

Learning optimal decisions with confidence

Jan Drugowitsch^{1,*}, André G. Mendonça², Zachary F. Mainen², Alexandre Pouget³

¹ Department of Neurobiology, Harvard Medical School, 220 Longwood Avenue, Boston, MA 02115, USA

² Champalimaud Neuroscience Programme, Champalimaud Centre for the Unknown, Avenida de Brasília, s/n, 1400-038 Lisbon, Portugal

³ Department of Basic Neuroscience, University of Geneva, Rue Michel Servet 1, CH-1211 Geneva, Switzerland

* Corresponding author:

Jan Drugowitsch

Department of Neurobiology

Harvard Medical School

220 Longwood Avenue

Boston, MA 02115

Telephone: +1 (617) 432 5026

E-Mail: jan_drugowitsch@hms.harvard.edu

1 **Abstract**

2 Diffusion decision models (DDMs) are immensely successful models for decision-making under
3 uncertainty and time pressure. In the context of perceptual decision making, these models
4 typically start with two input units, organized in a neuron-antineuron pair. In contrast, in the brain,
5 sensory inputs are encoded through the activity of large neuronal populations. Moreover, while
6 DDMs are wired by hand, the nervous system must learn the weights of the network through trial
7 and error. There is currently no normative theory of learning in DDMs and therefore no theory of
8 how decision makers could learn to make optimal decisions in this context. Here, we derive the
9 first such rule for learning a near-optimal linear combination of DDM inputs based on trial-by-trial
10 feedback. The rule is Bayesian in the sense that it learns not only the mean of the weights but
11 also the uncertainty around this mean in the form of a covariance matrix. In this rule, the rate of
12 learning is proportional (resp. inversely proportional) to confidence for incorrect (resp. correct)
13 decisions. Furthermore, we show that, in volatile environments, the rule predicts a bias towards
14 repeating the same choice after correct decisions, with a bias strength that is modulated by the
15 previous choice's difficulty. Finally, we extend our learning rule to cases for which one of the
16 choices is more likely a priori, which provides new insights into how such biases modulate the
17 mechanisms leading to optimal decisions in diffusion models.

18

19 **Significance Statement**

20 Popular models for the tradeoff between speed and accuracy of everyday decisions usually
21 assume fixed, low-dimensional sensory inputs. In contrast, in the brain, these inputs are
22 distributed across larger populations of neurons, and their interpretation needs to be learned from
23 feedback. We ask how such learning could occur and demonstrate that efficient learning is
24 significantly modulated by decision confidence. This modulation predicts a particular dependency
25 pattern between consecutive choices, and provides new insight into how a priori biases for
26 particular choices modulate the mechanisms leading to efficient decisions in these models.

27

28 **Introduction**

29 Decisions are a ubiquitous component of every-day behavior. To be efficient, they require
30 handling the uncertainty arising from the noisy and ambiguous information that the environment
31 provides (1). This is reflected in the trade-off between speed and accuracy of decisions. Fast
32 choices rely on little information and may therefore sacrifice accuracy. In contrast, slow choices
33 provide more opportunity to accumulate evidence and thus may be more likely to be correct, but
34 are more costly in terms of attention or effort and lost time and opportunity. Therefore, efficient
35 decisions require not only a mechanism to accumulate evidence, but also one to trigger a choice
36 once enough evidence has been collected. *Drift-diffusion models* (or *diffusion decision models*;
37 DDMs) are a widely-used model family (2) that provides both mechanisms. Not only do DDMs
38 yield surprisingly good fits to human and animal behavior (3–5), but they are also known to
39 achieve a Bayes-optimal decision strategy under a wide range of circumstances (4, 6–10).

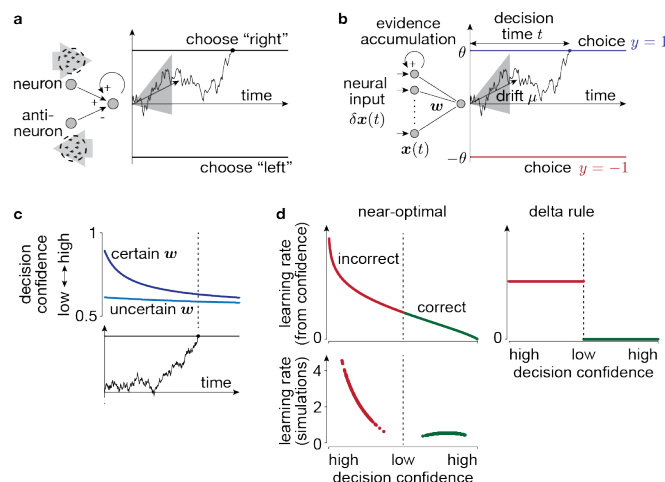
40 DDMs assume a particle that drifts and diffuses until it reaches one of two boundaries,
41 each triggering a different choice (**Fig. 1a**). The particle's drift reflects the net surplus of evidence
42 towards one of two choices. This is exemplified by the random-dot motion task, in which the
43 motion direction and coherence set the drift sign and magnitude. The particle's stochastic diffusion
44 reflects the uncertainty in the momentary evidence and is responsible for the variability in decision
45 times and choices widely observed in human and animal decisions (3, 5). A standard assumption
46 underlying DDMs is that the noisy momentary evidence that is accumulated over time is one-
47 dimensional — an abstraction of the momentary decision-related evidence of some stimulus. In
48 reality, however, evidence would usually be distributed across a larger number of inputs, such as
49 a neural population in the brain, rather than individual neurons (or neuron/anti-neuron pairs; **Fig.**
50 **1a**). Furthermore, the brain would not know a priori how this distributed encoding provides
51 information about the correctness of either choice. As a consequence, it needs to learn how to
52 interpret neural population activity from the success and failure of previous choices. How such an

53 interpretation can be efficiently learned over time, both normatively and mechanistically, is the
 54 focus of this work.

55 The multiple existing computational models for how humans and animals might learn to
 56 improve their decisions from feedback (e.g., 11–14) do not address the question we are asking,
 57 as they all assume that all evidence for each choice is provided at once, without considering the
 58 temporal aspect of evidence accumulation. This is akin to fixed-duration experiments, in which
 59 the evidence accumulation time is determined by the environment rather than the decision maker.
 60 We, instead, address a more general and natural case in which decision times are under the
 61 decision maker's control. In this setting, commonly studied using "reaction time" paradigms, the
 62 temporal accumulation of evidence needs to be treated explicitly, and – as we will show – the
 63 time it took to accumulate this evidence impacts how the decision strategy is updated after
 64 feedback. Some models for both choice and reaction times have addressed the presence of high-
 65 dimensional inputs (e.g., 15–17). However, they usually assumed as many choices as inputs,
 66 were mechanistic rather than normative, and did not consider how interpreting the input could be
 67 learned. We furthermore extend on previous work by considering the effect of a priori biases
 68 towards believing that one option is more correct than the other, and how such biases can be
 69 learned. This yields a new theoretical understanding of how choice biases impact optimal
 70 decision-making in diffusion models. Furthermore, it clarifies of how different implementations of
 71 this bias result in different diffusion model implementations, like the one proposed by Hanks et al.
 72 (18).

73

74 Results



75

76 **Figure 1. Learning the input weights from feedback in diffusion models.** In diffusion models, the input(s) provide at each point in
 77 time noisy evidence about the world's true state, here given by the drift μ . The decision maker accumulates this evidence over time
 78 (e.g., black example traces) to form a belief about μ . Bayes-optimal decisions choose according to the sign of the accumulated

79 evidence, justifying the two decision boundaries that trigger opposing choices. (a) In standard diffusion models, the momentary
 80 evidence either arises directly from noisy samples of μ , or, as illustrated here, from a neuron/anti-neuron pair that codes for opposing
 81 directions of evidence. The illustrated example assumes a random dot task, in which the decision maker needs to identify if most of
 82 the dots that compose the stimulus are moving either to the left or to the right. The two neurons (or neural pools) are assumed to
 83 extract motion energy of this stimulus towards the right (top) and left (bottom), such that their difference forms the momentary evidence
 84 towards rightward motion. A decision is made once the accumulated momentary evidence reaches one of two decision boundaries,
 85 triggering opposing choices. (b) Our setup differs from that in (a) in that we assume the input information $\delta x(t)$ to be encoded in a
 86 larger neural population whose activity is linearly combined with weights \mathbf{w} to yield the one-dimensional momentary evidence, and
 87 that the decision maker aims to learn these weights from feedback about the correctness of her choices. (c) Decision confidence (i.e.,
 88 the belief that the made choice was correct) in this kind of diffusion model drop as a function of time (horizontal axis) and with increased
 89 uncertainty about the input weights (different shades of blue). (d) For near-optimal learning, the learning rate (the term ξ_w in Eq. (6))
 90 is modulated by decision confidence (top left). High-confidence decisions lead to little learning if correct (green, right), and strong
 91 learning if incorrect (red, left). Low-confidence decisions result in a moderate confidence-related learning rate term (top, center). The
 92 learning rate in 1000 simulated trials (bottom) shows that the overall learning rate preserves this trend, with an additional suppression
 93 of learning for low-confidence decisions. Other learning heuristics (e.g., the delta rule, right) do not modulate their learning by
 94 confidence.
 95

96 **Bayes-optimal decision-making with diffusion models**

97 A standard way (8, 10, 19) to interpret diffusion models as mechanistic implementations of Bayes-
 98 optimal decision-making is to assume that, in each trial, an unobservable latent state μ (called
 99 *drift rate* in diffusion models) is drawn from a prior distribution, $\mu \sim N(0, \sigma_\mu^2)$, with zero mean and
 100 variance σ_μ^2 . The decision maker's aim is to infer whether this latent state is positive or negative
 101 (e.g., rightward vs. leftward motion in the random dot motion task), irrespective of its magnitude
 102 (e.g., the dot coherence level). The latent state itself is not directly observed, but is indirectly
 103 conveyed via a stream of noisy, momentary evidence values $\delta z_1, \delta z_2, \dots$, that, in each small time
 104 step of size δt , provide independent and identically distributed noisy information about μ through
 105 $\delta z_i | \mu \sim N(\mu \delta t, \delta t)$. Here, we have chosen a unit variance, scaled by δt . Any re-scaling of this
 106 variance by an additional parameter would result in a global re-scaling of the evidence that can
 107 be factored out (4, 8, 20), thus making such a re-scaling unnecessary.

108 Having after some time $t \equiv n\delta t$ observed n pieces of such evidence, $\delta z_{1:n}$, the decision
 109 maker's posterior belief about μ , $p(\mu | \delta z_{1:n})$, turns out to be fully determined by the accumulated
 110 evidence, $z(t) = \sum_{i=1}^n \delta z_i$, and time t (see Methods). Then, the posterior belief about μ being
 111 positive (e.g., leftward motion) results in (8)

112

$$p(\mu \geq 0 | z(t), t) = \int_0^\infty p(\mu | \delta z_{1:n}) d\mu = \Phi\left(\frac{z(t)}{\sqrt{t + \sigma_\mu^{-2}}}\right), \quad (1)$$

113

114 where $\Phi(\cdot)$ is the cumulative function of a standard Gaussian. The opposite belief about μ being
115 negative is simply $p(\mu < 0|z(t), t) = 1 - p(\mu \geq 0|z(t), t)$ (**Fig. 3a**). The accumulated evidence
116 follows a diffusion process, $z(t)|\mu \sim N(\mu t, t)$, and thus can be interpreted as the location of a
117 drifting and diffusing particle with drift μ and unit diffusion variance (**Fig. 1a**). By Eq. (1), the
118 posterior belief about $\mu \geq 0$ is $> 1/2$ for positive $z(t)$, and $< 1/2$ for negative $z(t)$. To make Bayes-
119 optimal decisions, Bayesian decision theory (21) requires that these decisions are chosen to
120 maximize the expected associated reward (or, more formally, to minimize the expected loss).
121 Assuming equally-rewarding correct choices, this implies choosing the option that is considered
122 more likely correct. Given the above posterior belief, this makes $y = \text{sign}(z(t)) \in \{-1, 1\}$ the
123 Bayes-optimal choice, which can be implemented mechanistically by (possibly time-varying)
124 boundaries $\pm\theta(t)$ on $z(t)$, associated with the two choices. At these boundaries, the posterior
125 belief about having made the correct choice, or decision confidence (22), is then given by Eq. (1)
126 with $z(t)$ replaced by $\theta(t)$. The sufficient statistics, $z(t)$ and t , of this posterior remain unchanged
127 by the introduction of such decision boundaries, such that Eq. (1) remains valid even in the
128 presence of these boundaries (8). Thus, under the above assumptions of prior and evidence,
129 diffusion models implement the Bayes-optimal decision strategy (**Fig. 3b**).

130 Note that $|\mu|$ (i.e., the momentary evidence's signal-to-noise ratio) controls the amount of
131 information provided about the sign of μ , and thus the difficulty of individual decisions. Thus, the
132 used prior $\mu \sim N(0, \sigma_\mu^2)$, which has more mass on small $|\mu|$, reflects that the difficulty of decisions
133 varies across trials, and that harder decisions are more frequent than easier ones. The prior width,
134 σ_μ^2 determines the spread of μ 's across trials, and therefore the overall difficulty of the task (larger
135 σ_μ^2 = overall easier task). We chose a Gaussian prior for mathematical convenience, and also
136 because hard trials are more frequent than easy ones in many experiments (e.g. (20)), even
137 though they don't commonly use Gaussian priors. In general, the important assumption is that the
138 difficulty varies across trials, but not exactly how it does so, which is to say that the shape of the
139 prior distribution is not critical (8). Different prior choice will not qualitatively change our results,
140 but would make it hard or impossible to derive interpretable closed-form expressions. Model
141 predictions would change qualitatively if we assume the difficulty to be fixed, or known a-priori
142 (see 8), but we will not consider this case, as it rarely if ever occurs in the real world.

143

144 **Using high-dimensional diffusion model inputs**

145 To extend diffusion models to multi-dimensional momentary evidence, we assume it to be given
146 the a k -dimensional vector $\delta\mathbf{x}_i$. This evidence might represent inputs from multiple sensors, or

147 the (abstract) activity of a neuronal population (**Fig. 1b**). As the activity of neurons in a population
148 that encodes limited information about the latent state μ is likely correlated across neurons (23,
149 24), we chose the momentary evidence statistics to also feature such correlations (see Methods).
150 In general, we choose these statistics such that $\mathbf{w}^T \delta \mathbf{x}_i = \delta z_i$, where the vector \mathbf{w} denote the k
151 input weights (for now assumed known). Defining the high-dimensional accumulated evidence by
152 $\mathbf{x}(t) = \sum_{i=1}^n \delta \mathbf{x}_i$, this implies $z(t) = \mathbf{w}^T \mathbf{x}(t)$, such that it is again Bayes-optimal to trigger decisions
153 as soon as $\mathbf{w}^T \mathbf{x}(t)$ equals one of two decision boundaries $\pm \theta(t)$. Furthermore, the posterior belief
154 about $\mu \geq 0$ is, similar to Eq. (1), given by

$$p(\mu \geq 0 | \mathbf{w}, \mathbf{x}(t), t) = \Phi(\mathbf{w}^T \tilde{\mathbf{x}}(t)), \quad (2)$$

156
157 where we have defined the time-attenuated accumulated evidence $\tilde{\mathbf{x}}(t) = \mathbf{x}(t) / \sqrt{t + \sigma_\mu^{-2}}$. As a
158 consequence, the decision-confidence for either choice, is, as before, given by Eq. (2), with
159 $\mathbf{w}^T \tilde{\mathbf{x}}(t)$ replaced by $\theta(t) / \sqrt{t + \sigma_\mu^{-2}}$. For time-independent decision bounds, $\theta(t) = \theta$, this
160 confidence decreases over time (**Fig. 1c**), reflecting the uncertainty about μ , and that late choices
161 are likely due to a low μ , which is associated with a hard trial, and thus low decision confidence.
162 This counter-intuitive drop in confidence with time has been previously described for diffusion
163 models with one-dimensional inputs (8, 25), and is a consequence of a trial difficulty that varies
164 across trials. Specifically, it arises from a mixture of easy trials associated with large $|\mu|$ that lead
165 to rapid, high-confidence choice, and hard trials associated with small $|\mu|$ that lead to slow, low-
166 confidence choices. Therefore, it doesn't depend on our choice of Gaussian prior, but is present
167 for any choice of symmetric prior over μ (see SI). The confidence remains constant over time only
168 when the difficulty is fixed across trials (i.e., $\mu \in \{-\mu_0, \mu_0\}$ for some fixed μ_0).

169

170 **Using feedback to find the posterior weights**

171 So far we have assumed the decision maker knows the linear input weights \mathbf{w} to make Bayes-
172 optimal choices. If they were not known, how could they be learned? Traditionally, learning has
173 been considered an optimization problem, in which the decision maker tunes some decision-
174 making parameters (here, the input weights \mathbf{w}) to maximize their performance. Here we will
175 instead consider it as an inference problem in which the decision maker aims to identify the
176 decision-making parameters that are most compatible with the provided observations. These two
177 views are not necessarily incompatible. For example, minimizing the mean squared error of a

178 linear model (an optimization problem) yields the same solution as sequential Bayesian linear
179 regression (an inference problem) (26). In fact, as we show in the SI, our learning problem can
180 also be formulated as an optimization problem. Nonetheless, we here take the learning-by-
181 inference route, as it provides a statistical interpretation of the involved quantities, which provides
182 additional insights. Specifically, we focus on learning the weights while keeping the diffusion
183 model boundaries fixed. The decision maker's reward rate (i.e., average number of correct
184 choices per unit time), which we use as our performance measure, depends on both weights and
185 the chosen decision boundaries. However, to isolate the problem of weight learning, we fix the
186 boundaries such that a particular set of optimal weights \mathbf{w}^* maximize this reward rate. The aim of
187 weight learning is to find these weights. Weight learning is a problem that needs to be solved
188 even if the decision boundaries are optimized at the same time. We have addressed how to best
189 tune these boundaries elsewhere (8, 27).

190 To see how learning can be treated as inference, consider the following scenario. Before
191 having observed any evidence, the decision maker has some belief, $p(\mathbf{w})$, about the input
192 weights, either as a prior or formed through previous experience. They now observe new
193 evidence, $\delta x_1, \delta x_2, \dots$ and use the mean of the belief over weights, $\langle \mathbf{w} \rangle$ (or any other statistics), to
194 combine this evidence and to trigger a choice y once the combined evidence reaches one of the
195 decision boundaries. Upon this choice, they receive feedback y^* about which choice was the
196 correct one. Then, the best way to update the belief about \mathbf{w} in light of this feedback is by Bayes'
197 rule,

$$198 \quad p(\mathbf{w}|\mathbf{x}(t), t, y^*) \propto p(y^*|\mathbf{w}, \mathbf{x}(t), t)p(\mathbf{w}), \quad (3)$$

199 where we have replaced the stream of evidence $\delta x_1, \delta x_2, \dots$ by the previously established
200 sufficient statistics $\mathbf{x}(t)$ and t .

202 The likelihood $p(y^*|\mathbf{w}, \mathbf{x}(t), t)$ expresses for any hypothetical weight vector \mathbf{w} the
203 probability that the observed evidence makes y^* the correct choice. To find its functional form,
204 consider that, for a known weight vector, we have shown that $p(\mu \geq 0|\mathbf{w}, \mathbf{x}(t), t)$, given by Eq. (2),
205 expresses the probability that $y = 1$ (associated with $\mu \geq 0$) is the correct choice. Therefore, $1 -$
206 $p(\mu \geq 0|\mathbf{w}, \mathbf{x}(t), t)$ corresponds to the probability that $y = -1$ (associated with $\mu < 0$) is the correct
207 choice. Therefore, it can act as the above likelihood function, which, by Eq.(2), is given by
208 $p(y^*|\mathbf{w}, \mathbf{x}(t), t) = \Phi(y^* \mathbf{w}^T \tilde{\mathbf{x}}(t))$, where we have used $1 - \Phi(a) = \Phi(-a)$. In summary, the decision
209 maker's belief is optimally updated after each choice by

210

$$p(\mathbf{w}|\mathbf{x}(t), t, y^*) \propto \Phi(y^* \mathbf{w}^T \tilde{\mathbf{x}}(t)) p(\mathbf{w}). \quad (4)$$

211
212 This update equation only requires knowing the accumulated evidence $\mathbf{x}(t)$, decision time t , and
213 feedback y^* , but is independent of the chosen option y , and how the decision maker came to this
214 choice. As a matter of fact, the decision maker could make random choices, irrespective of the
215 accumulated evidence, and still learn \mathbf{w} according to the above update equation, as long as they
216 keep track of $\mathbf{x}(t)$ and t , and acknowledge the feedback y^* . Therefore, learning and decision-
217 making aren't necessarily coupled. Nonetheless, we assume for all simulations that decision
218 makers perform decisions by using the mean estimate $\langle \mathbf{w} \rangle$, which is an intuitively sensible choice
219 if the decision maker's aim is to maximize their reward rate (see SI).

220 As in Eq. (4) the likelihood parameters, \mathbf{w} , are linear within a cumulative Gaussian
221 function, such problems are known as *Probit Regression* and don't have a closed-form expression
222 for the posterior. We could proceed by sampling from the posterior by Markov Chain Monte Carlo
223 methods, but that would not provide much insight into the different factors that modulate learning
224 the posterior weights. Instead, we proceed by deriving a closed-form approximation to this
225 posterior to provide such insight, as well as a potential mechanistic implementation.

226

227 **Confidence controls the learning rate**

228 To find an approximation to the posterior in Eq. (4), let us assume the prior to be given by the
229 Gaussian distribution, $p(\mathbf{w}) = N(\mathbf{w}|\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$, with mean $\boldsymbol{\mu}_w$ and covariance $\boldsymbol{\Sigma}_w$, which is the
230 maximum entropy distribution that specifies the mean and covariance (28). First, we investigated
231 how knowing \mathbf{w} with limited certainty, as specified by $\boldsymbol{\Sigma}_w$, impacts the decision confidence.
232 Marginalizing over all possible \mathbf{w} 's (see Methods) resulted in the choice confidence to be given
233 by

234

$$p(y|\mathbf{x}(t), t) = \Phi\left(\frac{y \boldsymbol{\mu}_w^T \tilde{\mathbf{x}}(t)}{\sqrt{1 + \tilde{\mathbf{x}}^T \boldsymbol{\Sigma}_w \tilde{\mathbf{x}}}}\right). \quad (5)$$

235

236 Compared to Eq. (2), the choice confidence is additionally attenuated by $\boldsymbol{\Sigma}_w$. Specifically, higher
237 weight uncertainty (i.e., an overall larger covariance $\boldsymbol{\Sigma}_w$) results in a lower decision confidence,
238 as one would intuitively expect (**Fig. 1c**).

239 Next, we found a closed-form approximation to the posterior, Eq. (4). For repeated
240 learning across consecutive decisions, the posterior over the weights after the previous decision

241 becomes the prior for the new decision. Unfortunately, a direct application of this principle would
242 lead to a posterior that changes its functional form after each update, making it intractable. We
243 instead used *Assumed Density Filtering* (ADF) (26, 29) that posits a fixed functional form
244 $q(\mathbf{w}|y^*, \mathbf{x}(t), t) = N(\mathbf{w}|\boldsymbol{\mu}_w^*, \boldsymbol{\Sigma}_w^*)$ of the posterior density – in our case Gaussian for consistency with
245 the prior – and then finds the posterior parameters $\boldsymbol{\mu}_w^*$ and $\boldsymbol{\Sigma}_w^*$ that make this approximate
246 posterior best match the “true” posterior $p(\mathbf{w}|y^*, \mathbf{x}(t), t)$, Eq. (4). Performing this match by
247 minimizing the Kullback-Leiber divergence $KL(p|q)$ results in the posterior mean (30, 31)
248

$$\boldsymbol{\mu}_w^* = \boldsymbol{\mu}_w + \frac{\xi_w}{\sqrt{1 + \tilde{\mathbf{x}}^T \boldsymbol{\Sigma}_w \tilde{\mathbf{x}}}} y^* \boldsymbol{\Sigma}_w \tilde{\mathbf{x}}, \quad (6)$$

249
250 and a similar expression for the posterior covariance (see Methods). Choosing $KL(p|q)$ to
251 measure the distance between p and q is to some degree arbitrary, but has beneficial properties,
252 such as that it causes the first two moments of q to match those of p (see SI). In Eq. (6), the factor
253 ξ_w modulates how strongly this mean is updated towards $y^* \boldsymbol{\Sigma}_w \tilde{\mathbf{x}}$, and turns out to be a
254 monotonically decreasing function of decision confidence (**Fig. 1d**, top; see Methods for
255 mathematical expression). For incorrect choices, for which the decision confidence is
256 $p(y^*|\mathbf{x}(t), t) < 1/2$, ξ_w is largest for choices made with high confidence, promoting significant
257 weight adjustments. For low-confidence choices it only promotes moderate adjustments, notably
258 irrespective of whether the choice was correct or incorrect. High-confidence, correct choices yield
259 a low ξ_w , and thus an intuitively minor strategy update. The update of the posterior covariance
260 follows a similar confidence-weighted learning rate modulation (**Fig. S1**; Methods).

261 Decision confidence is not the only factor that impacts the learning rate in Eq. (6). For
262 instance, $\tilde{\mathbf{x}}$ shrinks for longer, less confidence choices (because it is inversely proportional to
263 time) and results in overall less learning. Less certain weights, associated with larger magnitudes
264 of $\boldsymbol{\Sigma}_w$, have a similar effect. To investigate the overall impact of all of these factors combined on
265 the learning rate, we simulated a long sequence of consecutive choices and plotted the learning
266 rate for a random subset of these trials against the decision confidence (**Fig. 1d**, bottom). This
267 plot revealed a slight down-weighting of the learning rate for low-confidence choices when
268 compared to ξ_w , but left the overall dependency on ξ_w otherwise unchanged.

269

270 Performance comparison to optimal inference and to simpler heuristics

271 The intuitions provided by near-optimal ADF learning are only informative if its approximations do
272 not cause a significant performance drop. We quantified this drop by comparing ADF performance
273 to that of the Bayes-optimal rule, as found by Gibbs sampling (see Methods). Gibbs sampling is
274 biologically implausible as it requires a complete memory of inputs and feedbacks for past
275 decisions and is intractable for longer decision sequences, but nonetheless provides an optimal
276 baseline to compare against. We furthermore tested the performance of two additional
277 approximations. One was an ADF variant that assumes a diagonal covariance matrix Σ_w , yielding
278 a local learning rule that could be implemented by the nervous system. This variant furthermore
279 reduced the number of parameters from quadratic to linear in the size of w . The second was a
280 second-order Taylor expansion of the log-posterior, resulting in a learning rule similar to ADF, but
281 with a lower impact of weight uncertainty on the learning rate (see Methods).

282 Furthermore, we tested if simpler learning heuristics can match ADF performance. We
283 focused on three rules of increasing complexity. The *delta rule*, which can be considered a variant
284 of temporal-difference learning, or reinforcement learning (32), updates its weight estimate after
285 the n th decision by

286

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \frac{\alpha}{2\theta(0)} (y_n^* \theta(t) - \mathbf{x}_n(t)^T \mathbf{w}_n) \mathbf{x}_n(t), \quad (7)$$

287

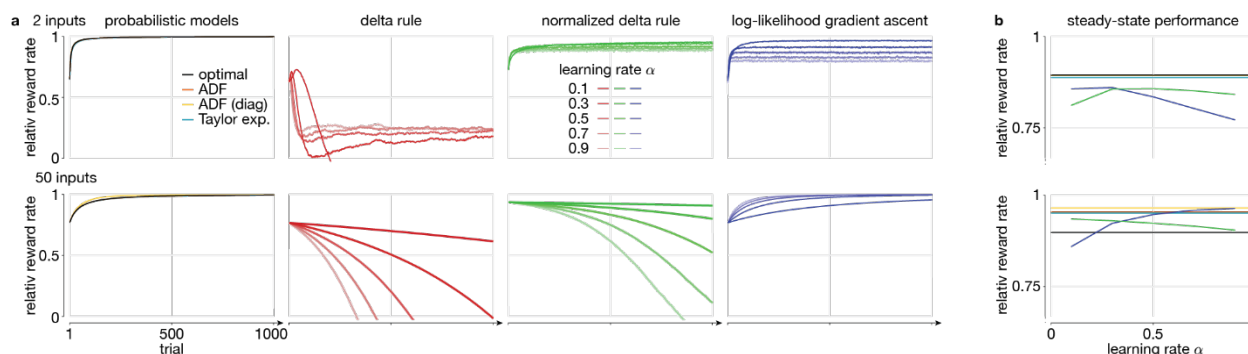
288 where $y_n^* \in \{-1, 1\}$ is the feedback about the correct choice provided after this decision, and we
289 have chosen to normalize the learning rate α by the initial bound height $\theta(0)$ to make it less
290 sensitive to this chosen height. As decisions are triggered at one of the two boundaries,
291 $\mathbf{x}_n(t)^T \mathbf{w}_n \in \{-\theta(t), \theta(t)\}$, the residual in brackets is zero for correct choices, and $\pm 2\theta(t)$ for
292 incorrect choices. As a result, and in contrast to ADF, weight adjustments are only performed
293 after incorrect choices, and with a fixed learning rate α rather than one modulated by confidence
294 (**Fig. 1d**; right). Our simulations revealed that the delta rule excessively and suboptimally
295 decrease in the weight size $\|\mathbf{w}\|$ over time, leading to unrealistically long reaction times and
296 equally unrealistic near-zero weights. To counteract this problem, we designed a *normalized delta*
297 *rule*, that updates the weight estimates as the delta rule, but thereafter normalizes them by $\mathbf{w} \leftarrow$
298 $\mathbf{w} \|\mathbf{w}^*\| / \|\mathbf{w}\|$ to ensure that its size matches that of the true weights \mathbf{w}^* . Access to these true
299 weights, \mathbf{w}^* , makes it an omniscient learning rule that can't be implemented by a decision maker

300 in practice. Lastly, we tested a learning rule that performs stochastic gradient ascent on the
 301 feedback log-likelihood,
 302

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \alpha \nabla_{\mathbf{w}} \log p(y_n^* | \mathbf{w}_n, \mathbf{x}_n(t), t) = \mathbf{w}_n + \alpha y_n^* \xi_{\mathbf{w}} \tilde{\mathbf{x}}_n(t). \quad (8)$$

303
 304 This rule introduces decision confidence weighting through $\xi_{\mathbf{w}}$, but differs from ADF in that it does
 305 not take the weight uncertainty ($\Sigma_{\mathbf{w}}$ in ADF) into account, and requires tuning of the learning rate
 306 parameter α .

307 We evaluated the performance of these learning rules by simulating weight learning
 308 across 1,000 consecutive decisions (called *trials*; see Methods for details) in a task in which use
 309 of the optimal weight vector maximizes the reward rate. This reward rate was the average reward
 310 for correct choices minus some small cost for accumulating evidence over the average time
 311 across consecutive trials and is a measure we would expect rational decision makers to optimize.
 312 For each learning rule we found its reward rate relative to random behavior and optimal choices.
 313



314
 315 **Figure 2. Input weight learning and tracking performance of different learning rules.** All plots show the relative reward rate (0 =
 316 immediate, random choices, 1 = optimal) averaged over 5,000 simulations with different true, underlying weights, and for 2 (top) and
 317 50 (bottom) inputs. (a) The relative reward rate for probabilistic and heuristic learning rules. The probabilistic learning rules include
 318 the optimal rule (Gibbs sampling), assumed density filtering (ADF), ADF with a diagonal covariance matrix (ADF (diag)), and a learning
 319 rule based on a second-order Taylor expansion of the log-posterior (Taylor exp.). For both 2 and 50 inputs, all rules perform roughly
 320 equally. For the heuristic rules, different color shadings indicate different learning rates. The initial performance shown is that *after* the
 321 first application of the learning rule, such that initial performances can differ across learning rule. (b) The steady-state performance
 322 across different heuristic rule learning rates. Steady state performance was measured as an average across 5,000 simulations,
 323 averaging over the last 100 of 1000 simulated trials in which the true weights slowly change across consecutive trials. An optimal
 324 relative reward rate of one corresponds to knowing the true weight in each trial, which, due to the changing weight, is not achievable
 325 in this setup. The color scheme is the same as in (a), but the vertical axis has a different scale. The delta rule did not converge and
 326 was not included in (b).

327
 328 **Figure 2a** shows this relative reward rate for all learning rules and different numbers of
 329 inputs. As can be seen, the performance of ADF and the other probabilistic learning rules is

330 indistinguishable from Bayes-optimal weight learning for all tested numbers of inputs. Surprisingly,
331 the ADF variant that ignores off-diagonal covariance entries even outperformed Bayes-optimal
332 learning for a large number of inputs (**Fig. 2a**, yellow line for 50 inputs). That reason that a simpler
333 learning rule could outperform the rule deemed optimal by Bayesian decision theory is that this
334 simpler rule has less parameters and a simpler underlying model that was nonetheless good
335 enough to learn the required weights. Learning fewer parameters with the same data resulted in
336 initially better parameter estimates, and better associated performance. Conceptually, this is
337 similar to a linear model outperforming a quadratic model when fitting a quadratic function if little
338 data is available, and if the function is sufficiently close to linear (as illustrated in Fig. S2). Once
339 more data is available, the quadratic model will outperform the linear one. Similarly, the Bayes-
340 optimal learning rule will outperform the simpler one once more feedback has been observed. In
341 our simulation, however, this does not occur within the 1000 simulated trials.

342 All other learning heuristics performed significantly worse. For low-dimensional input, the
343 delta rule initially improved its reward rate but worsens it again at a later stage across all learning
344 rates. The normalized delta rule avoided such performance drops for low-dimensional input, but
345 both delta rule variants were unable to cope with high-dimensional inputs. Only stochastic
346 gradient ascent on the log-likelihood provided a stable learning heuristic for high dimensional
347 inputs, but with the downside of having to choose a learning rate. Small learning rates lead to
348 slow learning, and an associated slower drop in angular error. Overall, the probabilistic learning
349 rules significantly outperformed all tested heuristic learning rules and matched (and in one case
350 even exceeded) the weight learning performance of the Bayes-optimal estimator.

351

352 **Tracking non-stationary input weights**

353 So far, we have tested how well our weight learning rule is able to learn the true, underlying
354 weights from binary feedback about the correctness of the decision maker's choices. For this we
355 assumed that the true weights remained constant across decisions. What would happen if these
356 weights change slowly over time? Such a scenario could occur if, for example, the world around
357 us changes slowly, or if the neural representation of this world changes slowly through neural
358 plasticity or similar. In this case, the true weights would become a moving target that we would
359 never be able to learn perfectly. Instead, we would after some initial transient expect to reach
360 steady-state performance that remains roughly constant across consecutive decisions. We
361 compared this steady-state performance of Bayes-optimal learning (now implemented by a
362 particle filter) to that of the probabilistic and heuristic learning rules introduced in the previous

363 section. The probabilistic rules were updated to take into account such a trial-by-trial weight
364 change, as modeled by a first-order autoregressive process (see Methods). The heuristic rules
365 remained unmodified, as their use of a constant learning rate already encapsulates the
366 assumption that the true weights change across decisions.

367 **Figure 2b** illustrates the performance of the different learning rules. First, it shows that,
368 for low-dimensional inputs the various probabilistic models yield comparable performances, but
369 for high-dimensional inputs the approximate probabilistic learning rules outperform Bayes-optimal
370 learning. In case of the latter, these approximations weren't actually harmful, but instead
371 beneficial, for the same reason discussed further above. In particular, the more neurally-realistic
372 ADF variant that only tracked the diagonal of the covariance matrix again outperformed all other
373 probabilistic models. Second, only the heuristic learning rule that performed gradient ascent on
374 the log-likelihood achieved steady-state performance comparable to the approximate probabilistic
375 rules, and then only for high input dimensionality and a specific choice of learning rate. This should
376 come as no surprise, as its use of the likelihood function introduces more task structure
377 information than the other heuristics use. The delta rule did not converge and therefore never
378 achieved steady-state performance. Overall, the ADF variant that focused only on the diagonal
379 covariance matrix achieved the best overall performance.

380

381 **Learning both weights and a latent state prior bias**

382 Our learning rule can be generalized to learn prior biases in addition to the input weights. The
383 prior we have used so far for the latent variable, $\mu \sim N(0, \sigma_\mu^2)$, is unbiased, as both $\mu \geq 0$ and $\mu <$
384 0 are equally likely. To introduce a prior bias, we instead used $\mu \sim N(m, \sigma_\mu^2)$, where m controls
385 the bias through $P^+ \equiv p(\mu \geq 0) = \Phi(m/\sigma_\mu)$. A positive (or negative) m causes $P^+ > 1/2$ (or $<$
386 $1/2$), thus making $y = 1$ (or $y = -1$) the more likely correct choice even before evidence is
387 accumulated. After evidence accumulation, such a prior results in the posterior

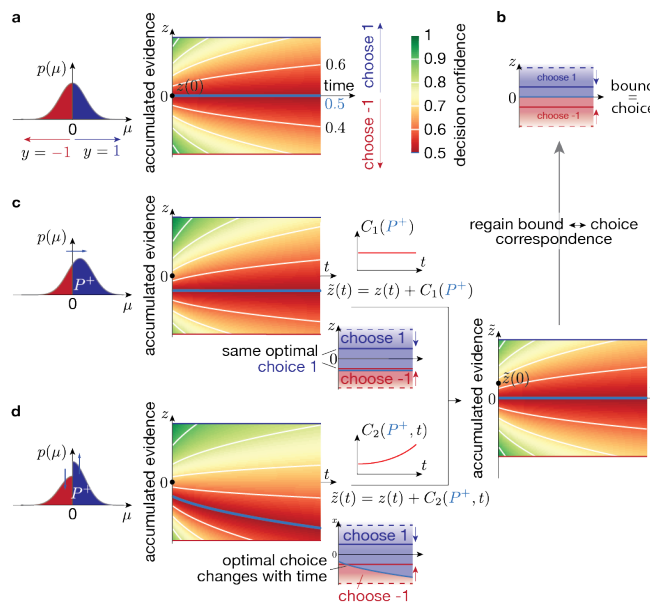
388

$$p(\mu \geq 0 | \mathbf{w}, \mathbf{x}(t), t) = \Phi\left(\frac{\mathbf{w}^T \mathbf{x}(t) + \sigma_\mu^{-2} m}{\sqrt{t + \sigma_\mu^{-2}}}\right). \quad (9)$$

389

390 Comparing this to the unbiased posterior Eq. (2) reveals the additional term $\sigma_\mu^{-2} m$ whose relative
391 influence wanes over time.

392



393
 394 **Figure 3. Decision confidence, prior biases, and the relation between decision boundary and choice.** (a) For an unbiased prior
 395 (i.e., $P^+ \equiv p(\mu \geq 0) = 1/2$), the decision confidence (color gradient) is symmetric around $z = 0$ for each fixed time t . The associated
 396 posterior belief $p(\mu \geq 0|z(t), t)$ (numbers above/below “time” axis label; constant along white lines; $1/2$ along light blue line) promote
 397 choosing $y = 1$ and $y = -1$ above (blue area in (b)) and below (red area in (b)) $z = 0$. (b) As a result, different choices are Bayes-
 398 optimal at the blue/red decision boundaries, as long as they are separated by $z = 0$, irrespective of the boundary separation (solid vs.
 399 dashed blue red lines). (c) If the prior is biased by an overall shift, the decision confidence is counter-shifted by the same constant
 400 across all t . In this case, both decision boundaries might promote the same choice, which can be counter-acted by a time-invariant
 401 shift of z by $C_1(P^+)$. (d) If the prior is biased by boosting one side while suppressing the other, the decision confidence shift becomes
 402 time-dependent, such that the optimal choice at a time-invariant boundary might change over time. Counteracting this effect requires
 403 a time-dependent shift of z by $C_2(P^+, t)$. In both (c) and (b) we have chosen $P^+ = 0.6$, for illustration.

404
 405 This additional term has two consequences. First, appending the elements m and σ_μ^{-2} to
 406 the vectors \mathbf{w} and $\mathbf{x}(t)$, respectively, shows that \mathbf{w} and m can be learned jointly by the same
 407 learning rule we have derived before (Methods). Second, the term requires us to rethink the
 408 association between decision boundaries and choices. As **Fig. 3c** illustrates, such a prior causes
 409 a time-invariant shift in the association between the accumulated evidence, $z(t) = \mathbf{w}^T \mathbf{x}(t)$, and
 410 the posterior belief of $\mu \geq 0$ and corresponding decision confidence. This shift makes it possible
 411 to have the same Bayes-optimal choice at both decision boundaries (**Fig. 3c**, blue/red decision
 412 areas). Hence, we have lost the mechanistically convenient unique association between decision
 413 boundaries and choices. We recover this association by a boundary counter-shift, such that these
 414 boundaries come to lie at the same decision confidence levels for opposite choices, making them
 415 asymmetric around $z = 0$. Mathematically, this is equivalent to shifting the evidence accumulation
 416 starting point, $\tilde{z}(0)$ away from zero in the opposite direction (**Fig. 3c**, shift by $C_1(P^+) = \sigma_\mu^{-2} m$; SI).
 417 Therefore, a prior bias is implemented by a bias-dependent simple shift of the accumulation

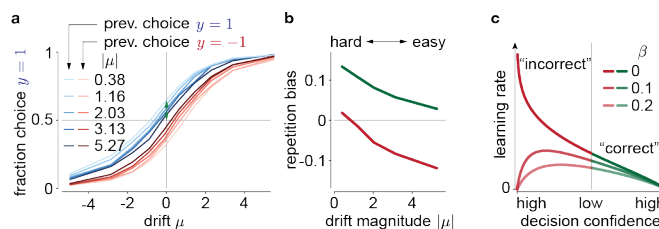
418 starting point, leading to a mechanistically straight-forward implementation of Bayes-optimal
 419 decision-making with biased priors.

420 A consequence of the shifted accumulation starting point is that, for some fixed decision
 421 time t , the decision confidence at both boundaries is the same (**Fig. 3c** right). This seems at odds
 422 with the intuition that a biased prior ought to bias the decision confidence in favor of the more
 423 likely option. However, this mechanism does end up assigning higher average confidence to the
 424 more likely option because of reactions times. As the starting point is now further away from the
 425 less likely correct boundary, it will on average take longer to reach this boundary, which lowers
 426 the decision confidence since confidence decreases with elapsed time. Therefore, even though
 427 the decision confidence at both boundaries is the same for the given decision time, it will on
 428 average across decision times be lower for the a-priori non-preferred boundary, faithfully
 429 implementing this prior (see SI for a mathematical demonstration).

430 Our finding that a simple shift in the accumulation starting point is the Bayes-optimal
 431 strategy appears at odds with previous work that suggested that the optimal shift of the
 432 accumulator variable $z(t)$ varies with time (18). This difference stems from a different
 433 implementation of the bias. While we have chosen an overall shift in the prior by its mean (**Fig.**
 434 **3c**), an alternative implementation is to multiply $p(\mu \geq 0)$ by P^+ , and $p(\mu < 0)$ by $1 - P^+$ (**Fig. 3d**),
 435 again resulting in $P^+ = p(\mu \geq 0)$. A consequence of this difference is that the associated shift of
 436 the posterior belief of $\mu \geq 0$ in the evidence accumulation space becomes time-dependent. Then,
 437 the optimal choice at a time-invariant boundary in that space might change over time (**Fig. 3d**).
 438 Furthermore, un-doing this shift to regain a unique association between boundaries and choices
 439 not only requires a shifted accumulation starting point, but additionally a time-dependent additive
 440 signal ($C_2(P^+, t)$ in **Fig. 3d**; SI), as was proposed in (18). Which of the two approaches is more
 441 adequate depends on how well it matches the prior implicit in the task design. Our approach has
 442 the advantage of a simpler mechanistic implementation, as well as yielding a simple extension to
 443 the previously derived learning rule. How learning prior biases in the framework of (18) could be
 444 achieved remains unclear (but see (33)).

445

446 Sequential choice dependencies due to continuous weight tracking



447

448 **Figure 4. Sequential choice dependencies due to continuous learning, and effects of noisy feedback.** Bayes-optimal learning
449 in a slowly changing environment predicts sequential choice dependencies with the following pattern. (a) After hard, correct choices
450 (low prev. $|\mu|$; light colors), the psychometric curve is shifted towards repeating the same choice (blue/red = choice $y = 1/-1$). This
451 shift decreases after easier, correct choices (high prev. $|\mu|$; dark colors). (b) We summarize these tuning curve shifts in the repetition
452 bias, which is the probability of repeating the same choice to a $\mu = 0$ stimulus (example green arrow for $\mu = -0.38$ in (a)). After
453 correct/incorrect choices (green/red curve), this leads to a win-stay/lose-switch strategy. Only the win-stay strategy is shown in (a).
454 (c) If choice feedback is noisy (inverted with probability β), the learning rate becomes overall lower. In particular for high-confidence
455 choices with “incorrect” feedback, the learning rate becomes zero, as the learner trusts her choice more than the feedback.

456

457 In every-day situations, no two decisions are made under the exact same circumstances.
458 Nonetheless, we need to be able to learn from the outcome of past choices to improve future
459 ones. A common assumption is that past choices become increasingly less informative about
460 future choices over time. One way to express this formally is to assume that the world changes
461 slowly over time – and that our aim is to track these changes. By ‘slow’ we mean that we can
462 consider it constant over a single trial but that it is unstable over the course of an hour-long
463 session. We implemented this tracking of the moving world, as in **Fig. 2b**, by slowly allowing the
464 weights mapping evidence to decisions to change. With such continuously changing weights,
465 weight learning never ends. Rather, the input weights are continuously adjusted to make correct
466 choices more likely in the close future. After correct choices, this means that weights will be
467 adjusted to repeat the same choice upon observing a similar input in the future. After incorrect
468 choices, the aim is to adjust the weights to perform the opposite choice, instead. Our model
469 predicts that, after an easy correct choice, in which confidence can be expected to be high, the
470 weight adjustments are lower than after hard correct choices (see **Fig. 1d** top, green line). As a
471 consequence, we would expect the model to be more likely to repeat the same choices after
472 correct and hard, than after correct and easy trials.

473 To test this prediction, we relied on the same simulation to generate **Fig. 2b** to measured
474 how likely the model repeated the same choice after correct decisions. **Figure 4a** illustrates that
475 this repetition bias manifests itself in a shift of the psychometric curve that makes it more likely to
476 repeat the previous choice. Furthermore, and as predicted, this shift is modulated by the difficulty
477 of the previous choice and is stronger if the previous choice was easy (i.e., associated with a large
478 $|\mu|$; **Fig. 4b**). Therefore, if the decision maker expects to operate in a volatile, slowly changing
479 world, our model predicts a repetition bias to repeat the same choices after correct decisions, and
480 that this bias is stronger if the previous choice was easy.

481

482 **Unreliable feedback reduces learning**

483 What would occur if choice feedback is less-than-perfectly reliable? For example, the feedback
484 itself might not be completely trustworthy, or hard to interpret. We simulated this situation by
485 assuming that the feedback is inverted with probability β . Here, $\beta = 0$ implies the so far assumed
486 perfectly reliable feedback, and $\beta = 1/2$ makes the feedback completely uninformative. This
487 change impacts how decision confidence modulates the learning rate (**Fig. 4c**) as follows. First,
488 it reduces the overall magnitude of the correction, with weaker learning for higher feedback noise.
489 Second, it results in no learning for highly confident choices that we are told are incorrect. In this
490 case, one's decision confidence overrules the unreliable feedback. This stands in stark contrast
491 to the optimal learning rule for perfectly reliable feedback, in which case the strongest change to
492 the current strategy ought to occur.

493

494 **Discussion**

495 Diffusion models are applicable to model decisions that require some accumulation of evidence
496 over time, which is almost always the case in natural decisions. We extended previous work on
497 the normative foundations of these models to more realistic situations in which the sensory
498 evidence is encoded by a population of neurons, as opposed to just two neurons, as has been
499 typically assumed in previous studies. We have focused on normative and mechanistic models
500 for learning the weights from the sensory neurons to the decision integrator without additionally
501 adjusting the decision boundaries, as weight learning is a problem that needs to be solved even
502 if the decision boundaries are optimized at the same time.

503 From the Bayesian perspective, weight learning corresponds to finding the weight
504 posterior given the provided feedback, and resulted in an approximate learning rule whose
505 learning rate was strongly modulated by decision confidence. It suppressed learning after high-
506 confidence correct decisions, supported learning for uncertain decisions irrespective of their
507 correctness, and promoted strong change of the combination weights after wrong decisions that
508 were made with high confidence (**Fig. 1d**). Evidence for such confidence-based learning has
509 already been identified in human experiments (34), but not in a task that required the temporal
510 accumulation of evidence in individual trials. Indeed, as we have previously suggested (22), such
511 a modulation by decision confidence should arise in all scenarios of Bayesian learning in N-AFC
512 tasks in which the decision maker only receives feedback about the correctness of their choices,
513 rather than being told which choice would have been correct. In the 2-AFC task we have

514 considered, being told that one's choice was incorrect automatically reveals that the other choice
515 was correct, making the two cases coincide. Moving from one-dimensional to higher-dimensional
516 inputs requires performing the accumulation of evidence for each input dimension separately (**Fig.**
517 **1b**; Eqs. (6) & (12) require $x(t)$ rather than only $w^T x(t)$), even if triggering choices only requires
518 a linear combination of $x(t)$. This is because uncertain input weights require keeping track of how
519 each input dimension contributed to the particle crossing the decision boundary in order to
520 correctly improve these weights upon feedback (i.e., proper credit assignment). The multi-
521 dimensional evidence accumulation predicted by our work arises naturally if inputs encode full
522 distributions across the task-relevant variables, such as in linear probabilistic population codes
523 (35) that trigger decisions by bounding the pooled activity of all units that represent the
524 accumulated evidence (36).

525 Continual weight learning predicts sequential choice dependencies that make the
526 repetition of a previous, correct choice more likely, in particular if this choice was difficult (**Fig. 4**).
527 Thus, based on assuming a volatile environment that promotes a continual adjustment of the
528 decision-making strategy, we provide a rational explanation for sequential choice dependencies
529 that are frequently observed in both humans and animals (e.g., 37, 38). In rodents making
530 decisions in response to olfactory cues we have furthermore confirmed that these sequential
531 dependencies are modulated by choice difficulty, and that the exact pattern of this modulation
532 depends on the stimulus statistics, as predicted by our theory (39) (but consistency with (40)
533 unclear).

534 Lastly, we have clarified how prior biases ought to impact Bayes-optimal decision-making
535 in diffusion models. Extending the work of Hanks et al. (18), we have demonstrated that the exact
536 mechanisms to handle these biases depend on the specifics of how these biases are introduced
537 through the task design. Specifically, we have suggested a variant that simplifies these
538 mechanisms and the learning of this bias. This variant predicts that the evidence accumulation
539 offset, that has previously been suggested to be time-dependent, to become independent of time,
540 and it would be interesting to see if LIP activity of monkeys performing the random-dot motion
541 task, as recorded by Hanks et al. (but see (41)), would change accordingly.

542

543 **Materials and Methods**

544 We here provide an outline of the framework and its results. Detailed derivations are provided in
545 the SI.

546 *Bayesian decision-making with one and multi-dimensional diffusion models*

547 We assume the latent state to be drawn from $\mu \sim N(m, \sigma_\mu^2)$, and the momentary evidence in each
 548 time step δt to provide information about this latent state by $\delta z_i | \mu \sim N(\mu \delta t, \delta t)$. The aim is to infer
 549 the sign of μ , and choose $y = 1$ if $\mu \geq 0$, and $y = -1$ otherwise. After having observed this
 550 evidence for some time $t \equiv n \delta t$, the posterior μ given all observed evidence $\delta z_{1:n}$ is by Bayes'
 551 rule given by

$$p(\mu | \delta z_{1:n}) \propto N(\mu | m, \sigma_\mu^2) \prod_{i=1}^n N(\delta z_i | \mu \delta t, \delta t) \propto N\left(\mu \frac{\sigma_\mu^{-2} m + z(t)}{\sigma_\mu^{-2} + t}, \frac{1}{\sigma_\mu^{-2} + t}\right). \quad (11)$$

553
 554 In the above, all proportionalities are with respect to μ , and we have defined $z(t) = \sum_{i=1}^n \delta z_i$ and
 555 have used $t = \sum_{i=1}^n \delta t$. How to find the posterior belief for about μ 's sign with $m = 0$ is described
 556 around Eq. (1).

557 We extend diffusion models to multi-dimensional inputs with momentary evidence
 558 $\delta \mathbf{x}_i | \mu, \mathbf{w} \sim N((\mathbf{a}\mu + \mathbf{b})\delta t, \mathbf{\Sigma}\delta t)$, with \mathbf{a} , \mathbf{b} and $\mathbf{\Sigma}$ chosen such that $\mathbf{w}^T \mathbf{x}(t) | \mu = z(t) | \mu \sim N(\mu t, t)$, as
 559 before. The posterior over μ and $\mu \geq 0$ is the same as for the one-dimensional case, with $z(t)$
 560 replaced by $\mathbf{w}^T \mathbf{x}(t)$. Defining $\tilde{\mathbf{x}}(t) = \mathbf{x}(t) / \sqrt{\sigma_\mu^{-2} + t}$, we find $p(\mu \geq 0 | \mathbf{w}, \mathbf{x}(t), t) = \Phi(\mathbf{w}^T \tilde{\mathbf{x}}(t))$. As
 561 $y = 1$ and $y = -1$ correspond to $\mu \geq 0$ and $\mu < 0$, and $y = 1$ is only chosen if $p(\mu \geq 0 | \mathbf{w}, \mathbf{x}(t), t) \geq$
 562 $1/2$, the decision confidence for $m = 0$ at some boundary $\mathbf{w}^T \mathbf{x}(t) = \pm \theta(t)$ is given by
 563 $\Phi\left(\theta(t) / \sqrt{\sigma_\mu^{-2} + t}\right)$. If input weights are unknown, and the decision maker holds belief $\mathbf{w} \sim$
 564 $N(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$ about these weights, the decision confidence needs to additionally account for weight
 565 uncertainty by marginalizing over \mathbf{w} , resulting in Eq. (5).

566 *Probabilistic and heuristic learning rules*

567 We find the approximate posterior $q(\mathbf{w}) = N(\mathbf{w} | \boldsymbol{\mu}_w^*, \boldsymbol{\Sigma}_w^*)$ that approximates the target posterior p
 568 Eq. (4) by Assumed Density Filter (ADF). This requires minimizing the Kullback-Leiber divergence
 569 $KL(p|q)$ (26, 29), resulting in Eq. (6) for the posterior mean, and

$$\boldsymbol{\Sigma}_w^* = \boldsymbol{\Sigma}_w + \xi_{cov}(\gamma) ((\boldsymbol{\Sigma}_w^{-1} + \tilde{\mathbf{x}} \tilde{\mathbf{x}}^T)^{-1} - \boldsymbol{\Sigma}_w), \quad (12)$$

571
 572 with learning rate modulators $\xi_w(\gamma) = N(\gamma | 0, 1) / \Phi(\gamma)$ and $\xi_{cov}(\gamma) = \xi_w(\gamma)^2 + \xi_w(\gamma)\gamma$, and where
 573 we have defined $\gamma \equiv y^* \boldsymbol{\mu}_w^T \tilde{\mathbf{x}} / \sqrt{1 + \tilde{\mathbf{x}}^T \boldsymbol{\Sigma}_w \tilde{\mathbf{x}}}$, which is monotonic in the decision confidence, Eq. (5).

574 Noisy choice feedback (**Fig. 4c**) changes the likelihood to assume reversed feedback with
575 probability β , and follow the same procedure as above to derive the posterior moments (see SI).
576 The ADF variant that only tracks the diagonal covariance elements assumes Σ_w to be diagonal,
577 and only computes the diagonal elements of Σ_w^* . A second-order Taylor expansion of the log of
578 Eq. (4) leads to update equations similar to Eqs. (6) and (12), but without the normalization by
579 weight uncertainty (see SI for details). All heuristic learning rules are described in the main text.

580 We modeled non-stationary input weights by $\mathbf{w}_n | \mathbf{w}_{n-1} \sim N(\mathbf{A}\mathbf{w}_{n-1} + \mathbf{b}, \Sigma_d)$ after a
581 decision in trial $n - 1$. This weight transition is taken into account by the probabilistic learning
582 rules by setting the parameter priors to $\mu_{w,n} = \mathbf{A}\mu_{w,n-1}^* + \mathbf{b}$ and $\Sigma_{w,n} = \mathbf{A}\Sigma_{w,n-1}^* \mathbf{A}^T + \Sigma_d$. For
583 stationary weights we have $\mathbf{A} = \mathbf{I}$, $\mathbf{b} = \mathbf{0}$, and $\Sigma_d = \mathbf{0}$.

584 Bayes-optimal weight inference for stationary weights performed by Gibbs sampling
585 for Probit models, and for non-stationary weights by particle filtering (see SI).

586 *Simulation details*

587 We used parameters $\mathbf{a} = \mathbf{w} / \|\mathbf{w}\|^2$ and $\mathbf{b} = \mathbf{0}$ for the momentary evidence δx . Its covariance Σ
588 was generated to feature eigenvalues that drop exponentially from $\sigma_x^2 = 2 / \|\mathbf{w}\|^2$ to zero until it
589 reaches a constant $\sigma_0^2 = 0.001 / \|\mathbf{w}\|^2$ noise baseline, as qualitatively observed in neural
590 populations. It additionally contains an eigenvector \mathbf{w} with eigenvalue set to guarantee $\mathbf{w}^T \Sigma \mathbf{w} =$
591 1, limiting the information that δx provides about μ . For non-stationary weights, all momentary
592 evidence parameters are adjusted after each weight change (see SI). The diffusion model bounds
593 $\pm\theta$ were time-invariant and tuned to maximize the reward rate when using the correct weights.
594 The reward rate is given by $(p(\text{correct}) - c_{\text{accum}} \langle t \rangle) / (t_{\text{iti}} + \langle t \rangle)$, where averages were across
595 trials, and we used evidence accumulation cost $c_{\text{accum}} = 0.01$ and inter-trial interval $t_{\text{iti}} = 2s$. We
596 used $\sigma_\mu^2 = 3^2$ to draw μ in each trial, and drew \mathbf{w} from $\mathbf{w} \sim N(\mathbf{1}, \mathbf{I})$ before each trial sequence. For
597 non-stationary weights, we re-sampled weight after each trial according to $\mathbf{w}_n | \mathbf{w}_{n-1} \sim N(\lambda \mathbf{w}_{n-1} +$
598 $(1 - \lambda), \sigma_d^2 \mathbf{I})$, with decay factor $\lambda = 1 - 0.01$ and $\sigma_d^2 = 1 - \lambda^2$ to achieve steady-state mean $\mathbf{1}$ and
599 identity covariance.

600 To compare the weight learning performance of ADF to alternative models (**Fig. 2a**), we
601 simulated 1,000 learning trials 5,000 times, and reported the reward rate per trial averaged across
602 these 5,000 repetitions. To assess steady-state performance (**Fig. 2b**), we performed the same
603 procedure with non-stationary weights, and reported reward rate averaged over the last 100 trials,
604 and over 5,000 repetitions. The same 100 trials were used to compute the sequential choice
605 dependencies in **Fig. 4a/b**. To simulate decision-making with diffusion models and uncertain
606 weights, we used the current mean estimate $\langle \mathbf{w} \rangle$ of the input weights to linearly combine the

607 momentary evidence. The probabilistic learning rules were all independent of the specific choice
608 of this estimate. The learning rate in **Fig. 1d** shows the pre-factor to $y^* \Sigma_w \tilde{x}$ in Eq. (6) over decision
609 confidence for a subsample of the last 10,000 trials of a single 15,000 trial simulation with non-
610 stationary weights. For the Gibbs sampler, we drew 10 burn-in samples, followed by 200 samples
611 in each trial. For the particle filter we simulated 1,000 particles.

612

613 **Acknowledgments**

614 This work was supported by a James S. McDonnell Foundation Scholar Award (#220020462,
615 JD), and grants from the NIMH (R01MH115554, JD), the Swiss National Science Foundation,
616 www.snf.ch, (#31003A_143707 and #31003A_165831, AP), Champalimaud Foundation (ZFM),
617 European Research Council (Advanced Investigator Grant 250334 & 67125, ZFM), Human
618 Frontier Science Program (Grant RGP0027/2010, ZFM & AP), Simons Foundation (Grant
619 325057, ZFM & AP), and Fundação para a Ciência e a Tecnologia (AGM).

620

621 **References**

- 622 1. Doya K, Ishii S, Pouget A, Rao RPN (2006) *Bayesian Brain: Probabilistic Approaches to*
623 *Neural Coding* (MIT Press).
- 624 2. Ratcliff R (1978) A theory of memory retrieval. *Psychol Rev* 85(2):59–108.
- 625 3. Ratcliff R, McKoon G (2008) The diffusion decision model: theory and data for two-choice
626 decision tasks. *Neural Comput* 20(4):873–922.
- 627 4. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD (2006) The physics of optimal
628 decision making: A formal analysis of models of performance in two-alternative forced-
629 choice tasks. *Psychol Rev* 113(4):700–765.
- 630 5. Ratcliff R, Smith PL (2004) A Comparison of Sequential Sampling Models for Two-Choice
631 Reaction Time. *Psychol Rev* 111(2):333–367.
- 632 6. Frazier PI, Yu AJ (2008) Sequential hypothesis testing under stochastic deadlines. *Adv*
633 *Neural Inf Process Syst*:1–8.
- 634 7. Tajima S, Drugowitsch J, Pouget A (2016) Optimal policy for value-based decision-making.
635 *Nat Commun* 7:12400.
- 636 8. Drugowitsch J, Moreno-Bote R, Churchland AK, Shadlen MN, Pouget A (2012) The cost of
637 accumulating evidence in perceptual decision making. *J Neurosci* 32(11):3612–3628.
- 638 9. Gold JI, Shadlen MN (2002) Banburismus and the brain: Decoding the relationship
639 between sensory stimuli, decisions, and reward. *Neuron* 36(2):299–308.

- 640 10. Drugowitsch J, Deangelis GC, Klier EM, Angelaki DE, Pouget A (2014) Optimal
641 multisensory decision-making in a reaction-time task. *Elife* 2014(3):1–19.
- 642 11. Dayan P, Kakade S, Montague PR (2000) Learning and selective attention. *Nat Neurosci*
643 3 Suppl(november):1218–1223.
- 644 12. Dayan P, Yu AJ (2003) Uncertainty and learning. *IETE J Res* 49(2–3):171–181.
- 645 13. Körding KP, Wolpert DM (2004) Bayesian integration in sensorimotor learning. *Nature*
646 427(6971):244–7.
- 647 14. Courville AC, Daw ND, Touretzky DS (2006) Bayesian theories of conditioning in a
648 changing world. *Trends Cogn Sci* 10(7):294–300.
- 649 15. Ratcliff R (1981) A theory of order relations in perceptual matching. *Psychol Rev*
650 88(6):552–572.
- 651 16. Gomez P, Ratcliff R, Perea M (2008) The overlap model: A model of letter position coding.
652 *Psychol Rev* 115(3):577–600.
- 653 17. Ratcliff R, Starns JJ (2013) Modeling confidence judgments, response times, and multiple
654 choices in decision making: Recognition memory and motion discrimination. *Psychol Rev*
655 120(3):697–719.
- 656 18. Hanks TD, Mazurek ME, Kiani R, Hopp E, Shadlen MN (2011) Elapsed decision time
657 affects the weighting of prior probability in a perceptual decision task. *J Neurosci*
658 31(17):6339–6352.
- 659 19. Moreno-Bote R (2010) Decision confidence and uncertainty in diffusion models with
660 partially correlated neuronal integrators. *Neural Comput* 22(7):1786–1811.
- 661 20. Palmer J, Huk AC, Shadlen MN (2005) The effect of stimulus strength on the speed and
662 accuracy of a perceptual decision. *J Vis* 5(5):376–404.
- 663 21. Berger JO (1993) *Statistical Decision Theory and Bayesian Analysis* (Springer). 2nd Editio.
- 664 22. Pouget A, Drugowitsch J, Kepecs A (2016) Confidence and certainty: distinct probabilistic
665 quantities for different goals. *Nat Neurosci* 19(3):366–374.
- 666 23. Averbeck BB, Latham PE, Pouget A (2006) Neural correlations, population coding and
667 computation. *Nat Rev Neurosci* 7(5):358–366.
- 668 24. Moreno-Bote R, et al. (2014) *Information-limiting correlations* (Nature Publishing Group)
669 doi:10.1038/nn.3807.
- 670 25. Kiani R, Shadlen MN (2009) Representation of Confidence Associated with a Decision by
671 Neurons in the Parietal Cortex. *Science (80-)* 324(5928):759–764.
- 672 26. Bishop C (2006) *Pattern Recognition and Machine Learning* (Springer).
- 673 27. Drugowitsch J, Deangelis GC, Angelaki DE, Pouget A (2015) Tuning the speed-accuracy

- 674 trade-off to maximize reward rate in multisensory decision-making. *Elife* 4(JUNE2015):1–
675 11.
- 676 28. Cover TM, Thomas JA (2006) *Elements of Information Theory* (Wiley). 2nd Editio.
- 677 29. Murphy KP (2012) *Machine Learning: a Probabilistic Perspective* (MIT Press).
- 678 30. Graepel T, Quiñero-Candela J, Borchert T, Herbrich R (2010) Web-Scale Bayesian
679 Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft’s Bing
680 Search Engine. *Proceedings of the 27th International Conference on Machine Learning*
681 *(ICML-10)* Available at:
682 [http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.165.5644&rep=rep1&](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.165.5644&rep=rep1&type=pdf)
683 [type=pdf%5Cnhttp://machinelearning.wustl.edu/mlpapers/paper_files/icml2010_](http://machinelearning.wustl.edu/mlpapers/paper_files/icml2010_GraepelC)
684 [BH10.pdf](http://machinelearning.wustl.edu/mlpapers/paper_files/icml2010_GraepelC).
- 685 31. Chu W, Zinkevich M, Li L, Thomas A, Tseng B (2011) Unbiased online active learning in
686 data streams. *Proceedings of the 17th ACM SIGKDD International Conference on*
687 *Knowledge Discovery and Data Mining - KDD '11* (ACM Press, New York, New York, USA),
688 p 195.
- 689 32. Sutton RS, Barto AG (2018) *Reinforcement learning: an introduction* (MIT Press). 2nd editio.
- 690 33. Zylberberg A, Wolpert DM, Shadlen MN (2018) Counterfactual Reasoning Underlies the
691 Learning of Priors in Decision Making. *Neuron* 99(5):1083-1097.e6.
- 692 34. Meyniel F, Dehaene S (2017) Brain networks for confidence weighting and hierarchical
693 inference during probabilistic learning. *Proc Natl Acad Sci* 114(19):E3859–E3868.
- 694 35. Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic
695 population codes. *Nat Neurosci* 9(11):1432–8.
- 696 36. Beck JM, et al. (2008) Probabilistic Population Codes for Bayesian Decision Making.
697 *Neuron* 60(6):1142–1152.
- 698 37. Busse L, et al. (2011) The Detection of Visual Contrast in the Behaving Mouse. *J Neurosci*
699 31(31):11351–11361.
- 700 38. Yu AJ, Cohen JD (2009) Sequential effects: Superstition or rational behavior? *Adv Neural*
701 *Inf Process Syst* 21:1873–80.
- 702 39. Mendonça AG, et al. (2019) The impact of learning on perceptual decisions and its
703 implication for speed-accuracy tradeoffs. *bioRxiv*:1–64.
- 704 40. Urai AE, Braun A, Donner TH (2017) Pupil-linked arousal is driven by decision uncertainty
705 and alters serial choice bias. *Nat Commun* 8:14637.
- 706 41. Rao V, DeAngelis GC, Snyder LH (2012) Neural correlates of prior expectations of motion
707 in the lateral intraparietal and middle temporal areas. *J Neurosci* 32(29):10063–74.

708

709 **Figure legends**

710 **Figure 1. Learning the input weights from feedback in diffusion models.** In diffusion models,
711 the input(s) provide at each point in time noisy evidence about the world's true state, here given
712 by the drift μ . The decision maker accumulates this evidence over time (e.g., black example
713 traces) to form a belief about μ . Bayes-optimal decisions choose according to the sign of the
714 accumulated evidence, justifying the two decision boundaries that trigger opposing choices. (a)
715 In standard diffusion models, the momentary evidence either arises directly from noisy samples
716 of μ , or, as illustrated here, from a neuron/anti-neuron pair that codes for opposing directions of
717 evidence. The illustrated example assumes a random dot task, in which the decision maker needs
718 to identify if most of the dots that compose the stimulus are moving either to the left or to the right.
719 The two neurons (or neural pools) are assumed to extract motion energy of this stimulus towards
720 the right (top) and left (bottom), such that their difference forms the momentary evidence towards
721 rightward motion. A decision is made once the accumulated momentary evidence reaches one of
722 two decision boundaries, triggering opposing choices. (b) Our setup differs from that in (a) in that
723 we assume the input information $\delta x(t)$ to be encoded in a larger neural population whose activity
724 is linearly combined with weights w to yield the one-dimensional momentary evidence, and that
725 the decision maker aims to learn these weights from feedback about the correctness of her
726 choices. (c) Decision confidence (i.e., the belief that the made choice was correct) in this kind of
727 diffusion model drop as a function of time (horizontal axis) and with increased uncertainty about
728 the input weights (different shades of blue). (d) For near-optimal learning, the learning rate (the
729 term ξ_w in Eq. (6)) is modulated by decision confidence (top left). High-confidence decisions lead
730 to little learning if correct (green, right), and strong learning if incorrect (red, left). Low-confidence
731 decisions result in a moderate confidence-related learning rate term (top, center). The learning
732 rate in 1000 simulated trials (bottom) shows that the overall learning rate preserves this trend,
733 with an additional suppression of learning for low-confidence decisions. Other learning heuristics
734 (e.g., the delta rule, right) do not modulate their learning by confidence.

735

736 **Figure 2. Input weight learning and tracking performance of different learning rules.** All
737 plots show the relative reward rate (0 = immediate, random choices, 1 = optimal) averaged over
738 5,000 simulations with different true, underlying weights, and for 2 (top) and 50 (bottom) inputs.
739 (a) The relative reward rate for probabilistic and heuristic learning rules. The probabilistic learning
740 rules include the optimal rule (Gibbs sampling), assumed density filtering (ADF), ADF with a
741 diagonal covariance matrix (ADF (diag)), and a learning rule based on a second-order Taylor

742 expansion of the log-posterior (Taylor exp.). For both 2 and 50 inputs, all rules perform roughly
743 equally. For the heuristic rules, different color shadings indicate different learning rates. The initial
744 performance shown is that after the first application of the learning rule, such that initial
745 performances can differ across learning rule. (b) The steady-state performance across different
746 heuristic rule learning rates. Steady state performance was measured as an average across 5,000
747 simulations, averaging over the last 100 of 1000 simulated trials in which the true weights slowly
748 change across consecutive trials. An optimal relative reward rate of one corresponds to knowing
749 the true weight in each trial, which, due to the changing weight, is not achievable in this setup.
750 The color scheme is the same as in (a), but the vertical axis has a different scale. The delta rule
751 did not converge and was not included in (b).

752
753 **Figure 3. Decision confidence, prior biases, and the relation between decision boundary**
754 **and choice.** (a) For an unbiased prior (i.e., $P^+ \equiv p(\mu \geq 0) = 1/2$), the decision confidence (color
755 gradient) is symmetric around $z = 0$ for each fixed time t . The associated posterior belief $p(\mu \geq$
756 $0|z(t), t)$ (numbers above/below “time” axis label; constant along white lines; $1/2$ along light blue
757 line) promote choosing $y = 1$ and $y = -1$ above (blue area in (b)) and below (red area in (b)) $z =$
758 0 . (b) As a result, different choices are Bayes-optimal at the blue/red decision boundaries, as long
759 as they are separated by $z = 0$, irrespective of the boundary separation (solid vs. dashed blue
760 red lines). (c) If the prior is biased by an overall shift, the decision confidence is counter-shifted
761 by the same constant across all t . In this case, both decision boundaries might promote the same
762 choice, which can be counter-acted by a time-invariant shift of z by $C_1(P^+)$. (d) If the prior is biased
763 by boosting one side while suppressing the other, the decision confidence shift becomes time-
764 dependent, such that the optimal choice at a time-invariant boundary might change over time.
765 Counteracting this effect requires a time-dependent shift of z by $C_2(P^+, t)$. In both (c) and (b) we
766 have chosen $P^+ = 0.6$, for illustration.

767
768 **Figure 4. Sequential choice dependencies due to continuous learning, and effects of noisy**
769 **feedback.** Bayes-optimal learning in a slowly changing environment predicts sequential choice
770 dependencies with the following pattern. (a) After hard, correct choices (low prev. $|\mu|$; light colors),
771 the psychometric curve is shifted towards repeating the same choice (blue/red = choice $y =$
772 $1/-1$). This shift decreases after easier, correct choices (high prev. $|\mu|$; dark colors). (b) We
773 summarize these tuning curve shifts in the repetition bias, which is the probability of repeating the
774 same choice to a $\mu = 0$ stimulus (example green arrow for $\mu = -0.38$ in (a)). After correct/incorrect
775 choices (green/red curve), this leads to a win-stay/lose-switch strategy. Only the win-stay strategy

776 is shown in (a). (c) If choice feedback is noisy (inverted with probability β), the learning rate
777 becomes overall lower. In particular for high-confidence choices with “incorrect” feedback, the
778 learning rate becomes zero, as the learner trusts her choice more than the feedback.