabilistic  
1115 inference by neural sampling. *Neuron*. 2016;90(3):649–60.

1116 55. Lee H, Lee HJ, Choe KW, Lee SH. Neural evidence for boundary updating as the  
1117 source of the repulsive bias in classification. *J Neurosci*. 2023;43(25):4664–83.

1118 56. Wilken P, Ma WJ. A detection theory account of change detection. *J Vis.*  
1119 2004;4(12):11.

1120 57. Bays PM, Gorgoraptis N, Wee N, Marshall L, Husain M. Temporal dynamics of  
1121 encoding, storage, and reallocation of visual working memory. *J Vis.* 2011;11(10):6–6.

1122 58. Luu L, Stocker AA. Post-decision biases reveal a self-consistency principle in  
1123 perceptual inference. *eLife.* 2018;7:e33334.

1124 59. Luu L, Stocker AA. Categorical judgments do not modify sensory representations in  
1125 working memory. *PLoS Comput Biol.* 2021;17(6):e1008968.

1126 60. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for  
1127 exploratory decisions in humans. *Nature.* 2006;441(7095):876–9.

1128 61. Opheusden B van, Acerbi L, Ma WJ. Unbiased and efficient log-likelihood estimation  
1129 with inverse binomial sampling. *PLoS Comput Biol.* 2020;16(12):e1008483.

1130 62. Acerbi L, Ma WJ. Practical Bayesian optimization for model fitting with Bayesian  
1131 adaptive direct search. *Advances in Neural Information Processing Systems.*  
1132 2017;30:1836–46.

1133 63. Daunizeau J, Adam V, Rigoux L. VBA: A Probabilistic Treatment of Nonlinear  
1134 Models for Neurobiological and Behavioural Data. *PLoS Comput Biol.*  
1135 2014;10(1):e1003441.

## 1136 Supporting Information

### 1137 **S1 Fig. Schematic illustration of BMBU's account of how the joint contribution of** 1138 **the sensory and feedback evidence to boundary updating leads to the reversal of** 1139 **choice bias as a function of sensory evidence strength.**

1140 **(A)** Reversal of subsequent choice bias—expressed in PSE—as a function of sensory  
1141 evidence strength and boundary inference—expressed in likelihood computation—  
1142 based on a PDM episode. Left panel: The circles with different colors (indicated by (b-  
1143 d), which points to the corresponding panels below **(B-D)**) represent the PSEs  
1144 associated with the boundary updating for three example PDM episodes, where the  
1145 stimulus ( $S_t$ ) varies from 0 to 2 while the choice ( $C_t$ ) and feedback ( $F_t$ ) are *large* and  
1146 *correct*, respectively. Right panel: At the core of boundary inference is the computation  
1147 of the likelihood of the class boundary based on the mnemonic measurement ( $m'_t$ ) and  
1148 the informed state of the class variable ( $CL_t$ ), where  $CL_t$  is jointly determined by  $F_t$  and  
1149  $C_t$  (see **Materials and methods** for the full computation of boundary inference in  
1150 BMBU). **(B-D)** The likelihoods of the class boundary given the three example PDM  
1151 episodes defined in **(A)**, where sensory evidence varies from the low **(B)**, to the  
1152 intermediate **(C)**, and to the high **(D)** level. To help understand why and how, given the  
1153 same feedback evidence, the direction of boundary updating reverses as the sensory  
1154 evidence strengthens, we visualize the boundary likelihoods as a product of two  
1155 functions (Equation 12), indicated by sub-panels marked as (1) and (2). Top row: As  
1156 indicated by (1), we plot each boundary likelihood when only the mnemonic  
1157 measurement is considered, assuming that no feedback is provided. Note that these  
1158 likelihood functions are centered around the values of  $m'_t$ , by attracting the class  
1159 boundary toward themselves, driving a shift towards the *large* side (i.e. positive side on  
1160 the boundary axis). Middle-Bottom rows: When the feedback evidence is given—i.e.,  
1161 when the informed state of  $CL_t$  is revealed as *large*—in addition to the mnemonic  
1162 measurement, an additional piece of information about the class boundary arises. As  
1163 indicated by (1)×(2), we plot each boundary likelihood (defined in **(A)**). As indicated by  
1164 (2), we plot each function (Middle row), as the result of (Bottom row) divided by (Top  
1165 row). The complementary cumulative distribution functions shown here are also  
1166 centered around  $m'_t$  because the *large* state of  $CL_t$  means that the class boundary is  
1167 located somewhere smaller than  $m'_t$ . Note that these skewed distributions push the  
1168 inferred class boundary away from the state of  $CL_t$  informed by feedback, driving a shift  
1169 towards the *small* side (i.e. negative side on the boundary axis). Consequently, the  
1170 influences from the sensory evidence and the feedback evidence counteract each other  
1171 (Bottom row). Note that the likelihood functions are centered in the *small* side when the  
1172 sensory evidence is weak **(B)**, in the neutral side when intermediate **(C)**, and in the  
1173 large side when strong **(D)**. These systematic shifts of the class-boundary likelihood as  
1174 a function of the strength of sensory evidence predict that the PSE of the psychometric  
1175 curve for the subsequent trial ( $t+1$ ) reverses its sign from negative to positive as a  
1176 function of the stimulus size, as shown in **(A)**.

1177 (TIF)

1178 **S2 Fig. Example trial courses of estimated class boundary.** (A) An example trial  
1179 history to show how a temporal trajectory of the class boundary inferred by BMBU. For  
1180 example, at trial #1 (x-axis), a physical stimulus (symbol x) was 0, a sensory  
1181 measurement (symbol o) was a positive value when the boundary belief (solid black  
1182 bar; y-axis) was centered at 0. BMBU's choice was *large* (symbol square on the top of  
1183 y-axis), and correct feedback (same square filled with green color) was provided, which  
1184 indicates that the class variable at trial #1  $CL_1$  was *large* (arrow's direction indicates the  
1185 effect of the trial class variable on the subsequent boundary-updating). BMBU updates  
1186 one's belief based on evidence from stimulus (colored symbol o) and feedback ( $CL_1$ ),  
1187 available at the time of boundary-updating. To illustrate cases where the bias reversal  
1188 we defined in Fig 3D in the main text happen and do not happen, same examples were  
1189 intentionally used as those we used in S1 Fig where we further detailed on the model's  
1190 mechanisms. Depending on colors, sensory evidence is weak (yellow symbol o) or  
1191 strong (purple symbol o), which leads to whether or not the reversal happens. Trial  
1192 cases featured in a red box indicates that the "Reinforcement" principle is held  
1193 (predicting subsequent choices to repeat *large* choice) while those featured in a green  
1194 box indicates that the "Reversal" happens (predicting subsequent choices to reverse the  
1195 previously made *large* choice). (B) Temporal trajectories of the class boundary when  
1196 the same 6-trial sequence of physical stimuli in (A) was simulated for 100 times. This  
1197 means different  $m$  and  $m'$  were realized. The data underlying this figure (A, B) can be  
1198 found in S1 Data.

1199 (TIF)

1200

1201 **S3 Fig. Model recovery analysis.** Each square represents exceedance probability  $p_{\text{exc}}$   
1202 from model recovery procedure. The 'ground-truth' model to simulate synthetic behavior  
1203 was correctly recovered with  $p_{\text{exc}} > 0.9$  for all 4 models considered in the study. The light  
1204 shade of the diagonal squares indicates that the ground-truth model was the best-fitting  
1205 model, leading to a successful model recovery. Numerical values can also be found in  
1206 S1 Data.

1207 (TIF)

1208 **S4 Fig. Histograms of classification accuracies of the human participants and**  
1209 **their model partners in the *ex ante* simulations.**

1210 **(A, B)** Across-individual distributions of the classification accuracy of the belief-based  
1211 RL model (A) and BMBU (B) overlaid on those of the human participants. The models'  
1212 choices were generated via *ex ante* simulations with a specific set of model parameters  
1213 (Table A in S1 Appendix), the results of which are depicted in Fig 4 and Fig 5. The  
1214 classification accuracy is measured by calculating the percentage of the trials in which  
1215 the choice matched the feedback used in the actual experiment. The empty bars  
1216 correspond to the histogram of human performances, the range of which is demarcated  
1217 by the dashed vertical lines ([min, max]=[60.65%, 73.94%]). The average human  
1218 classification accuracy was 67.85%. (A) Comparison of classification accuracy between  
1219 the belief-based RL model's simulation (red color) and the human choices. The model's  
1220 *ex ante* simulation accuracy was not different from the human accuracy ( $t(53) = 1.4429$ ,  
1221  $P = 0.1549$ ; Null hypothesis: model's performance vector and humans' performance

1222 vector come from populations with equal means, unpaired two-tailed  $t$ -test). **(B)**  
1223 Comparison of classification accuracy between BMBU's simulation (green color) and  
1224 the human choices. The model's *ex ante* simulation accuracy was not different from the  
1225 human accuracy ( $t(53) = 0.9707, P = 0.3361$ , unpaired two-tailed  $t$ -test). There was no  
1226 significant difference in classification accuracy between the value-updating model and  
1227 BMBU ( $t(48) = 0.5733, P = 0.5691$ , unpaired two-tailed  $t$ -test). The data underlying this  
1228 figure **(A, B)** can be found in S1 Data.  
1229 (TIF)

1230 **S5 Fig. Retrospective (left columns), prospective (middle columns), and**  
1231 **subtractive (right columns) history effects in PSE for the 'Hybrid' model's *ex post***  
1232 **model simulations.**

1233 Top and bottom rows in each panel show the PSEs associated with the *toi* episodes  
1234 involving *correct* and *incorrect* feedback at *toi*. Symbols with error bars,  $\text{mean} \pm \text{s.e.m.}$   
1235 across the 30 model agents, which correspond to their 30 human partners. The colors  
1236 of the symbols and lines label choices (blue: *small* and yellow: *large*). The data  
1237 underlying this figure can be found in S1 Data.

1238 (TIF)

1239 **S6 Fig. Maps of frequency deviations of the value-updating (A) and world-**  
1240 **updating (B) model agents' classifications in the *ex post* simulations from the**  
1241 **human decision-makers in the retrospective (left) and prospective (right) history**  
1242 **effects.**

1243 Each cell represents a pair of PDM episodes, as specified by the column and row  
1244 labels. At each cell, the color represents how much the episode frequency observed in  
1245 the model agents deviates from that observed in the corresponding human decision-  
1246 makers. The results of statistical tests on these deviations are summarized in Fig 7E  
1247 and 7F. The data underlying this figure **(A, B)** can be found in S1 Data.  
1248 (TIF)

1249 **S7 Fig. Retrospective (left columns), prospective (middle columns), and**  
1250 **subtractive (right columns) history effects in PSE for the human classification**  
1251 **performances of Urai et al. (2017)'s work [37] (A) and Hachen et al. (2021)'s work**  
1252 **[31] (B).**

1253 **(A, B)** We downloaded both publicly available datasets, analyzed them in the same way  
1254 that we analyzed human observers in our work, and plotted the results in the same  
1255 format used for Fig 7A. Top and bottom rows in each panel show the PSEs associated  
1256 with the *toi* episodes involving *correct* and *incorrect* feedback. Symbols with error bars,  
1257  $\text{mean} \pm \text{s.e.m.}$  across human observers. The colors of the symbols and lines label  
1258 choices (blue: *small* and yellow: *large*). The overall patterns of the PSEs plotted here  
1259 appear similar to those plotted in Fig 7A, displaying the reversals in direction of  
1260 stimulus-dependent feedback effects. When the same statistical tests used in our work  
1261 were carried out, some of the data points at the stimuli with strong sensory evidence at  
1262 *toi* significantly deviated from zero in the direction opposite to the feedback effect  
1263 predicted by the value-updating scenario, as indicated by the asterisks. **(A)** Sequential

1264 features of human observers (N=27) analyzed in our way from human dataset that once  
1265 had been published [37], which is openly available  
1266 (<http://dx.doi.org/10.6084/m9.figshare.4300043>), then analyzed in the previous study  
1267 [9]. In this study, the participants performed a binary classification task on the difference  
1268 in motion coherence by sorting the pairs of random-dot-kinematogram stimuli shown in  
1269 two intervals (s1 and s2) into one of the two classes ('s1<s2' vs. 's1>s2') over  
1270 consecutive trials. The presented stimuli were taken from 3 sets of difficulty levels (the  
1271 difference between motion coherence of the test and the reference stimulus; easy: [2.5,  
1272 5, 10, 20, 30], medium: [1.25, 2.5, 5, 10, 30], hard: [0.625, 1.25, 2.5, 5, 20]). As done in  
1273 the original study [9], we binned the trials into 8 levels by merging the trials of two  
1274 neighboring coherence levels (e.g., the coherence levels of [0.625, 1.25]) into a single  
1275 bin. Note that the coherence bins of [20, 35, 45, 48.75, 51.25, 55, 65, 80] (%s1) on the  
1276 x-axis (50% represents the equal coherence between s1 and s2) are matched to the x-  
1277 axis in Figure 8 of the previous study in which the same dataset had been used.  
1278 Asterisks mark the significance of one-sample *t*-tests (uncorrected  $P<0.05$ , one-tailed in  
1279 the direction of feedback effects) on the panel *toi+1* (stimulus 80%:  $t(26) = 2.0138$ ,  $P =$   
1280 0.0272) and on the panel *subtracted* (stimulus 20%:  $t(26) = -3.1900$ ,  $P = 0.0018$ ,  
1281 stimulus 80%:  $t(26) = 3.8810$ ,  $P = 0.0003$ ). **(B)** Sequential features of human observers  
1282 (N=8) published in another previous study [31]. We used the human dataset openly  
1283 available as part of the repository (<https://osf.io/hux4n>). In this study, the participants  
1284 performed a binary classification task on the speed of vibrotactile stimuli by classifying  
1285 the speed of the presented vibration as 'low-speed (weak)' or 'high-speed (strong)'.  
1286 Note that the nine-level stimuli of [-4, -3, -2, -1, 0, 1, 2, 3, 4] on the x-axis followed how  
1287 data were encoded by the original study [31]. Asterisks mark the significance of one-  
1288 sample *t*-tests (uncorrected  $P<0.05$ , one-tailed in the direction of feedback effects) on  
1289 the panel *toi+1* (stimulus -4:  $t(7) = -3.6757$ ,  $P = 0.004$ , stimulus -3:  $t(7) = -3.5252$ ,  $P =$   
1290 0.0048, and stimulus -2:  $t(7) = -2.0325$ ,  $P = 0.04$ ) and on the panel *subtracted*  
1291 (stimulus -4:  $t(7) = -1.9848$ ,  $P = 0.044$ ). The data underlying this figure **(A, B)** can be  
1292 found in S1 Data.  
1293 (TIF)

1294  
1295

## 1296 **S1 Appendix. Supporting details**

1297 **Supplemental details (Text) on additional model specifications of BMBU** are  
1298 provided. **Supplementary tables (A-D Tables) to support the Results section** are  
1299 provided. **Table A. Parameters used for *ex ante* simulations. Table B. Parameters**  
1300 **recovered from fitting the main models, world-updating and value-updating**  
1301 **models, to human choices (N=30). Table C. Parameters recovered from fitting the**  
1302 **rest of the models to human choices (N=30). Table D. Statistical results on model**  
1303 **behavior versus human behavior in terms of PSE measures.**

1304 (DOCX)

1305

1306 **S1 Data. Excel spreadsheet containing, in separate sheets, the underlying**  
1307 **numerical data for Figs 2D, 4B, 4D, 4E, 4G, 4H, 5B, 5D, 5E, 5G, 5H, 6B, 6C, 6D, 7A,**  
1308 **7B, 7C, 7D, 7E, 7F, S2A, S2B, S3, S4A, S4B, S5, S6A, S6B, S7A, and S7B.**

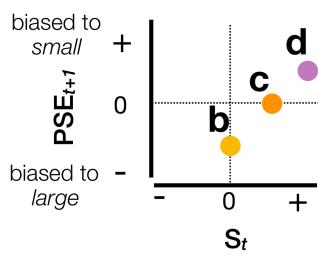
1309 (XLSX)

1310

1311 **S2 Data. Excel spreadsheet containing detailed statistical information comparing**  
1312 **alternative PSE estimation methods.**

1313 (XLSX)

**A**

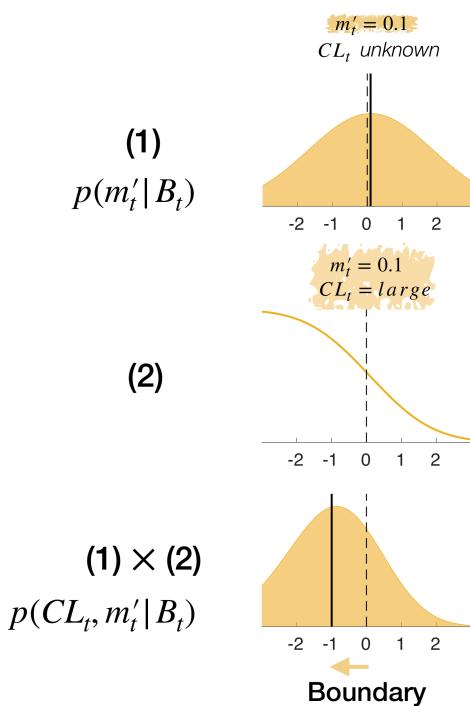


**Boundary likelihood,**

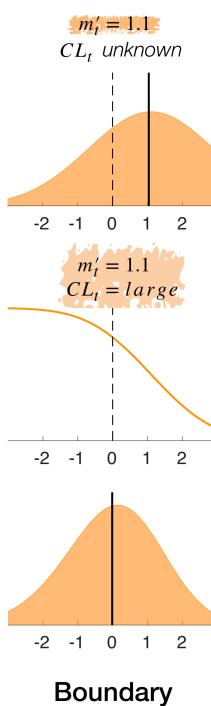
$$p(m'_t, CL_t = \text{large} | B_t)$$

$$= \int_{B_t}^{+\infty} p(m'_t | S_t) p(S_t | B_t) dS_t = (1) \times (2)$$

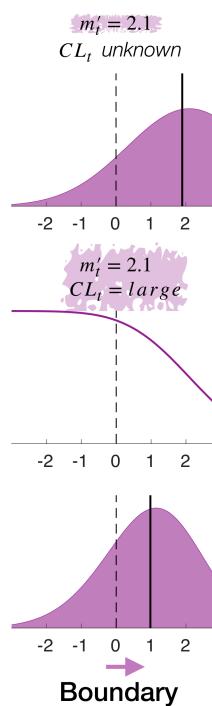
**B**



**C**



**D**

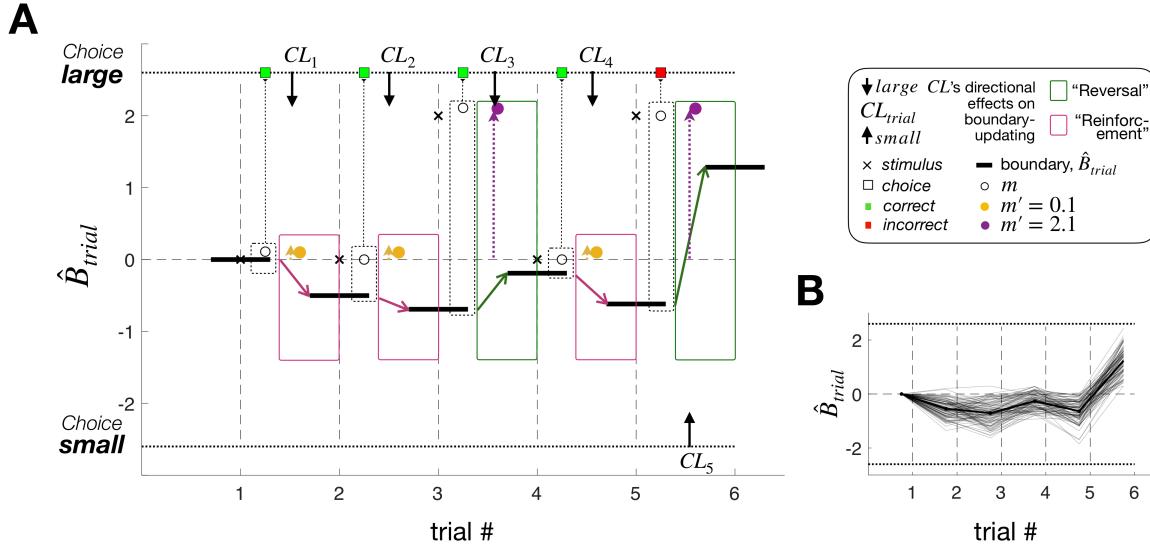


1314

1315 **S1 Fig. Schematic illustration of BMBU's account of how the joint contribution of**  
 1316 **the sensory and feedback evidence to boundary updating leads to the reversal of**  
 1317 **choice bias as a function of sensory evidence strength.**

1318 **(A)** Reversal of subsequent choice bias—expressed in PSE—as a function of sensory  
 1319 evidence strength and boundary inference—expressed in likelihood computation—  
 1320 based on a PDM episode. Left panel: The circles with different colors (indicated by (b-  
 1321 d), which points to the corresponding panels below (B-D)) represent the PSEs  
 1322 associated with the boundary updating for three example PDM episodes, where the  
 1323 stimulus ( $S_t$ ) varies from 0 to 2 while the choice ( $C_t$ ) and feedback ( $F_t$ ) are *large* and  
 1324 *correct*, respectively. Right panel: At the core of boundary inference is the computation  
 1325 of the likelihood of the class boundary based on the mnemonic measurement ( $m'_t$ ) and  
 1326 the informed state of the class variable ( $CL_t$ ), where  $CL_t$  is jointly determined by  $F_t$  and  
 1327  $C_t$  (see **Materials and methods** for the full computation of boundary inference in  
 1328 BMBU). **(B-D)** The likelihoods of the class boundary given the three example PDM

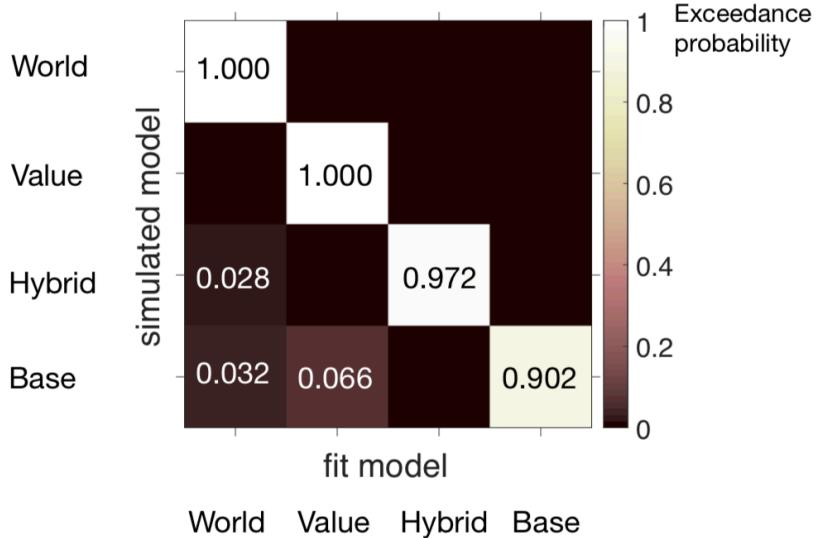
1329 episodes defined in **(A)**, where sensory evidence varies from the low **(B)**, to the  
1330 intermediate **(C)**, and to the high **(D)** level. To help understand why and how, given the  
1331 same feedback evidence, the direction of boundary updating reverses as the sensory  
1332 evidence strengthens, we visualize the boundary likelihoods as a product of two  
1333 functions (Equation 12), indicated by sub-panels marked as (1) and (2). Top row: As  
1334 indicated by (1), we plot each boundary likelihood when only the mnemonic  
1335 measurement is considered, assuming that no feedback is provided. Note that these  
1336 likelihood functions are centered around the values of  $m'_t$ , by attracting the class  
1337 boundary toward themselves, driving a shift towards the *large* side (i.e. positive side on  
1338 the boundary axis). Middle-Bottom rows: When the feedback evidence is given—i.e.,  
1339 when the informed state of  $CL_t$  is revealed as *large*—in addition to the mnemonic  
1340 measurement, an additional piece of information about the class boundary arises. As  
1341 indicated by (1)×(2), we plot each boundary likelihood (defined in **(A)**). As indicated by  
1342 (2), we plot each function (Middle row), as the result of (Bottom row) divided by (Top  
1343 row). The complementary cumulative distribution functions shown here are also  
1344 centered around  $m'_t$  because the *large* state of  $CL_t$  means that the class boundary is  
1345 located somewhere smaller than  $m'_t$ . Note that these skewed distributions push the  
1346 inferred class boundary away from the state of  $CL_t$  informed by feedback, driving a shift  
1347 towards the *small* side (i.e. negative side on the boundary axis). Consequently, the  
1348 influences from the sensory evidence and the feedback evidence counteract each other  
1349 (Bottom row). Note that the likelihood functions are centered in the *small* side when the  
1350 sensory evidence is weak **(B)**, in the neutral side when intermediate **(C)**, and in the  
1351 large side when strong **(D)**. These systematic shifts of the class-boundary likelihood as  
1352 a function of the strength of sensory evidence predict that the PSE of the psychometric  
1353 curve for the subsequent trial ( $t+1$ ) reverses its sign from negative to positive as a  
1354 function of the stimulus size, as shown in **(A)**.  
1355



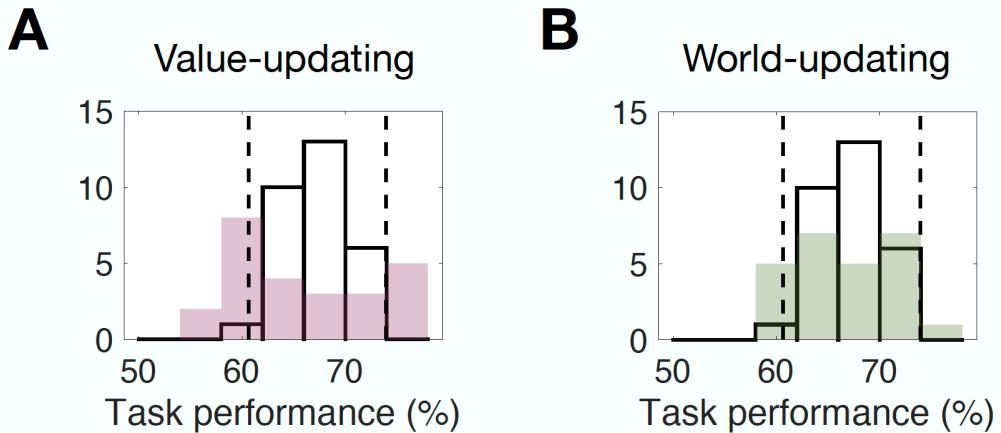
1356  
1357

1358 **S2 Fig. Example trial courses of estimated class boundary. (A)** An example trial  
 1359 history to show how a temporal trajectory of the class boundary inferred by BMBU. For  
 1360 example, at trial #1 (x-axis), a physical stimulus (symbol x) was 0, a sensory  
 1361 measurement (symbol o) was a positive value when the boundary belief (solid black  
 1362 bar; y-axis) was centered at 0. BMBU's choice was *large* (symbol square on the top of  
 1363 y-axis), and correct feedback (same square filled with green color) was provided, which  
 1364 indicates that the class variable at trial #1  $CL_1$  was *large* (arrow's direction indicates the  
 1365 effect of the trial class variable on the subsequent boundary-updating). BMBU updates  
 1366 one's belief based on evidence from stimulus (colored symbol o) and feedback ( $CL_1$ ),  
 1367 available at the time of boundary-updating. To illustrate cases where the bias reversal  
 1368 we defined in Fig 3D in the main text happen and do not happen, same examples were  
 1369 intentionally used as those we used in S1 Fig where we further detailed on the model's  
 1370 mechanisms. Depending on colors, sensory evidence is weak (yellow symbol o) or  
 1371 strong (purple symbol o), which leads to whether or not the reversal happens. Trial  
 1372 cases featured in a red box indicates that the "Reinforcement" principle is held  
 1373 (predicting subsequent choices to repeat *large* choice) while those featured in a green  
 1374 box indicates that the "Reversal" happens (predicting subsequent choices to reverse the  
 1375 previously made *large* choice). **(B)** Temporal trajectories of the class boundary when  
 1376 the same 6-trial sequence of physical stimuli in **(A)** was simulated for 100 times. This  
 1377 means that different  $m$  and  $m'$  were realized. The data underlying this figure **(A, B)** can  
 1378 be found in S1 Data.

1379  
1380



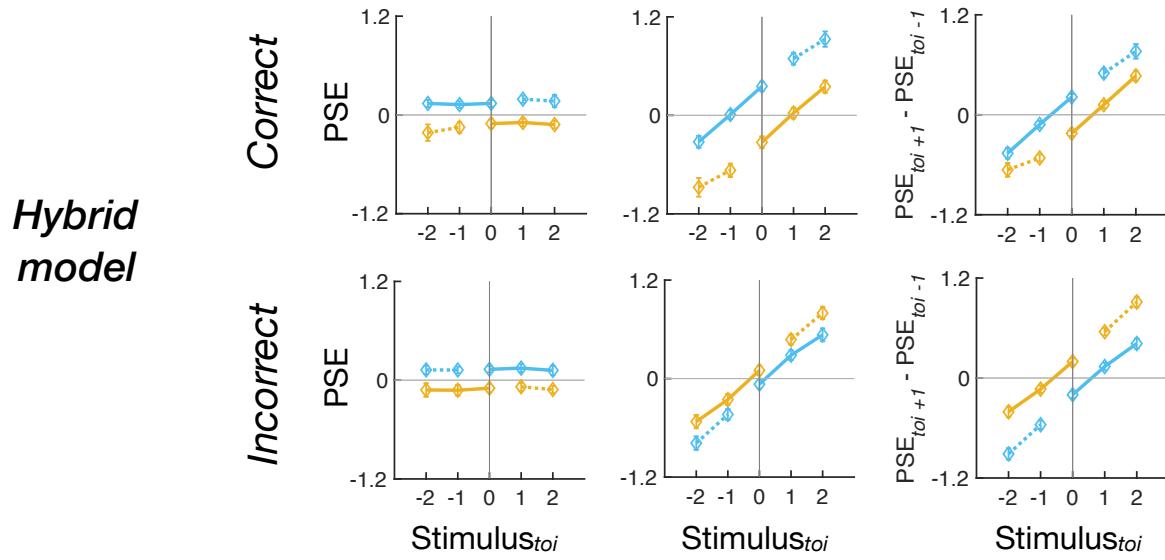
1381  
1382  
1383 **S3 Fig. Model recovery analysis.** Each square represents exceedance probability  $p_{exc}$   
1384 from model recovery procedure. The 'ground-truth' model to simulate synthetic behavior  
1385 was correctly recovered with  $p_{exc} > 0.9$  for all 4 models considered in the study. The light  
1386 shade of the diagonal squares indicates that the ground-truth model was the best-fitting  
1387 model, leading to a successful model recovery. Numerical values can also be found in  
1388 S1 Data.  
1389  
1390



1391

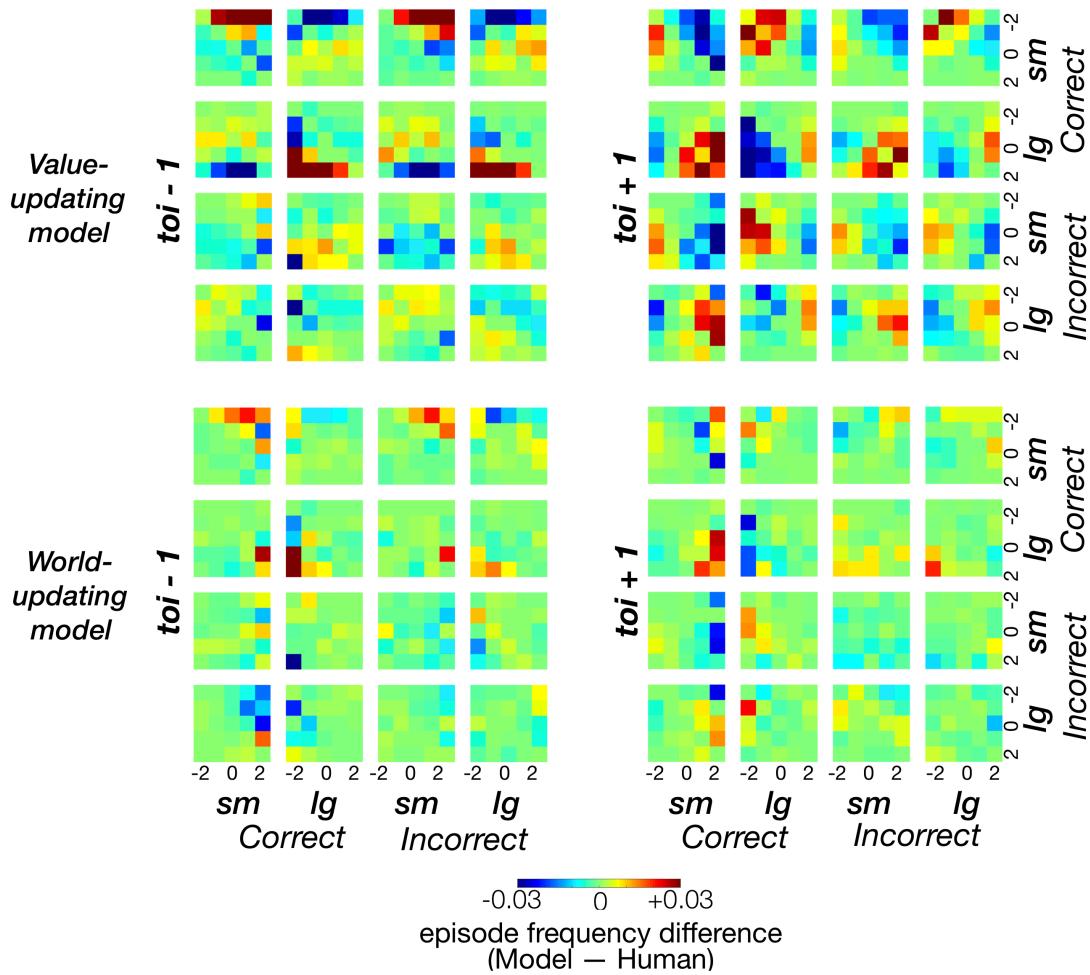
1392 **S4 Fig. Histograms of classification accuracies of the human participants and**  
1393 **their model partners in the *ex ante* simulations.**

1394 **(A, B)** Across-individual distributions of the classification accuracy of the belief-based  
1395 RL model **(A)** and BMBU **(B)** overlaid on those of the human participants. The models'  
1396 choices were generated via *ex ante* simulations with a specific set of model parameters  
1397 (Table A in S1 Appendix), the results of which are depicted in Fig 4 and Fig 5. The  
1398 classification accuracy is measured by calculating the percentage of the trials in which  
1399 the choice matched the feedback used in the actual experiment. The empty bars  
1400 correspond to the histogram of human performances, the range of which is demarcated  
1401 by the dashed vertical lines ( $[min, max]=[60.65\%, 73.94\%]$ ). The average human  
1402 classification accuracy was 67.85%. **(A)** Comparison of classification accuracy between  
1403 the belief-based RL model's simulation (red color) and the human choices. The model's  
1404 *ex ante* simulation accuracy was not different from the human accuracy ( $t(53) = 1.4429$ ,  
1405  $P = 0.1549$ ; Null hypothesis: model's performance vector and humans' performance  
1406 vector come from populations with equal means, unpaired two-tailed  $t$ -test). **(B)**  
1407 Comparison of classification accuracy between BMBU's simulation (green color) and  
1408 the human choices. The model's *ex ante* simulation accuracy was not different from the  
1409 human accuracy ( $t(53) = 0.9707$ ,  $P = 0.3361$ , unpaired two-tailed  $t$ -test). There was no  
1410 significant difference in classification accuracy between the value-updating model and  
1411 BMBU ( $t(48) = 0.5733$ ,  $P = 0.5691$ , unpaired two-tailed  $t$ -test). The data underlying this  
1412 figure **(A, B)** can be found in S1 Data.  
1413



1414  
1415 **S5 Fig. Retrospective (left columns), prospective (middle columns), and**  
1416 **subtractive (right columns) history effects in PSE for the 'Hybrid' model's ex post**  
1417 **model simulations.**  
1418 Top and bottom rows in each panel show the PSEs associated with the *toi* episodes  
1419 involving *correct* and *incorrect* feedback at *toi*. Symbols with error bars, mean±s.e.m  
1420 across the 30 model agents, which correspond to their 30 human partners. The colors  
1421 of the symbols and lines label choices (blue: *small* and yellow: *large*). The data  
1422 underlying this figure can be found in S1 Data.  
1423

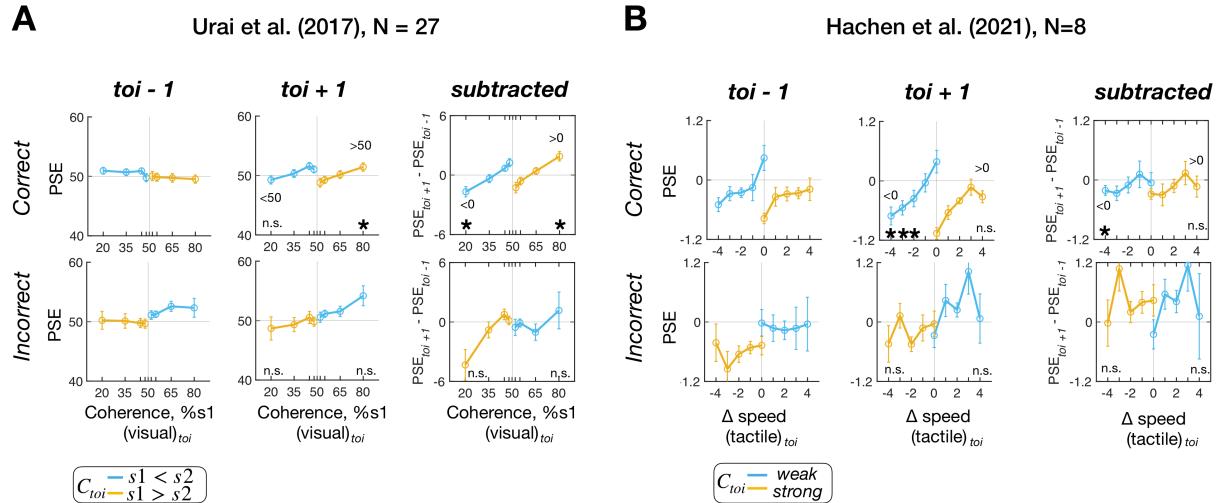
**A**



1424

1425 **S6 Fig. Maps of frequency deviations of the value-updating (A) and world-**  
1426 **updating (B) model agents' classifications in the ex post simulations from the**  
1427 **human decision-makers in the retrospective (left) and prospective (right) history**  
1428 **effects.**

1429 Each cell represents a pair of PDM episodes, as specified by the column and row  
1430 labels. At each cell, the color represents how much the episode frequency observed in  
1431 the model agents deviates from that observed in the corresponding human decision-  
1432 makers. The results of statistical tests on these deviations are summarized in Fig 7E  
1433 and 7F. The data underlying this figure (A, B) can be found in S1 Data.  
1434



1435  
1436

1437 **S7 Fig. Retrospective (left columns), prospective (middle columns), and**  
1438 **subtractive (right columns) history effects in PSE for the human classification**  
1439 **performances of Urai et al. (2017)'s work [37] (A) and Hachen et al. (2021)'s work**  
1440 **[31] (B).**

1441 **(A, B)** We downloaded both publicly available datasets, analyzed them in the same way  
1442 that we analyzed human observers in our work, and plotted the results in the same  
1443 format used for Fig 7A. Top and bottom rows in each panel show the PSEs associated  
1444 with the *toi* episodes involving *correct* and *incorrect* feedback. Symbols with error bars,  
1445 mean $\pm$ s.e.m. across human observers. The colors of the symbols and lines label  
1446 choices (blue: *small* and yellow: *large*). The overall patterns of the PSEs plotted here  
1447 appear similar to those plotted in Fig 7A, displaying the reversals in direction of  
1448 stimulus-dependent feedback effects. When the same statistical tests used in our work  
1449 were carried out, some of the data points at the stimuli with strong sensory evidence at  
1450 *toi* significantly deviated from zero in the direction opposite to the feedback effect  
1451 predicted by the value-updating scenario, as indicated by the asterisks. **(A)** Sequential  
1452 features of human observers (N=27) analyzed in our way from human dataset that once  
1453 had been published [37], which is openly available  
1454 (<http://dx.doi.org/10.6084/m9.figshare.4300043>), then analyzed in the previous study  
1455 [9]. In this study, the participants performed a binary classification task on the difference  
1456 in motion coherence by sorting the pairs of random-dot-kinematogram stimuli shown in  
1457 two intervals (s1 and s2) into one of the two classes ('s1<s2' vs. 's1>s2') over  
1458 consecutive trials. The presented stimuli were taken from 3 sets of difficulty levels (the  
1459 difference between motion coherence of the test and the reference stimulus; easy: [2.5,  
1460 5, 10, 20, 30], medium: [1.25, 2.5, 5, 10, 30], hard: [0.625, 1.25, 2.5, 5, 20]). As done in  
1461 the original study [9], we binned the trials into 8 levels by merging the trials of two  
1462 neighboring coherence levels (e.g., the coherence levels of [0.625, 1.25]) into a single  
1463 bin. Note that the coherence bins of [20, 35, 45, 48.75, 51.25, 55, 65, 80] (%s1) on the  
1464 x-axis (50% represents the equal coherence between s1 and s2) are matched to the x-  
1465 axis in Figure 8 of the previous study in which the same dataset had been used.  
1466 Asterisks mark the significance of one-sample *t*-tests (uncorrected  $P<0.05$ , one-tailed in  
1467 the direction of feedback effects) on the panel *toi+1* (stimulus 80%:  $t(26) = 2.0138$ ,  $P =$

1468 0.0272) and on the panel *subtracted* (stimulus 20%:  $t(26) = -3.1900$ ,  $P = 0.0018$ ,  
1469 stimulus 80%:  $t(26) = 3.8810$ ,  $P = 0.0003$ ). **(B)** Sequential features of human observers  
1470 ( $N=8$ ) published in another previous study [31]. We used the human dataset openly  
1471 available as part of the repository (<https://osf.io/hux4n>). In this study, the participants  
1472 performed a binary classification task on the speed of vibrotactile stimuli by classifying  
1473 the speed of the presented vibration as ‘low-speed (weak)’ or ‘high-speed (strong)’.  
1474 Note that the nine-level stimuli of  $[-4, -3, -2, -1, 0, 1, 2, 3, 4]$  on the x-axis followed how  
1475 data were encoded by the original study [31]. Asterisks mark the significance of one-  
1476 sample  $t$ -tests (uncorrected  $P < 0.05$ , one-tailed in the direction of feedback effects) on  
1477 the panel *toi+1* (stimulus  $-4$ :  $t(7) = -3.6757$ ,  $P = 0.004$ , stimulus  $-3$ :  $t(7) = -3.5252$ ,  $P$   
1478 = 0.0048, and stimulus  $-2$ :  $t(7) = -2.0325$ ,  $P = 0.04$ ) and on the panel *subtracted*  
1479 (stimulus  $-4$ :  $t(7) = -1.9848$ ,  $P = 0.044$ ). The data underlying this figure **(A, B)** can be  
1480 found in S1 Data.

1481  
1482

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<sup>1</sup>, Heeseung Lee<sup>1</sup>, Chae Young Lim<sup>2</sup>, Issac Rhim<sup>3</sup>, Sang-Hun Lee<sup>1</sup>

<sup>1</sup>Department of Brain and Cognitive Sciences, Seoul National University, Seoul, South Korea

<sup>2</sup>Department of Statistics, Seoul National University, Seoul, South Korea

<sup>3</sup>Institute of Neuroscience, University of Oregon, Eugene, OR, USA

\*Correspondence: [visionsl@snu.ac.kr](mailto: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  $p(m_t^i | S_t)$  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

1513 probability of  $m'_t$  if that hypothetical state is true. Here, we denote  $p(m'_t|S_t)$  by a likelihood function,

1514  $L_{m'_t}(S_t)$ .

1515 Through marginalization over  $m_t$  from the learned generative model, which is described by

1516 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,

1517 as follows:

$$1518 \begin{aligned} p(m'_t|S_t) &= \int p(m'_t|m_t)p(m_t|S_t)dm_t = \int N(m'_t; m_t, \sigma_{m'}^2)N(m_t; S_t, \sigma_m^2)dm_t \\ &= \int \frac{1}{\sqrt{2\pi\sigma_{m'}^2}} e^{-\frac{(m'_t-m_t)^2}{2\sigma_{m'}^2}} \frac{1}{\sqrt{2\pi\sigma_m^2}} e^{-\frac{(m_t-S_t)^2}{2\sigma_m^2}} dm_t = N(m'_t; S_t, \sigma_{m'}^2 + \sigma_m^2), \end{aligned} \quad (S2)$$

1519 where the expressions  $p(m'_t|m_t)p(m_t|S_t)$  are substituted by the two Gaussian noise distributions as

1520 defined in the learned generative model (Equation 5-6 in the main text) and integrated over the possible

1521 states of the sensory measurement variable  $m_t$ , which is now unknown—inaccessible any longer. Thus,

1522 we find that  $L_{m'_t}(S_t)$ , the likelihood of  $S_t$ , follows  $N(S_t; m'_t, \sigma_{m'}^2 + \sigma_m^2)$ .

1523 Next, the factor  $p(CL_t|S_t, B_t)$  in Equation S1 corresponds to the information about the stimulus

1524 and the class boundary offered by the state of the class variable  $CL_t$  informed by feedback, either *small*

1525 or *large*. The likelihood of any hypothetical joint states of  $S_t$  and  $B_t$  given  $CL_t$  refers to the probability of

1526  $CL_t$  if that hypothetical state is true. Here, we denote  $p(CL_t|S_t, B_t)$  by a likelihood function,  $L_{CL_t}(S_t, B_t)$ .

1527  $L_{CL_t}(S_t, B_t)$  is determined depending on the inequality between  $S_t$  and  $B_t$ : in the case of  $CL_t = small$ ,

$$1528 L_{CL_t=small}(S_t, B_t) = p(CL_t = small|S_t, B_t) = \begin{cases} 1, & S_t < B_t \\ 0, & S_t > B_t \\ 0.5, & S_t = B_t \end{cases} \quad (S3)$$

1529 ; in the case of  $CL_t = large$ ,

$$1530 L_{CL_t=large}(S_t, B_t) = p(CL_t = large|S_t, B_t) = \begin{cases} 0, & S_t < B_t \\ 1, & S_t > B_t \\ 0.5, & S_t = B_t \end{cases} . \quad (S4)$$

1531 Now, let us get back to Equation S1 and denote the boundary likelihood function in the case of

1532  $CL_t = \text{small}$  (abbreviated as  $s$ ) by  $L_{m'_t, CL_t=s}(B_t)$ , which can further be decomposed into two integrals

1533 with finite limits (as similarly done for Equation 9 in the main text), as follows:

1534 
$$L_{m'_t, CL_t=s}(B_t) \equiv p(m'_t, CL_t = s | B_t)$$

1535 
$$= \int_{S_t=-\infty}^{S_t=B_t} p(m'_t | S_t) p(CL_t = s | S_t, B_t) p(S_t | B_t) dS_t + \int_{S_t=B_t}^{S_t=+\infty} p(m'_t | S_t) p(CL_t = s | S_t, B_t) p(S_t | B_t) dS_t,$$

1536 which can be rewritten as follows:

1537 
$$= \lim_{r \rightarrow \infty, b \rightarrow B_t^-} \int_{-r}^b p(m'_t | S_t) p(CL_t = s | S_t, B_t) p(S_t | B_t) dS_t$$

1538 
$$+ \lim_{r \rightarrow \infty, b \rightarrow B_t^+} \int_b^r p(m'_t | S_t) p(CL_t = s | S_t, B_t) p(S_t | B_t) dS_t.$$

1539 (S5)

1540 Since the last term on the right-hand side of Equation S5 becomes zero by Equation S3 (for any  
1541 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  
1542  $CL_t = \text{small}$  is reduced as follows:

1543 
$$L_{m'_t, CL_t=s}(B_t) \equiv p(m'_t, CL_t = s | B_t) = \int_{-\infty}^{B_t} p(m'_t | S_t) p(S_t | B_t) dS_t. \quad (S6)$$

1544 From Equation 3 in the main text and Equation S2, we know the distributions  $p(S_t | B_t)$   
1545 and  $p(m'_t | S_t)$ , respectively. Substituting the expression for these distributions gives:

1546 
$$L_{m'_t, CL_t=s}(B_t)$$

1547 
$$= \int_{-\infty}^{B_t} \frac{1}{\sqrt{2\pi(\sigma_{m'}^2 + \sigma_m^2)}} e^{-\frac{(m'_t - S_t)^2}{2(\sigma_{m'}^2 + \sigma_m^2)}} \frac{1}{\sqrt{2\pi\sigma_S^2}} e^{-\frac{(S_t - B_t)^2}{2\sigma_S^2}} dS_t = \frac{1}{\sqrt{2\pi\left(\frac{\sigma_M^2\sigma_S^2}{(\sigma_M^2 + \sigma_S^2)}\right)}} \int_{-\infty}^{B_t} e^{-\frac{\left(S_t - \frac{B_t\sigma_M^2 + m'_t\sigma_S^2}{\sigma_M^2 + \sigma_S^2}\right)^2}{2\frac{\sigma_M^2\sigma_S^2}{(\sigma_M^2 + \sigma_S^2)}}} dS_t \times \frac{1}{\sqrt{2\pi(\sigma_M^2 + \sigma_S^2)}} e^{-\frac{(B_t - m'_t)^2}{2(\sigma_M^2 + \sigma_S^2)}},$$

1548 (S7)

1549 where  $\sigma_M^2 = \sigma_{m'}^2 + \sigma_m^2$ . Equation S7 is equivalent to Equation 11 in the main text.

1550 To aid in intuitive comprehension of  $L_{m'_t, CL_t=s}(B_t)$ , we can express it as the product of two terms  
1551 from Equation S7 by rewriting it as follows:

1552  $L_{m'_t, CL_t=s}(B_t) = F_X(B_t) \mathcal{L}_{m'_t}(B_t),$  (S8)

1553 where  $F_X$  denotes the first term on the right-hand side of Equation S7 (to be detailed in Equation S10),  
 1554 and  $\mathcal{L}_{m'_t}$  denotes the second term on the right-hand side of Equation S7 (to be detailed in Equation S9),  
 1555 which equals to the likelihood function for  $B_t$  given  $m'_t$  defined under the assumption that no  $CL_t$  variable  
 1556 exists in the generative process. To make this point explicit, we used a different likelihood notation  $\mathcal{L}$   
 1557 from the notation  $L$  used throughout the paper.

1558 Note that, according to the learned generative model in our study, the decision-maker acquires the  
 1559 knowledge about  $B_t$  by evaluating how probable the two observations,  $m'_t$  (sensory evidence in memory;  
 1560 simply as sensory evidence, hereinafter) and  $CL_t$  (feedback evidence), are for each possible value of  $B_t$ .  
 1561 Contrastingly,  $\mathcal{L}_{m'_t}(B_t)$  in Equation S8 can be considered the information about  $B_t$  solely based on the  
 1562 sensory evidence by computing  $p(m'_t|B_t)$ . Thus,  $\mathcal{L}_{m'_t}(B_t)$  represents the “sensory influence” that drives  
 1563 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  
 1564 (see the first row in S1B-D Fig), as follows:

1565 
$$\mathcal{L}_{m'_t}(B_t) = \frac{1}{\sqrt{2\pi(\sigma_M^2 + \sigma_S^2)}} e^{-\frac{(B_t - m'_t)^2}{2(\sigma_M^2 + \sigma_S^2)}} = N(B_t; m'_t, \sigma_M^2 + \sigma_S^2). \quad (S9)$$

1566 On the other hand, the term  $F_X(B_t)$  in Equation S8 can be considered the information about  $B_t$   
 1567 jointly based on the feedback and sensory evidence, while parting out the aforementioned “sensory  
 1568 influence” (divided by  $\mathcal{L}_{m'_t}(B_t)$ ). Let  $X$  be a random variable with cumulative distribution function (CDF)  
 1569  $F_X$

1570 
$$F_X(B_t) = \int_{-\infty}^{B_t} f_X(S_t) dS_t, \quad (S10)$$

1571 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  
 1572 is derived specifically for  $CL_t = \text{small}$ , we can interpret this term as the “feedback influence” that drives  
 1573 the boundary update with a CDF multiplied, by pushing the boundary likelihood  $L_{m'_t, CL_t=s}(B_t)$  in a more  
 1574 positive direction on the  $B_t$  axis. Owing to this contribution by the feedback evidence, the boundary

1575 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  
 1576  $b < m'_t$  for the current state of the class boundary that generated  $m'_t$  and  $CL_t = small$ . This aligns with  
 1577 the intuition from the temperature example described in the main text (Fig 1C).

1578 Similarly, from Equations S5-S7, we derive the boundary likelihood for the case  $CL_t = large$ :

1579 
$$L_{m'_t, CL_t=l}(B_t) = \int_{B_t}^{\infty} p(m'_t | S_t) p(S_t | B_t) dS_t = (1 - F_X(B_t)) \mathcal{L}_{m'_t}(B_t), \quad (S11)$$

1580 Contrary to the  $CL_t = small$  case, the boundary likelihood  $L_{m'_t, CL_t=l}(B_t = b)$  would support that  $b < m'_t$   
 1581 is more plausible than  $b > m'_t$  for  $B_t$  that generated  $m'_t$  and  $CL_t = large$ , since the multiplication is  
 1582 performed, instead with a complementary CDF (see the second row in S1B-D Fig).

1583  
 1584  
 1585  
 1586  
 1587  
 1588  
 1589

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

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

World		Value	
$\sigma_m$	25 levels [0.15, 3.27]	$\sigma_m$	25 levels [0.15, 3.27]
$\mu_0$	0	$\mu_0$	0
$\sigma_0$	5	$\alpha$	0.35
$\sigma_s$	1.5811	$\beta$	5
$\sigma_{m'}$	2.5	$V_{init}$	0.5
$\sigma_{diffusion}$	0.8		

1590  
 1591  
 1592  
 1593  
 1594

**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)
$\sigma_m$	0.86 (0.46)	$\sigma_m$	1.20 (0.30)
$\mu_0$	0.04 (0.62)	$\mu_0$	0.02 (0.43)
$\sigma_0$	4.27 (1.22)	$\alpha$	0.17 (0.20)
$\sigma_s$	0.84	$\beta$	7.43

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

1595  
1596  
1597  
1598

**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)
$\sigma_m$	0.68 (0.33)	$\sigma_m$	1.54 (0.39)	$\sigma_m$	1.90 (0.50)
$\mu_0$	0.01 (0.61)	$\mu_0$	0.02 (0.31)		
$\sigma_0$	3.55 (1.28)				
$\sigma_s$	0.86 (1.03)				
$\sigma_{m'}$	3.84 (1.22)				
$\sigma_{diffusion}$	3.77 (1.87)				
$\alpha$	0.15 (0.11)				
$\beta$	6.96 (2.83)				
$V_{init}$	0.882 (0.15)				

1599  
1600  
1601  
1602

**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)	
	n.s.		significant		n.s.		significant		n.s.		significant		n.s. # conditions	
Test type	<i>t</i>	<i>w</i>	<i>t</i>	<i>w</i>	<i>t</i>	<i>w</i>	<i>t</i>	<i>w</i>	<i>t</i>	<i>w</i>	<i>t</i>	<i>w</i>	<i>t</i>	<i>w</i>
Value	(13)	(13)	*(1) **(2) ***(4)	*(1) **(5) ***(1)	(10)	(9)	*(1) **(1) ***(8)	*(1) **(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)

1603  
1604

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      \*\*\*:  $P<0.0000167$