

1 **Tonic dopamine and biases in value learning linked through a**
2 **biologically inspired reinforcement learning model**

3

4

5 Sandra Romero Pinto^{1,2}, Naoshige Uchida¹

6

7 Affiliations:

8 ¹Department of Molecular and Cellular Biology, Center for Brain Science, Harvard University,

9 Cambridge, MA 02138, USA

10 ²Program in Speech and Hearing Bioscience and Technology, Division of Medical Sciences,

11 Harvard Medical School, Boston, MA 02115, USA

12

13 Correspondence: uchida@mcb.harvard.edu (N.U); sromeropinto@g.harvard.edu (S.R.P)

14

15

16

17 **Abstract**

18

19 A hallmark of various psychiatric disorders is biased future predictions. Here we examined the
20 mechanisms for biased value learning using reinforcement learning models incorporating recent
21 findings on synaptic plasticity and opponent circuit mechanisms in the basal ganglia. We show
22 that variations in tonic dopamine can alter the balance between learning from positive and
23 negative reward prediction errors, leading to biased value predictions. This bias arises from the
24 sigmoidal shapes of the dose-occupancy curves and distinct affinities of D1- and D2-type
25 dopamine receptors: changes in tonic dopamine differentially alters the slope of the dose-
26 occupancy curves of these receptors, thus sensitivities, at baseline dopamine concentrations. We
27 show that this mechanism can explain biased value learning in both mice and humans and may
28 also contribute to symptoms observed in psychiatric disorders. Our model provides a foundation
29 for understanding the basal ganglia circuit and underscores the significance of tonic dopamine in
30 modulating learning processes.

31

32

33 **Introduction**

34

35 Our ability to predict the outcomes of our actions is crucial in selecting and motivating
36 appropriate actions. Systematic biases in future predictions or expectations, however, can lead to
37 maladaptive behaviors, such as those observed in patients with various psychiatric disorders¹⁻⁴.
38 For example, overly negative or pessimistic predictions can contribute to major depression^{1,5},
39 whereas excessively positive or optimistic predictions may be associated with pathological
40 gambling, addiction, and mania^{3,4,6-8}. Despite the importance of understanding the causes of
41 biased future predictions, the biological mechanisms underlying them remain poorly understood.

42

43 Our future expectations and decisions are shaped by experiences of positive and negative events.
44 The process of learning from outcomes has been modeled using reinforcement learning (RL)
45 models⁹⁻¹², where value predictions are updated based on reward prediction errors (RPEs), that is
46 the discrepancy between received and expected outcomes. In addition to its role in learning,

47 recent studies have indicated the importance of RPEs in mood; these studies have suggested that
48 mood depends not on the absolute goodness of outcomes, but rather on the recent history of
49 RPEs^{13,14}.

50
51 In the brain, dopamine is thought to be a key regulator in this process of learning from positive
52 and negative outcomes. The dynamics of dopamine are often categorized into two modes: tonic
53 and phasic. Tonic dopamine refers to “baseline” dopamine that operates on a long timescale such
54 as tens of seconds or minutes, while phasic activity refers to transient changes that occur at a
55 much shorter, sub-second timescale, often triggered by external stimuli^{15–18}. A significant body
56 of evidence has shown that phasic responses of dopamine neurons convey RPEs and drive
57 learning of values and actions^{17–20}. On the other hand, changes in tonic dopamine might also
58 modulate value learning, yet whether and how the level of tonic dopamine modulates learning
59 remain poorly understood.

60
61 Previous studies have reported that patients with psychiatric disorders exhibit biased learning
62 from positive versus negative outcomes. For one, some studies have shown that patients with
63 major depression have a reduced sensitivity in learning from rewarding events, while their ability
64 to learn from negative events remains relatively intact^{1,5,21}. Similarly, patients with Parkinson’s
65 disease are better at learning from negative than positive outcomes^{22,23}. Analysis of these patients
66 using RL models has suggested that biases in learning can be explained by alterations in specific
67 parameters in RL models, such as the learning rate parameters or the sensitivity to positive and
68 negative outcomes. For example, some studies have suggested that anhedonia in major
69 depressive disorder may correspond to a reduced learning rate from positive compared to
70 negative outcomes¹.

71
72 Mechanistically, some of these changes in RL parameters can be linked to altered functions of
73 dopamine. First, it has been shown that dopamine synthesis capacity, an approximate indicator of
74 baseline dopamine levels, in the striatum, as measured using positron emission tomography

75 (PET), correlates with learning rate parameters²⁴. Second, dopamine medications can change the
76 balance between learning from positive and negative outcomes^{22,24,25}. Third, responses to
77 positive outcomes in the nucleus accumbens (NAc), as measured based on blood oxygenation-
78 dependent (BOLD) signals, are reduced in patients with psychiatric disorders such as
79 depression²⁶⁻²⁹. These observations point to important roles of reinforcement learning processes
80 and dopamine in regulating value learning. However, the parameters in RL models remain an
81 abstract entity, and biological processes underlying changes in these parameters are still largely
82 unknown.

83
84 One limitation in most RL models used in previous studies is that they do not reflect key neural
85 circuit architectures in the brain (but see ³⁰⁻³²) nor recent findings on intracellular signaling and
86 plasticity rules that can constrain how dopamine functions in biological circuits³³⁻³⁵.
87 Incorporating these key biological factors may lead to better understanding of how changes in
88 RL parameters may arise in psychiatric disorders. Furthermore, recent studies have found that
89 the activity of dopamine neurons is consistent with a novel RL algorithm called distributional
90 RL³⁶⁻³⁸. Distributional RL takes into account the diversity in dopamine signals, and a population
91 of dopamine neurons together encodes the entire distribution of rewards, not just the average.
92 Although distributional RL has shown to be efficient in solving various RL problems in artificial
93 intelligence^{37,39}, how distributional RL can be implemented in biological neural circuits and how
94 distributional RL relates to biased value learning remain to be examined.

95
96 In this study, we sought to identify potential mechanisms that cause biased value predictions
97 using biologically inspired RL models. To this goal, we first construct an RL model that
98 incorporates recent biological findings, such as intracellular signaling and synaptic plasticity
99 rules as well as the basic circuit architecture in the brain³². Based on this model, we propose two
100 potential biological mechanisms that can cause optimistic or pessimistic biases in value
101 predictions. We will then show that some existing data can be explained by one of these models.

102 Finally, we will show how our model can provide an account of how biases in value predictions
103 arise in psychiatric disorders.

104

105

106 **Results**

107

108 **Basic reinforcement learning algorithms**

109 Here we first formulate basic RL algorithms that will become the basis of our later models. In
110 RL, an agent learns to predict the expectation of future rewards associated with a given state, a
111 quantity termed as *value*¹¹. For simplicity, we will drop the dependency on time here, but note
112 that the basic results hold even if time is considered (Methods 1.1). Learning of value is driven
113 by RPEs (δ), the discrepancy between the actual and expected reward (r and V , respectively)
114 (Eq1). To improve the accuracy of the value prediction, RPEs are utilized to update the estimate
115 of V . This is done iteratively by adding a fraction (α) of δ (Eq2) where α defines the learning
116 rate.

117

$$118 \quad \delta = r - V \quad (1)$$

$$119 \quad V \leftarrow V + \alpha \cdot \delta \quad (2)$$

120

121 When the magnitude of reward r is fixed (i.e., deterministic environment), the value V learned
122 through this algorithm (Eq 1 and 2) converges on r and the RPE converges on zero. When the
123 magnitude of reward r varies stochastically trial-to-trial, the value at convergence fluctuates
124 around the expected value of the reward distribution (see Methods 1) (Fig. 1a) and the RPE
125 around zero.

126

127 **Risk-sensitive RL.** In the framework called risk-sensitive RL⁴⁰, learning rates are defined
128 separately for positive and negative RPEs (denoted by α^+ , α^-).

129

$$130 \quad V \leftarrow V + \alpha^+ \cdot \delta \quad \text{if } \delta > 0 \quad (3)$$

131 $V \leftarrow V + \alpha^- \cdot \delta$ if $\delta < 0$

132

133 In the presence of stochastic rewards, when the learning rates between positive and negative
134 RPEs are different, the value learned through this algorithm (Eq 1 and 3) does not converge on
135 the expected value, but instead on a value higher or lower than the expected value depending on
136 the relative amplitude of the learning rates α^+ , α^- . This algorithm, therefore, develops optimistic
137 or pessimistic value expectations, respectively. This learning algorithm is called “risk-sensitive”
138 because values of probabilistic (risky) rewards are biased compared to deterministic (certain)
139 rewards, and, therefore, the agent develops a preference between risky and certain rewards even
140 when the expected values are the same (Fig. 1b).

141

142 **Distributional RL.** The concept of asymmetric updates has been utilized in a novel RL
143 framework called distributional RL^{36,37,41}. This algorithm allows an agent to learn the entire
144 probability distribution of rewards, instead of the expected value which is typically the learning
145 target in traditional RL algorithms (Fig. 1c). In distributional RL, an agent is equipped with a set
146 of value predictors (V_i), where i corresponds to the index of the value predictor (or “value
147 neuron”). The value of the i -th neuron (V_i) is updated based on the learning rates (α_i^+ , α_i^-) and
148 the RPE (δ_i) for that neuron i :

149

150 $V_i \leftarrow V_i + \alpha_i^+ \cdot \delta_i$ if $\delta_i > 0$ (3)

151 $V_i \leftarrow V_i + \alpha_i^- \cdot \delta_i$ if $\delta_i < 0$

152

153 Similar to risk-sensitive RL, the learned value of each value predictor converges on estimates
154 larger or lower than the expected value, determined by the ratio between α_i^+ and α_i^- .

155 Mathematically, each V_i converges on the τ_i -th expectile of the distribution (Fig. 1c) where τ_i
156 (asymmetric scaling factor) is defined by:

157

158 Asymmetric scaling factor: $\tau_i = \frac{\alpha_i^+}{\alpha_i^- + \alpha_i^+}$ (4)

159

160 Expectiles are the solutions to asymmetric least squares minimization and generalize the mean of
161 a distribution (with the mean being the 0.5th expectile) as quantiles generalize the median (with
162 the median being the 0.5th quantile)⁴². Since a set of expectiles can define a distribution, the
163 diversity of τ_i across the population enables learning of the entire probability distribution.

164

165 **Problem.** In both risk-sensitive RL and distributional RL, unbalance in learning rate parameters
166 for positive and negative RPEs gives rise to optimistic and pessimistic biases in learned values.
167 Importantly, however, the underlying biological mechanism regulating learning rate parameters
168 (α^+ , α^-) and asymmetry thereof (τ) remains unclear.

169

170 In the following sections, we will discuss potential biological mechanisms that regulate
171 asymmetric learning rates (α^+ , α^-). We will first modify the above RL algorithms to incorporate
172 important neural circuit architectures in the brain. We will then propose two key biological
173 mechanisms that can give rise to asymmetric learning rates (called Model 1 and 2). We will then
174 show that our model can explain previous experimental data and psychiatric conditions.

175

176 **Incorporating biological features into RL models**

177 The above RL models provide algorithmic-level formulations, yet they do not recapitulate
178 fundamental characteristics of the neural circuits thought to perform RL in the brain^{43–46}. We
179 next incorporate some of the important circuit and synaptic properties into the model.

180

181 In the brain, it is thought that dopamine neurons in the ventral tegmental area (VTA) broadcast
182 RPEs¹⁷ and modulate synaptic plasticity in dopamine-recipient areas^{33,47}. The striatum is the
183 major target of dopaminergic projections. It has been thought that spiny projection neurons
184 (SPNs) in the striatum represent values, and dopamine modulates plasticity of synapses on

185 SPNs^{33,34,47,48} (Fig. 2a). Under this framework, the value representations in SPNs are updated by
186 dopaminergic RPEs. In most RL models, each value predictor is typically updated by both
187 positive and negative RPEs. If the value is computed based on a weighted sum of some inputs
188 (i.e., using linear function approximation¹¹), the update rules described above (Eq 3 and 4) are
189 equivalent to performing a semi-gradient descent that minimizes RPEs¹¹ (see Methods).

190

191 The basic architectural assumptions of these RL models are, however, at odds with the RL
192 circuitry in the brain. Importantly, in the striatum, there are two major classes of dopamine-
193 recipient SPNs characterized based on the type of dopamine receptor that they express: D1- or
194 D2-type dopamine receptors (D1R and D2R)⁴⁸. SPNs expressing D1R and D2R constitute the so-
195 called direct and indirect pathways and exert opposing effects on downstream “output” neurons,
196 with each pathway promoting or opposing a certain output (e.g., movement).

197

198 In addition to the presence of direct and indirect pathways, there are two additional properties in
199 these opposing populations that need to be considered³². First, D1R and D2R have different
200 affinities to dopamine: high in D2R and low in D1R (EC₅₀ affinity constant is 1 μ M for D1R and
201 10 nM for D2R)^{49,50}. The dose-occupancy relationship of D1R and D2R are sigmoidal but they
202 are shifted with one another with respect to dopamine concentration (Fig. 2b). Importantly, at
203 normal dopamine levels (approx. 50-100nM)^{51,52}, D2Rs are mostly occupied while D1Rs are
204 mostly unoccupied (Fig. 2b). Although whether the affinities of D1R and D2R differ at the
205 molecular level has been questioned⁵³, a recent study showed that intracellular signaling through
206 protein kinase A (PKA) in D1- and D2-SPNs is triggered by a phasic increase and a decrease in
207 dopamine, respectively, in behaving animals³⁵. These results are consistent with (apparent)
208 difference in affinities of D1R and D2R assumed in previous studies⁴⁹, although the exact reason
209 for the difference remains to be clarified⁵³.

210

211 The second important property pertains to different learning rules in D1- and D2-SPNs which are
212 predicted from different affinities of the receptors. Consistent with the observed PKA signals in

213 these cells, recent studies have shown that glutamatergic inputs on D1-SPNs are potentiated by a
214 transient *increase* in dopamine, whereas those on D2-SPNs are potentiated by a transient
215 *decrease* in dopamine^{34,35} (Fig. 2c), supporting opposing plasticity rules between D1- and D2-
216 SPNs.

217

218 There have been previous efforts to incorporate in RL models the direct and indirect pathways
219 (also called “Go” and “NoGo” pathways, respectively) such as Opponent Actor Learning
220 (OpAL³⁰, OpAL*⁵⁴) and Actor learning Uncertainty (AU)³² models. These previous models
221 were developed as *Actor-Critic models*¹¹. Here, we will build on the AU model to focus on the
222 problem of value learning and extend it to support risk-sensitive RL and distributional RL. Our
223 model has two separate populations of value predictors corresponding to D1R- and D2R-SPNs,
224 that store the quantities P_i and N_i respectively (Eq 6, Fig. 2d). Mimicking dopamine's effect on
225 potentiation, P_i or N_i will increase their estimates if an RPE is positive or negative, respectively,
226 with the learning rates defined by α_i^+ , α_i^- (Eq. 6). Importantly, the value V_i can be obtained
227 simply by taking the difference between P_i and N_i . (Eq. 7).

228

$$229 \quad \text{D1R-SPN: } P_i \leftarrow P_i + \alpha_i^+ \cdot |\delta_i| - \beta \cdot P_i \quad \text{if } \delta_i > 0 \quad (5)$$

$$230 \quad \text{D2R-SPN: } N_i \leftarrow N_i + \alpha_i^- \cdot |\delta_i| - \beta \cdot N_i \quad \text{if } \delta_i < 0 \dots$$

$$231 \quad \text{Value: } V_i = P_i - N_i \quad (6)$$

232

233 where β is a decay parameter which represents synaptic decay in the absence of RPEs. This
234 model (Eq 6 and 7) preserves various essential properties of the previous RL models: (1)
235 learning in P and N can be combined to provide a simple update rule for value V , and (2) this
236 update rule approximates the gradient descent that minimizes RPEs (when $\beta = 0$, the update rule
237 is equivalent to the gradient descent). Importantly, with $\beta > 0$, we can show that these simple
238 learning rules guarantee convergence of value, without the need for additional mechanisms to
239 modulate the learning rates over iterations (Methods 1.3).

240

241 For instance, in a stochastic environment where there is a probability p of receiving a reward of a
242 fixed magnitude $r = 1$, the stochastic fixed point of the learned value V_i (i.e., convergence point)
243 will be defined by Eq 7.

244

$$245 \quad V_i = \frac{\frac{\tau_i \cdot p}{1-\tau_i \cdot 1-p}}{\frac{\tau_i \cdot p}{1-\tau_i \cdot 1-p} + 1 + C} \cdot r, \text{ where } C = \frac{\beta}{(1-p) \cdot (1-\tau)} \quad (7)$$

246

247 Note that Eq.7 contains an additional term C which depends on β and this decay factor β is
248 important to stabilize the P_i and N_i estimates (avoid infinite increases) (Methods, 1.3.1, Extended
249 Data Fig. 1).

250

251 This formulation now provides a mechanistic model suitable for risk-sensitive RL (when there is
252 one value predictor) as well as distributional RL (when there are multiple value predictors),
253 which incorporate the neural circuit architecture and plasticity rules of D1R- and D2R-SNPs
254 found in the brain.

255

256 With this model at hand, we now discuss potential mechanisms that produce an asymmetry in
257 learning rates α_i^+ , α_i^- , which, in turn, causes biases in value predictions. In principle, learning
258 rate parameters can be a function of (1) the scaling of RPEs, i.e., the slope of dopamine
259 responses as a function of RPE (δ), and (2) the scaling of value updates, i.e., the efficacy of
260 dopamine-dependent synaptic plasticity at the level of SPNs. In the following, we discuss each
261 scenario, emphasizing the role of either tonic or phasic dopamine activity in each of these
262 mechanisms (Model 1 and 2, respectively). For simplicity, we will start with a model in which
263 α^+ , α^- are equal for all neurons within both P and N populations, equivalent to risk-sensitive
264 RL. We will then relax this assumption and introduce heterogeneity by allowing α_i^+ , α_i^- to vary
265 across neurons, implementing a form of distributional RL.

266

267 **Model 1: The role of baseline dopamine in asymmetric learning**

268 As discussed above, D1R and D2R have different affinities to dopamine which leads to different
269 levels of receptors' occupancy at a given baseline dopamine level (Fig. 2b). Crucially, due to the
270 sigmoidal shape of the dose-occupancy curves, the slope of the curve changes with baseline
271 dopamine level, which means that a given dopamine transient leads to a different change in
272 receptor occupancy depending on the baseline dopamine level (Fig. 3a,b). That is, the receptors'
273 *sensitivity* changes with baseline dopamine (Fig. 3c). In addition, a key consequence of the
274 distinct receptors' affinities is that an increase and decrease in baseline dopamine will cause
275 opposite changes in the sensitivity of D1R and D2R. Specifically, an increase in dopamine will
276 decrease D1R sensitivity relative to D2R, whereas a decrease in dopamine will increase D2R
277 sensitivity relative to D1R (Fig. 3c,d).

278

279 Building on this insight in Model 1, we postulate that the learning rates for positive and negative
280 RPEs are a function of the D1R and D2R sensitivity, respectively. This is supported by previous
281 studies that have reported that the effect of dopamine transients of a given magnitude in SPNs'
282 plasticity can be modulated by the level of dopamine baseline³⁴. In addition, it has been reported
283 that the level of potentiation in SPNs^{33,34} or plasticity, which are related to intracellular signals³⁵,
284 scale with the magnitude of dopamine transients, keeping all else fixed. These observations can
285 be summarized with the following rule:

$$286 \quad LTP_{D1} \approx \alpha^+ \cdot |DA_{burst}|$$

$$287 \quad LTP_{D2} \approx \alpha^- \cdot |DA_{pause}|$$

288 Where α^+ , α^- correspond to the receptors' sensitivities and depend on the dopamine baseline
289 level. This rule can be directly related to the update equations for the P and N populations in our
290 model:

$$291 \quad \Delta P = \alpha^+ \cdot |\delta| \dots \text{if } \delta \geq 0$$

$$292 \quad \Delta N = \alpha^- \cdot |\delta| \dots \text{if } \delta < 0$$

293

294 It can be shown that this learning rule is in agreement with the normative solution for the
295 problem of value learning¹¹ (Methods 1.6).

296

297 In short, in Model 1, a shift in the baseline dopamine level causes asymmetries in scaling of the
298 value updates for positive versus negatives RPEs via the modulation of receptors' sensitivities,
299 which leads to value learning biases. This is a direct consequence of the dose occupancy
300 relationships of D1R and D2R (Fig. 3b-d).

301

302 **Model 2: Asymmetric scaling of phasic dopamine responses, inspired by distributional RL**

303 In Model 2, we postulate that the learning rates α^+ and α^- are a function of the scaling (i.e.,
304 'slope') of dopamine responses evoked by positive and negative RPEs, respectively:

$$305 \quad DA_{burst} = \alpha^+ \cdot \delta \dots \text{if } \delta \geq 0$$

$$306 \quad DA_{pause} = \alpha^- \cdot \delta \dots \text{if } \delta < 0$$

307 This is supported by a previous study on distributional RL that demonstrated that individual
308 dopamine neurons vary in terms of how the magnitude of reward responses is scaled as a
309 function of positive and negative RPEs (Fig. 3e)³⁶.

310

311 In the distributional RL framework, individual dopamine neurons vary in terms of their
312 asymmetric scaling factor τ_i and each of the multiple value predictors (V_i) converges on the τ_i -th
313 expectile of the reward distribution (Eq. 4). However, in most applications of distributional RL,
314 action selection is still based on the expected value of the reward distribution. Thus, the quantity
315 relevant to action selection can be described using the population level average $\tau_{population}$, and
316 biased value learning at the behavioral level could arise if $\tau_{population}$ is higher or lower than 0.5.
317 This can occur from a differential loss of optimistic or pessimistic dopamine neurons. Another
318 possibility is an overall upward or downward shift in the distribution of τ_i across the population
319 due to, for example, intrinsic factors modulating the gain of dopamine phasic responses.

320

321 Risk-sensitive RL can be thought of as a special case of distributional RL which has only one
322 value predictor. Here, the slope of the average dopamine evoked transient to positive and
323 negative RPEs, may correspond to the population level learning rates for positive and negative
324 RPEs (α^+ , α^-), respectively. If the asymmetric scaling factor τ is higher or lower than 0.5, value
325 learning will be biased (Fig. 3e).

326

327 **Testing for evidence of either model in experimental data**

328 *Tian and Uchida (2015).*

329 We next examined whether Model 1 or 2 can explain empirical data obtained in experimental
330 animals or humans. We first examined the data obtained in mice in our previous study⁵⁵. In this
331 study, the authors tested the effect of lesioning the habenula, a brain structure implicated in
332 depression⁵⁶⁻⁵⁸, on the activity of dopamine neurons and on reward-seeking behavior. Head-fixed
333 mice were trained in a Pavlovian conditioning task in which odor cues predicted reward with
334 different probabilities (10%, 50%, 90%). After performing habenula (n=5) or sham (n=7) lesions
335 (Fig. 4a), the spiking activity of VTA dopamine neurons was recorded while mice performed the
336 task.

337

338 After lesions, mice exhibited an elevated reward-seeking behavior (anticipatory licking) in
339 response to cues predictive of probabilistic rewards, consistent with an optimistic bias in reward
340 expectation (Fig. 4b, right). Importantly, anticipatory licking gradually increased over several
341 sessions after lesions, suggesting that the optimistic bias developed through learning (Fig. 4b,
342 left). To bring insight into the underlying cause of these biases, we fit two different RL models to
343 the anticipatory lick responses on a trial-by-trial basis (Extended Data Fig. 2), assuming a linear
344 relationship between value predictions and anticipatory licking. These models considered either a
345 change in the sensitivity to rewards (Extended Data Fig. 2b) or asymmetric learning rates
346 (Extended Data Fig. 2c). This analysis showed that the biases observed in the behavior could be
347 explained by asymmetric learning rates, but not by reward sensitivity because the reward

348 sensitivity was unchanged in the lesion group with respect to the control group (Extended Data
349 Fig. 2c).

350
351 Dopamine neurons' responses to reward-predictive cues reflect the increases in value expectation
352 predicted by the cue with respect to baseline. The overall magnitudes of cue-evoked responses
353 were not elevated in lesioned animals compared to control animals (Fig. 4d). However, the shape
354 of the response curve indicated an 'optimistic' bias: although in control animals, cue responses
355 scaled linearly with the expected value (i.e., reward probability), the response function of the
356 lesioned animals was convex. In other words, in control animals the response to the 50%-reward
357 cue was not significantly different from the quantity that results from the linear interpolation
358 between the responses to 10%- and 90%-reward cues. In lesioned animals, however, the response
359 to the 50%-reward cue was significantly greater than this quantity, which is indicative of an
360 optimistic bias in value predictions (Fig. 4d, see Methods 1.3.3 for analysis of value predictions
361 curve convexity). Such a change was observed at the level of the population average. Further
362 analysis using individual neurons showed that when calculating a single-cell level metric that
363 compares the 50%-reward cue to the same linear interpolation point, there was a broad
364 distribution in this metric below and above the interpolated point both in the control and lesion
365 groups (Fig. 4e-f). The distribution was, however, shifted in its mean in the lesion group (Fig.
366 4e). These analyses indicated that both anticipatory licking and dopamine cue responses have an
367 optimistic bias as characterized by an overvaluation of probabilistic rewards, without still
368 pointing to the underlying mechanism.

369
370 **Model 2 cannot explain the optimistic biases in behavior and cue-evoked dopamine**
371 **responses after Hb lesions**

372
373 In Model 2, an optimistic bias in reward expectation can arise if the average of the asymmetric
374 scaling factor at the population level ($\tau_{population}$) becomes greater than 0.5 (Fig. 5a,b).

375

376 To test this idea, we obtained the asymmetric scaling factors (τ_i) from dopamine neurons based
377 on their outcome responses: for each neuron, we constructed outcome response functions against
378 the magnitude of RPEs (Fig. 5c, Extended Data Fig. 3a,b). The response functions were obtained
379 based on (1) whether reward was delivered (positive RPEs) or not (negative RPEs), and on (2)
380 the magnitude of the reward expectation given by the reward probabilities predicted by each cue
381 (0.1, 0.5, 0.9) (Extended Data Fig. 3a,b). We then obtained the point at which the responses are
382 more likely to be below or above baseline (i.e., ‘zero-crossing points’)³⁶ (Extended Data Fig. 3c),
383 and computed α_i^+ and α_i^- as the slopes of the responses in the positive and negative domains
384 with respect to this zero-crossing point (Extended Data Fig. 3d), respectively. In both control and
385 lesioned animals, asymmetric scaling factors tiled a wide range between 0 and 1 and presented
386 other signatures consistent with distributional RL³⁶ (Extended Data Fig.4). Nonetheless, although
387 the variance of the distribution of asymmetric scaling factors was greater in lesioned animals, the
388 mean did not change, indicating a lack of bias between α_i^+ and α_i^- at the population level (Fig.
389 5d). This was also the case when the asymmetric scaling factor was derived directly from the
390 population average response (Fig. 5c). Thus, contrary to the conclusion in our previous study¹⁵,
391 these analyses indicated that changes in reward responses (and the resulting scaling factor τ) do
392 not explain the optimistic biases in behavior nor cue responses in lesioned animals (Fig. 5e,f).

393

394 **Model 1 can explain the optimistic biases in behavior and cue-evoked dopamine responses** 395 **by Hb lesion**

396 In addition to changes in the magnitude of dopamine RPEs, we observed that the baseline firing
397 rates of dopamine neurons were elevated in lesioned animals (Fig. 6a). According to Model 1, if
398 these changes are followed by an increase in the baseline dopamine levels in the striatum, this
399 should give rise to biased value learning ($\alpha^+ > \alpha^-$) and an optimistic bias in value expectation.
400 In this way, this change in baseline firing can explain optimistic biases observed in lesioned
401 animals. However, it remains unclear whether the observed change in baseline firing can result in
402 functionally relevant levels of changes in the receptor occupancies discussed above.

403

404 To quantitatively predict dopamine concentrations in the striatum and resulting receptor
405 occupancies of D1R and D2R, we used a biophysical model commonly used in the field⁵⁹
406 (Fig.6a). This model has the firing rate of dopamine neurons as its input, and considers diffusion
407 of dopamine, dopamine reuptake, and D2-autorreceptor-mediated inhibition of dopamine release
408 to predict the dopamine concentration in the striatum (Fig. 6b,e). In addition, it considers the
409 affinities of D1R and D2R to estimate their occupancy levels (Fig. 6c,f). After estimating these
410 two variables (dopamine concentration and receptor occupancy), we derived the receptor
411 sensitivities (Fig. 6g-h). The receptor sensitivities were quantified as the slope of the resultant
412 changes in receptor occupancy given the observed baseline and phasic responses of dopamine
413 neurons. We then trained Model 1 using the receptor sensitivities as learning rates (α^+ and α^-)
414 for both control and lesioned animals.

415
416 The biophysical model indeed supported that the observed change in dopamine neuron firing can
417 cause a significant increase in dopamine concentration (Fig. 6e) and in D1 and D2 receptor
418 occupancies at baseline (Fig. 6g). These changes are expected to cause a significant asymmetry
419 in receptor sensitivities favoring D1 receptors over D2 receptors (Fig. 6h-i).

420
421 These receptor sensitivities were directly used as the asymmetric learning rates in a temporal-
422 difference (TD) learning version of Model 1 (see Methods 1.3, 3.3). After training, the model
423 incorporating the predicted asymmetries in learning rates (α^+ , α^-) produced optimistic biases in
424 value predictions and in normalized cue responses, similar to those observed in lesioned animals
425 (Fig. 6k-l). The model simulating control animals developed no significant biases.

426
427 Additionally, the overall decrease in the magnitude of cue responses, observed in lesioned
428 animals, was reproduced in Model 1 using TD learning (Fig. 6k). This occurs because TD
429 learning calculates RPEs based on the change in values between before and after cue
430 presentation, and the “baseline” (pre-cue) reward expectation was also increased by optimistic
431 value learning (Fig. 6k). These results, together, indicate that Model 1 provides a parsimonious

432 account of the data: a change in baseline firing of dopamine neurons, rather than changes in
433 phasic responses, is the likely mechanism that led to optimistic biases in reward-seeking
434 behavior as well as cue-evoked dopamine responses in habenula lesioned animals.

435

436 **Model 1 and model 2 play complementary roles in the encoding of asymmetric learning**
437 **rates**

438 Although Model 2 did not explain the optimistic biases in the data in habenula-lesioned mice, the
439 distributional RL version of Model 2 explained other features of the data (Extended Data Fig. 3-
440 4). As mentioned, in both control and lesioned animals, asymmetric scaling factors tiled a wide
441 range between 0 and 1³⁶ (Extended Data Fig.4). Furthermore, cue-evoked responses of individual
442 neurons showed a wider distribution than what is expected by noise (Figure 4d). Finally, the core
443 prediction of distributional RL – a positive correlation between the asymmetric scaling factors of
444 the RPE responses of individual dopamine neurons and their zero-crossing points³⁶ – was also
445 present in controls and after Hb lesions. Together these results support that the basic features of
446 distributional RL are present in a way consistent with Model 2.

447

448 To complement this analysis, we tested whether Model 2 could have explained the signatures of
449 the data if asymmetric scaling factors (τ) derived from dopamine responses were indeed overall
450 biased (Extended Data Fig. 5). As expected from the model's fixed-point analysis (Methods 1.3),
451 if we imposed a shift in the mean of the distribution of asymmetric scaling factors (i.e.,
452 $\tau_{population} > 0.5$), the value predictors indeed exhibited optimistic biases (Extended Data Fig.
453 5e,f). However, the model did not reproduce the optimistic bias in cue-induced TD errors
454 observed in the data (Extended Data Fig.5g,h). This is due to an interaction of the biases in
455 prediction at “baseline” (pre-cue) and the cue, together with the optimistic asymmetry in the
456 scaling of the TD errors at cue themselves. Importantly, this was found in both versions of Model
457 2, distributional and risk-sensitive RL (Extended Data Fig. 5a-d and e-h). The difficulty of
458 explaining biased dopaminergic cue responses further makes the Model 2 an unlikely mechanism
459 to explain the optimistic biases in the data.

460

461 Altogether the data supports a model in which the mechanisms of Model 1 and 2 play
462 complementary roles in the encoding of asymmetric learning rates. The mechanism of Model 2
463 explains the variability in single neuron responses, consistent with the expectile code in
464 distributional RL. On the other hand, the mechanism of Model 1 at the population level,
465 generating asymmetries in learning rates and biases in value expectations, which might require
466 context-dependent regulation⁶⁰.

467

468 Taken together, the above results suggest that Model 1 and 2 coexist in the brain. This can be
469 formalized as follows:

470

$$471 \quad P_i \leftarrow P_i + \hat{\alpha}_i^+ \cdot \delta_i \dots \text{if } \delta_i \leq 0$$

$$472 \quad N_i \leftarrow N_i + \hat{\alpha}_i^- \cdot \delta_i \dots \text{if } \delta_i < 0$$

473 where:

$$474 \quad \hat{\alpha}_i^+ = \alpha_P \cdot \alpha_i^+$$

$$475 \quad \hat{\alpha}_i^- = \alpha_N \cdot \alpha_i^-$$

476

477 where α_i^+ and α_i^- correspond to the asymmetric scaling of dopamine RPEs at the single-cell level
478 (Model 2) and α_P and α_N correspond to the asymmetric scaling of synaptic plasticity in D1R and
479 D2R at the population level (Model 1).

480

481 **Linking asymmetric learning and baseline dopamine levels in healthy subjects**

482 *Cools et al., (2009)*²⁴.

483 There have been very few studies that examined the relationship between baseline dopamine
484 levels and asymmetry in learning from positive and negative outcomes. As a rare case for such
485 examinations, Cools et al.²⁴ provided intriguing data in humans. They compared the performance
486 in reversal learning and the quantity called ‘dopamine synthesis capacity’. Dopamine synthesis
487 capacity is estimated by injecting the positron emission tomography (PET) tracer

488 [¹⁸F]fluorometatyrosine (FMT) and is thought to be correlated with baseline dopamine levels^{61,62}.
489 This study found that higher dopamine synthesis capacity was correlated with better learning
490 from gains but not with learning from losses (Fig. 7b). As a result, in reversal learning, subjects
491 with higher dopamine synthesis capacity learned from gains than losses, reported as the ‘relative
492 reversal learning (RRL)’ index in their study (Fig. 7b). This result, thus, provides direct evidence
493 supporting our Model 1.

494
495 In addition, they found that dopamine synthesis capacity predicts the effectiveness of
496 bromocriptine (D2 partial agonist) in altering learning rate asymmetry: bromocriptine’s ability to
497 bias learning from gains over losses (i.e., positive change in RRL) was negatively correlated with
498 dopamine synthesis capacity (Fig. 7c). We found that this result can also be explained by Model
499 1. For this, we simulated the effects of bromocriptine with the biophysical model used above,
500 and derived the asymmetric learning rates from the slopes of the D2R occupancy (Fig. 7d,
501 Extended Data Fig. 6a,b) or activation curves (Fig. 7d, Extended Data Fig. 6c,d). The RRL
502 parameter reported by Cools et al. corresponds to the asymmetric scaling factor τ , and is
503 equivalent to $(2\tau - 1)$ (as described in the Methods 4.1). We then computed what would be the
504 change in this parameter $\Delta(2\tau - 1)$ induced by bromocriptine (Fig. 7e-f, Extended Data Fig. 6e-
505 1).

506
507 This analysis revealed that by considering the asymmetries in learning rates induced by changes
508 in the baseline occupancy of the receptors, our model can capture their results in a qualitative
509 manner. Intuitively, the less dopamine there is at baseline, the lower the occupancy of D2R at
510 placebo conditions. This leads to a larger increase in D2R occupancy induced by D2 agonist in
511 low dopamine baseline conditions (Fig. 7d, Extended Data Fig. 6a) and, thus, a larger increase in
512 asymmetry in learning from gains over losses, if D1R occupancy is kept fixed. These effects still
513 hold even if we consider, in addition to bromocriptine’s effects in postsynaptic receptors (D2
514 long or D2l), its effect on inhibition of dopamine release via presynaptic (D2 short or D2s)
515 autoreceptors^{63,64} (Fig. 7d, Extended Data Fig. 6b). This can be simulated as a decrease in

516 dopamine level, which leads to a shift in the occupancy curves to the right. Finally, we can
517 consider effect of the *partial* agonism of the drug, that leads to a lower activation level of
518 receptors even if the occupancy is maximal (Fig. 7d, Extended Data Fig. 6c-d). Even after
519 considering this last factor, the results remain qualitatively the same as those found in the
520 original study. These results were robust to a relatively wide range of values in the simulation's
521 parameters (Extended Data Fig. 7, 8).

522

523 **Linking psychiatric conditions to baseline dopamine levels**

524 *Timmer et al., 2018*

525 Various psychiatric disorders are characterized by abnormal future predictions or mood. Our
526 Model 1 raise the possibility that an overall decrease in baseline dopamine level in the striatum
527 would enhance learning from negative outcomes over learning from positive outcomes leading to
528 persistent pessimistic future value expectations, a hallmark of depressive-like symptoms (Fig.
529 3a,b). A piece of evidence supporting this in the human literature is the greater learning rates for
530 losses over gains in patients with Parkinson's disease (PD)^{22,25}, its comorbidity with
531 depression^{25,65} that can precede the PD diagnosis⁶⁵⁻⁶⁷, and the reports of decreased dopamine
532 transporter binding in the ventral striatum in depressed PD patients compared to non-depressed
533 PD patients^{68,69}.

534

535 In addition, the progression of dopaminergic axonal loss in PD is topographically unbalanced:
536 the axonal loss is more prominent in the dorsal striatal regions⁷⁰ than in the ventral ones. This
537 leads to uneven dopamine baseline levels across the striatum that would interact with the global
538 increases in dopamine induced by dopaminergic medications in PD patients. We hypothesize that
539 a behavioral readout of the degree of this unevenness might be the presence or absence of
540 depression as a comorbidity: *patients with depression might have lower dopamine levels in the*
541 *ventral striatum*. Thus, if indeed baseline dopamine levels are correlated with depression, this
542 comorbidity could be predictive of the effects of PD medication.

543

544 We examined a previous study that provided evidence for this hypothesis ⁷¹. Here, PD patients
545 with and without depression history were tested in a gambling task, under presence or absence of
546 medication ('ON' and 'OFF' medication states). The authors fitted a 'loss aversion' parameter to
547 the behavioral performance, which is equivalent to $1 - \tau$ in our model, under some assumptions
548 (Methods). Their results were consistent with our model predictions. In the OFF-medication
549 state, there was a (near-significant) main effect of depression group (with or without depression)
550 on the learning rate asymmetry: patients with a depression history tended to be more loss averse
551 than nondepressed patients ($P = 0.052$). This is consistent with a decrease of dopamine levels in
552 the ventral striatum and thus a regime of $\alpha^+ < \alpha^-$ in value learning. Importantly, in the ON-
553 medication state, the medication effects on the asymmetry in learning rates were predicted by the
554 degree of severity of depression: patients with larger depression scores exhibited greater drug-
555 induced decreases in loss aversion (Fig. 7g), which would correspond to an increase in $\tau =$
556 $\frac{\alpha^+}{\alpha^+ + \alpha^-}$ in our model This is consistent with our Model 1: higher degrees of depression might be
557 correlated with lower levels of baseline dopamine, making the D1R sensitivity more susceptible
558 to an artificial increase in baseline dopamine with L-DOPA medication (Fig. 7h; further details
559 discussed in Methods).

560

561

562 **Discussion**

563

564 A hallmark of various psychiatric disorders is overly optimistic or pessimistic predictions about
565 the future. Using RL models, we sought to identify potential biological mechanisms that give rise
566 to biased value predictions, with a particular focus on the roles of phasic versus tonic dopamine.
567 Our results demonstrate that variations in tonic dopamine levels can modulate the efficacy of
568 synaptic plasticity induced by positive versus negative RPEs, thereby resulting in biased value
569 learning (Model 1). This effect arises due to sigmoidal shapes of the dose-occupancy curves and
570 different affinities of dopamine receptors (D1R and D2R); alterations in the tonic dopamine level

571 result in changes in the slope of the dose-occupancy curve (and thus, sensitivity) of dopamine
572 receptors at the baseline dopamine concentration. We show that this mechanism offers a simple
573 explanation for how changes in tonic dopamine levels can result in biased value learning in a few
574 examples of value learning in mice and humans. Additionally, we show that this mechanism may
575 underlie symptoms of various psychiatric and neurological disorders. Although altered phasic
576 dopamine responses could have been a natural suspect as a candidate mechanism for biased
577 value learning^{37,38}, our study provides a novel mechanism; the interaction between tonic and
578 phasic dopamine can give rise to biased value learning, even when phasic dopamine responses
579 remain relatively unchanged.

580

581 **The impact of properties of dopamine receptors on reinforcement learning (RL)**

582 Our results highlight the importance of considering properties of dopamine receptors and neural
583 circuit architecture (i.e., direct and indirect pathways) in RL models. Based on different affinities
584 of dopamine D1 and D2 receptors, it has been proposed that D1- and D2-SPNs play predominant
585 roles in learning from positive and negative dopamine responses^{32,72–75}. In support of this idea,
586 recent experiments have demonstrated that PKA signaling in D1- and D2-SPNs is primarily
587 driven by a phasic increase and decrease of dopamine, respectively³⁵. Furthermore, LTP-like
588 changes in D1- and D2-SPNs are triggered by a phasic increase and decrease of dopamine,
589 respectively^{33,34}. These recent pieces of evidence suggest that these plasticity rules are a basic
590 principle of the RL circuitry in the brain. Here we explored the properties of this RL model and
591 found the impact of the shape (slope) of receptor occupancy curves and showed that the tonic
592 dopamine levels can modulate the relative efficacy of learning from positive versus negative
593 RPEs.

594

595 One assumption in our model is that after a change in the tonic dopamine level, intracellular
596 signaling reaches a steady inactive state, and it is the *change* in receptor occupancy that matters
597 for inducing synaptic plasticity, rather than the *absolute* level of receptor occupancy reached
598 during phasic dopamine responses. We note that absolute level might also contribute, yet it is

599 expected that an increase or decrease in absolute occupancy levels will cause effects in the same
600 direction as the effects of relative change that we explored in this study.

601
602 Additionally, our model, which incorporates the new plasticity rules, the opponent circuit
603 architecture and properties of D1/D2 dopamine receptors, provides insights into the basic design
604 principle of the brain's RL circuit. It should be noted that the dose occupancy curves were
605 plotted as a function of the logarithm of dopamine concentration, which makes the occupancy
606 curves into sigmoidal shapes (Fig. 3, Extended Data Fig. 9). This logarithmic scaling is
607 important in two ways. First, considering two sigmoidal curves for D1R and D2R together, the
608 curves are approximately *symmetric* around the normal baseline dopamine level (Fig. 3a,
609 Normal). Second, logarithmic scaling means that a fold-change in dopamine concentration will
610 lead to the same leftward or rightward shift in these plots. It has long been argued that signaling
611 of RPEs by dopamine neurons is curtailed by the fact that dopamine neurons have relatively low
612 firing rates (2-8 spikes per second), and inhibitory responses of dopamine neurons tend to be
613 smaller than excitatory responses^{76,77}. Importantly, if we consider logarithmic scaling of
614 dopamine concentration, the problem of this asymmetry is substantially mitigated (Extended
615 Data Fig. 10). For example, with the baseline firing of 6 spikes per second, a phasic increase to
616 18 spikes per second and a phasic decrease to 2 spikes per second will cause the identical *fold-*
617 *changes* in spiking (i.e., 3-fold changes in both directions), which would lead to a similar *fold-*
618 *changes* in dopamine levels (Extended Data Fig. 11) and similar *percent* increase and decrease in
619 receptor occupancy in D1R and D2R, respectively (Fig. 3a). Consequently, the system achieves
620 symmetry in its response to positive and negative dopamine responses of observed magnitudes.

621
622 This may help understand *why* the basal ganglia circuit employs the opponent circuit architecture
623 in the first place. In the model used in the present study, the value is encoded as the difference
624 between the activity of D1- and D2-SPNs ($V = P - N$)³². We propose that this opponent circuit
625 architecture, together with the logarithmic scaling of dopamine concentration, allows the system
626 to effectively learn and encode both positive and negative values, which are contributed by the

627 increase of firing in D1- and D2-SPNs, respectively. This would allow to expand the dynamic
628 range of value coding, without requiring high baseline firing rates. Thus, at the normal dopamine
629 baseline, learning from positive and negative dopamine responses is well balanced. When the
630 tonic dopamine level deviates from the normal level, however, then the symmetry is broken and
631 value learning becomes biased, as explored in the present study.

632

633 **The role of tonic dopamine levels in psychiatric disorders**

634 As mentioned above, our modeling results provide an account for biased value predictions
635 observed in various psychiatric and neurological conditions. For one, our model provides a link
636 between findings in depressive-like states in animal models and the value learning biases
637 exhibited by humans.

638

639 In a rodent model of depression, it has been reported that spontaneous activity of dopamine
640 neurons is decreased⁷⁸ (but see^{79,80}). In addition, decreased spontaneous firing of dopamine
641 neurons has been observed as a result of chronic pain-induced adaptations that correlate with
642 anhedonia-like behavior⁸¹. Furthermore, maternal deprivation, which increases susceptibility to
643 anhedonia, led to an upregulation of D2R expression in the VTA⁸², which is expected to decrease
644 the excitability of dopamine neurons via its autoreceptor function. Finally, chronic
645 administration of corticosteroids, a method to mimic anxiety and anhedonia-like states, results in
646 an increase in somatodendritic dopamine concentration which then decreases dopamine
647 excitability via D2R hyper-activation⁸³. These results of decreased dopamine excitability
648 correlated with anhedonia-like states are consistent with findings of increased burst firing of
649 lateral habenula (LHb) neurons⁵⁶ and potentiation of glutamatergic inputs onto the habenula⁵⁷ in
650 depression models. This is further supported by reports that depressive-like behavioral
651 phenotypes can be ameliorated by optogenetic activation of dopamine neurons⁸⁴ and the anti-
652 depressant effects of ketamine might be mediated by the inhibition of bursting in the LHb⁵⁸

653

654 The mechanism by which a broad change in dopamine excitability could lead to depressive-like
655 states remains to be revealed. Just by assuming that a decrease in spontaneous firing leads to a
656 decrease in baseline dopamine level in the striatum, our model readily predicts that learning from
657 negative outcomes will be emphasized over learning from positive outcomes (Fig. 3a,b), as has
658 been reported in some studies of patients with major depressive disorder (MDD)¹. In addition,
659 RL agents learning in these conditions exhibit enhanced risk-averse behavior, pessimistic
660 outcome expectations, and increased sensitivity to losses compared to gains, all of which are
661 signatures of depressive-like conditions^{1,5,21,85,86}. This contrasts with findings of increased
662 dopamine synthesis capacity in pathological gambling patients⁸⁷, who show the opposite
663 behavioral signatures³.

664
665 An additional line of research relevant to our proposal is PD patients and pathological gambling
666 as a comorbidity. Previous work has emphasized the interaction between the degree of
667 dopaminergic loss and the effects of PD medications⁸⁸⁻⁹⁰, which can sometimes result in the
668 development of addictive disorders such as pathological gambling. As mentioned, the loss of
669 dopaminergic axons in PD patients has been reported to happen predominantly in the dorsal
670 regions of the striatum⁷⁰. Thus, at the onset of the motor impairment symptoms, which is when
671 L-DOPA medication tends to be prescribed, dopamine level is expected to be low in the dorsal
672 striatum while it might be relatively intact in the ventral striatum. This can lead to ‘overdose’ of
673 dopamine by medication: while L-DOPA might take dopamine levels in the dorsal striatum back
674 to its original set-point, it might cause an ‘overdose’ in the ventral striatum^{89,91}. Our model
675 predicts that this overdose would lead to decreases in D2R sensitivity relative to D1R. Assuming
676 that the ventral striatal regions have a dominant role in value learning, this would result in
677 excessive optimistic expectations and risk seeking, two key behavioral features of pathological
678 gambling and addictive disorders. We provided indirect evidence for this hypothesis; future work
679 should directly test these predictions.

680

681 It should be noted that we did not consider changes in dopamine receptors density, which have
682 also been related to value learning biases⁹² and psychiatric conditions⁹³. Future studies should
683 explore the influence of this additional factor in the encoding of asymmetric learning rates (i.e.,
684 $(\hat{\alpha}_i^+, \hat{\alpha}_i^-)$).

685

686 **Tonic dopamine as a modulator of ‘mood’**

687 Mood refers to a person’s emotional state as it relates to their overall sense of well-being.

688 Although the exact neural substrate of mood remains unknown, recent studies have indicated that
689 mood reflects not the absolute goodness of outcomes but rather on the discrepancy between
690 actual and expected outcomes in recent history^{13,14}. That is, mood depends on the cumulative
691 sum of RPEs that occurred recently¹³. It has also been proposed that mood, in turn, affects the
692 way we perceive and learn from positive and negative outcomes (RPEs)¹³.

693

694 Our model provides a unified mechanism for these two aspects of mood; both subjective feeling
695 of mood and biased learning from positive versus negative outcomes can arise from changes in
696 baseline dopamine levels which can be modulated by recent history of phasic dopamine
697 responses. It was proposed that this history dependent modulation of learning is an adaptive
698 mechanism that allows organisms to adapt quickly to slow changes in environments based on the
699 “momentum” of whether the situation is changing in a better or worse direction on a slow
700 timescale (e.g. seasonal change)^{13,14}. The models presented in the present study may provide
701 mechanistic insights into such mood-dependent modulation of learning and perception.

702

703 **Neural circuits for distributional reinforcement learning (RL)**

704 We examined the possibility that optimistic biases in reward seeking behavior and dopamine cue
705 responses observed in habenula-lesioned mice can be explained by Model 2, either based on risk-
706 sensitive RL (the average response) or distributional RL (responses of a diverse set of individual
707 dopamine neurons). We did not find evidence supporting this possibility. However, the present
708 study makes two important contributions with respect to distributional RL. First, we can show

709 that our model, which incorporated direct and indirect pathway architecture, can support
710 distributional RL (Extended Data Fig. 12, 13). It would be interesting to examine what additional
711 features and functions could be gained by having this opponent architecture. Second, we largely
712 replicated the previous results³⁶ using an independent data set. That is, the signatures of
713 distributional RL were present in this data set (Extended Data Fig. 3-4), and dopamine cue-
714 evoked responses did show an optimistic bias. This provides further evidence for a distributional
715 code in dopamine neurons, and shows that there is an overall elevated distributional
716 representation in dopamine cue responses in habenula lesioned animals.

717

718 **Concluding remarks**

719 Taken together, our biologically inspired RL model provides a foundation to link findings in the
720 brain and formal models of RL. Our work highlights a causal impact of baseline dopamine on
721 biasing future value predictions, which may underlie mood and some abnormalities observed in
722 psychiatric patients and could be used to regulate risk sensitive behavior.

723

724

725 **Methods**

726 **1. Reinforcement learning model**

727 Here we provide formal definitions and the framework of reinforcement learning used in this study. We
728 have focused our model formulations to the problem of *prediction*, in which an agent learns to predict the
729 value function¹¹. The problem of *control* (the problem of how an agent selects and executes actions) is not
730 considered. In RL, an agent's objective is to maximize the total cumulative rewards. It does so by learning
731 the value associated with each state in an environment. For now, we will develop the model dropping the
732 dependency on time within each episode. Here, the target to learn is the value function as defined by

$$733 \quad V(s_i) := \mathbf{E}[r^{(n)} | s^{(n)} = s_i]$$

734 Where $r^{(n)}$ is the reward experienced in the episode n (i.e., trial) of visiting state s_i . Learning of $V(s)$ is
735 driven by reward prediction errors (RPEs, δ), the discrepancy between the actual and expected reward:

$$736 \quad \delta^{(n)} = r^{(n)} - V(s_i)$$

737 The value is updated for the experienced state according to:

$$738 \quad V^{(n+1)}(s_i) \leftarrow V^{(n)}(s_i) + \alpha \cdot \delta^{(n)}$$

739 This is also known as the Rescorla-Wagner (RW) delta rule⁹⁴. The reward in each trial is sampled from a
740 reward distribution specific to a given state: $r^{(n)} \sim R(s_i)$. With the learning rule above, the value
741 converges on the expected value of this reward distribution. This can be shown with a stochastic fixed-
742 point approach; the convergence point is derived by obtaining the value of $V(s_i)$ at which the change in
743 $V(s_i)$ from trial n to trial $(n + 1)$ is expected to be zero (i.e., is zero on average):

$$744 \quad \mathbf{E}[V^{(n+1)}(s_i) - V^{(n)}(s_i)] = 0$$

$$745 \quad \mathbf{E}[\alpha \cdot \delta^{(n)}] = 0$$

$$746 \quad \mathbf{E}[\alpha \cdot (r^{(n)} - V(s_i))] = 0$$

$$747 \quad \alpha \cdot \mathbf{E}[r] - \alpha \cdot \mathbf{E}[V(s_i)] = 0$$

$$748 \quad \mathbf{E}[V(s_i)] = \mathbf{E}[r_t]$$

$$749 \quad V^*(s_i) = \mathbf{E}[r_t]$$

750 Where $V^*(s_i)$ is the stochastic fixed-point: the value around which $V(s_i)$ is expected to fluctuate after
751 learning and corresponds to the learning target above.

752 1.1. Temporal difference learning

753 Now we will consider time and extend the models to the temporal difference (TD) learning framework¹¹.
754 Dopamine responses have been shown to present key signatures of TD errors⁹⁵. Therefore, TD learning
755 models allow us to directly link the model variables to dopamine neural responses.

756 We can derive TD learning by defining a different environmental structure and learning objective. We
757 start by considering arbitrary states (s_t), which transition at each time step following a Markov process,
758 and at each time step the agent samples a random reward from a probability distribution $r_t \sim R(s_t)$.

759 The learning objective is now the value of a given state $V(s_t)$ defined as the *expected cumulative sum of*
760 *all future rewards* starting from state s . Rewards are discounted by a constant discounting factor (γ , with
761 $0 \leq \gamma \leq 1$) each time step. The expectation is taken over stochastic state transitions and sampled rewards:

$$762 \quad V(s_t) := \mathbf{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} \dots | s_t = S]$$

763 Where s_t is the state at time t , r_t is the reward sampled at time t and $V(s_t)$ is the value of the state s_t .
764 Since the environment and transitions are assumed to follow a Markov process, the equation above can be
765 rewritten in a recursive manner. This is known as the Bellman equation¹¹ :

$$766 \quad V(s_t) := \mathbf{E}[r_t + \gamma \cdot V(s_{t+1}) | s_t = S]$$

767 The agent approximates the true value $V(s_t)$ with a learned estimate $\hat{V}(s_t)$. With this approximation,
768 before learning converges, the estimates for the left- and right-hand sides are not equal. Thus, after
769 sampling a reward $r_t \sim R(s_t)$ from the environment, the difference between the two terms in the Bellman
770 equation represents the error in value prediction, called the temporal difference reward prediction error
771 (TD RPE, δ below),

$$772 \quad \delta_t = r_t + \gamma \cdot \hat{V}(s_{t+1}) - \hat{V}(s_t)$$

773 With α as the learning rate, the updates for the value estimates are:

$$774 \quad \hat{V}(s_t) \leftarrow \hat{V}(s_t) + \alpha \cdot \delta_t$$

775 With this definition, the TD RPE contains the difference between the estimated value of states evaluated
776 at consecutive time points. If we fix the discounting factor to be $\gamma = 1$, then $\gamma \cdot \hat{V}(s_{t+1}) - \hat{V}(s_t)$ is the

777 temporal derivative of the value function. As a result of this property, unexpected increases and decreases
778 in value result in positive and negative transient changes in TD RPE, respectively⁹⁵.

779 If dopamine responses encode TD RPEs, then cue-evoked responses can be formulated as:

$$780 \quad \delta_{cue} = \gamma \cdot \hat{V}(s_{cue}) - \hat{V}(s_b)$$

781 Where δ_{cue} is the TD RPE induced by the cue, $\hat{V}(s_b)$ is the value prediction at baseline and $\hat{V}(s_{cue})$ is
782 the value prediction elicited by the cue (which reflects the expected value predicted by each trial type). As
783 the $\hat{V}(s_b)$ is the same across all trial types and represents the average value predictions across them, then
784 δ_{cue} is dominated by the expected value of each trial type. This is a useful feature that we used in our
785 simulations for the habenula lesion experiment.

786 1.2. Distributional TD learning

787 In Results, we used a distributional TD learning model to test whether the subtle changes in the
788 distribution of asymmetric scaling factors observed after lesions could lead to the observed changes in cue
789 responses after learning.

790 In distributional TD learning, our learning objective is the entire distribution over cumulative discounted
791 future rewards, instead of the value defined above^{36,37,39}. We will call this the *return distribution*, $Z(s_t)$.
792 We can thus write an analogue of the Bellman equation, the ‘distributional Bellman equation’:

$$793 \quad Z(s_t) := R(s_t) + \gamma \cdot Z(s_{t+1})$$

794 The target to learn in distributional TD is now $V_i(s_t)$ that minimizes for the expectile regression loss:

$$795 \quad V_i(s_t) := \operatorname{argmin}_v E[(Z(s_t) - v)^2 \cdot (\tau_i - \mathbf{1}_{(Z(s_t)-v)<0})]$$

796 Where $Z(s_t)$ is a random variable, representing the return distribution, and $\mathbf{1}_f$ is the indicator functions
797 that is equal to 1 if the condition in the subscript $\{f := (Z(s_t) - v) < 0\}$ is met, and 0 otherwise.
798 Minimizing the expectile regression loss makes $V_i(s_t)$ to converge on the τ_i^{th} expectile of the return
799 distribution³⁹.

800 The target is learned by taking samples from the estimated return distribution³⁹ $\tilde{z}(s_{t+1}) \sim Z(s_{t+1})$ and
801 from the reward distribution $r_t \sim R(s_t)$, to compute the TD error:

$$802 \quad \delta_{i,t} := r_t + \gamma \cdot \tilde{z}(s_{t+1}) - V_i(s_t)$$

803 Note that $\tilde{z}(s_{t+1})$ is random so the TD error is also random, and $\delta_{i,t} \neq r_t + \gamma \cdot V_i(s_{t+1}) - V_i(s_t)$. For
 804 more information regarding the sampling method employed in the simulations see Methods Section 3.3.
 805 In addition, the updates are performed with different learning rates (α_i^+ , α_i^-) for positive and negative δ_i .
 806 This asymmetry in the weighting of the errors used to update $V_i(s_t)$ is essential to minimize the expectile
 807 regression loss.

$$808 \quad \hat{\delta}_{i,t} = \alpha_i^+ \cdot \delta_{i,t} \dots \text{ if } \delta_{i,t} > 0$$

$$809 \quad \hat{\delta}_{i,t} = \alpha_i^- \cdot \delta_{i,t} \dots \text{ if } \delta_{i,t} < 0$$

810 The reliance on a single sample for $\tilde{z}(s_{t+1})$ suffers from high variance. Therefore, for performing the
 811 updates we average across a set of M updates, each depending on a single sample $\delta_{i,t}$.

$$812 \quad \mathbf{E}[\Delta V_i(s_t)] = \frac{1}{M} \sum_j^M \alpha_i^- \cdot \delta_{i,j} \cdot \mathbf{1}_{\delta_{i,j} < 0} + \alpha_i^+ \cdot \delta_{i,j} \cdot \mathbf{1}_{\delta_{i,j} > 0}$$

$$813 \quad V_i(s_t) \leftarrow V_i(s_t) + \mathbf{E}[\Delta V_i(s_t)]$$

814 This learning rule will asymptotically converge to the τ_i -th expectile of the return distribution³⁹.

815 1.3. TD learning with D1 and D2 populations

816 It is straightforward to extend the TD learning algorithm to have separate populations for D1 and D2
 817 SPNs³². We employed this model to derive dopamine cue responses with Model 1 (Fig. 6i). In this model,
 818 the same computation of TD RPE of standard TD learning is still used. Yet, this model differs in the
 819 updates and computation of $\hat{V}(s_t)$.

820 As mentioned previously, the updates in the P_i and N_i populations happen exclusively with positive or
 821 negative TD RPEs, respectively:

$$822 \quad P(s_t) \leftarrow P(s_t) + \alpha^+ \cdot |\delta_t| - \beta \cdot P(s_t) \dots \text{ if } \delta_t > 0$$

$$823 \quad N(s_t) \leftarrow N(s_t) + \alpha^- \cdot |\delta_t| - \beta \cdot N(s_t) \dots \text{ if } \delta_t < 0$$

824 Where α^+ and α^- are the learning rates for the P and N populations, that we postulate is modulated by
 825 baseline dopamine levels. The variable $\beta \in (0,1)$ is the decay factor, which we keep constant throughout
 826 the simulations and serves to stabilize $P(s_t)$, $N(s_t)$.

827 The computation of value estimate $\hat{V}(s_t)$ is given by:

828
$$\hat{V}(s_t) = P(s_t) - N(s_t)$$

829 **1.3.1. Convergence of risk sensitive TD learning**

830 We now discuss the convergence of the proposed TD learning algorithm with D1 and D2 populations.
 831 This analysis builds on the work in risk-sensitive reinforcement learning⁴⁰ and the already established
 832 results of convergence for stochastic iterative algorithms (e.g., TD learning) (Bertsekas & Tsitsiklis,
 833 1996⁹⁶, Proposition 4.4, p. 156) .

834 **Theorem:** The results by Bertsekas & Tsitsiklis (1996)⁹⁶ establish that, given a sequence $r_t \in$
 835 \mathbb{R}^m generated by the iterative algorithm:

836
$$a_{n+1}(s) = (1 - \sigma_n(s))a_n(s) + \sigma_n(s)((Ha_n)(s) + \omega_n(s)) \quad \forall s \in 1, \dots, m \quad \text{Eq. I}$$

837 The variable a_n converges to the unique solution a^* of the equation: $Ha^* = a^*$ with probability = 1,
 838 assuming the following conditions are fulfilled:

839 1. The step sizes $\sigma_i(i)$ are non-negative and satisfy:

840
$$\sum_{n=0}^{\infty} \sigma_n(s) = \infty \quad \forall s \in 1, \dots, m$$

841
$$\sum_{n=0}^{\infty} \sigma_n(s)^2 < \infty \quad \forall s \in 1, \dots, m$$

842 2. The noise term $\omega_n(s)$ satisfies:

843 - $E[\omega_n(s)|\mathcal{F}_n] = 0 \quad \forall s, n$, where \mathcal{F}_n denotes the history of the process up to and including time
 844 step n

845 - Given any norm $\|\cdot\|$ on \mathbb{R}^m there exist constants A and B such that: $E[\omega_n^2(s)|\mathcal{F}_n] \leq A +$
 846 $B\|r_n\|^2 \quad \forall s, n$

847 3. The mapping H is a *maximum norm contraction* (see below for definition)

848 To prove convergence, we will first discuss the case of risk-sensitive TD learning following⁴⁰ and then
 849 discuss TD learning with D1 and D2 populations.

850 We define the risk sensitive TD-learning rule as:

851
$$\hat{V}_n(s) \leftarrow \hat{V}_{n-1}(s) + \sigma \cdot \mathcal{X}^\tau(\delta_{s_{n-1}, s_n})$$

852 Where:

853 $\delta_{s_{n-1}, s_n} = r_{n-1, n} + \gamma \cdot \hat{V}_{n-1}(s_n) - \hat{V}_n(s_{n-1})$

854 - The step index is $n \in 0, \dots, \infty$

855 - The step size σ is kept constant across iterations.

856 - For simplicity in calculations we follow ⁴⁰ and make use of the operator \mathcal{X}^K with $K \in (-1, 1)$

857
$$\mathcal{X}^K(x) = (1 - K) \cdot x \quad \dots \text{if } x > 0$$

858
$$\mathcal{X}^K(x) = (1 + K) \cdot x \quad \dots \text{if } x \leq 0$$

859 It is simple to show that the asymmetric scaling factor used in this paper is a scaled version of
860 the operator. That is: $\tau = 0.5(1 - K)$ and $1 - \tau = 0.5(1 + K)$.

861 - In addition, as in ⁴⁰, given that the function $\mathcal{X}^\tau(x)$ is piece-wise differentiable we can apply
862 the mean value theorem to show that for each pair of numbers (a, b) there exists a $\varepsilon_{a,b,K} \in$
863 $[1 - |K|, 1 + |K|]$, such that: $\varepsilon_{a,b,K} = \frac{\mathcal{X}^\tau(a) - \mathcal{X}^\tau(b)}{a - b}$. This relationship will become useful in
864 the future.

865 We will re-format the update rule to better match the iterative algorithm above:

866 Adding and subtracting $\sigma \cdot \hat{V}_{n-1}(s) / \alpha$

867
$$\hat{V}_n(s) \leftarrow (1 - \sigma/\alpha) \hat{V}_{n-1}(s) + \sigma/\alpha \left(\alpha \cdot \mathcal{X}^\tau(\delta_{n-1}) + \hat{V}_{n-1}(s) \right)$$

868 Defining an operator that will become useful:

869
$$\mathcal{J}_{\alpha K}[V](s) := V(s) + \alpha \cdot \sum_{i,j \in \mathcal{S}} p_{ij} \cdot \mathcal{X}^K \cdot \delta_{ij}$$

870 Defining the noise term as:

871
$$\omega_{n-1}(s) = \hat{V}_{n-1}(s) + \alpha \cdot \mathcal{X}^K(\delta_{s_{n-1}, s_n}) - \mathcal{J}_{\alpha K}[\hat{V}_{n-1}](s)$$

872 Then our update rule above becomes:

873
$$\hat{V}_n(s) \leftarrow (1 - \sigma/\alpha) \hat{V}_{n-1}(s) + \sigma/\alpha \left(\mathcal{J}_{\alpha K}[\hat{V}_{n-1}](s) + \omega_{n-1}(s) \right)$$

874 The formulation above can be directly compared to the one of stochastic iterative algorithm theorem

875 (Eq.I), and now we can check whether the conditions for convergence are met.

- 876 1. The conditions for the learning rate, are a direct consequence of our choice of the parameter
877 which is a constant in our model and $0 < \alpha < 1$.
- 878 2. It has been shown that showed that the conditions for the noise term $\omega_{n-1}(s)$ as formulated
879 above are satisfied⁴⁰.
- 880 3. Finally, the operator $\mathcal{T}_{\alpha\tau}[V](s)$ is a contraction mapping as also shown in Bersekas (1996)⁴⁰.

881 Therefore, the variable V_n converges to the unique solution V^* for which:

$$882 \quad V^* = \mathcal{T}_{\alpha K}[V^*] = V^* + \alpha \cdot \sum_{i,j \in \mathcal{S}} p_{ij} \cdot \mathcal{X}^K \cdot \delta_{ij}$$

883 We elaborate now on the proof for the contraction mapping of the operator $\mathcal{T}_{\alpha\tau}[V](s)$, as this will be
884 useful for the proof of the D1 D2 TD learning model.

885 **Definition of contraction mapping.** Let (X, d) be a metric space (a set X , with a notion of distance, d ,
886 between points). A mapping $\mathcal{T}: X \rightarrow X$ is a contraction mapping if there exists a constant $c: 0 \geq c > 1$
887 such that for all $x \in X$:

$$888 \quad d(\mathcal{T}[x_i], \mathcal{T}[x_j]) \leq cd(x_i, x_j)$$

889 That is, a contraction mapping maps points closer together.

890 Elaborating now on the operator $\mathcal{T}_{\alpha\tau}[V](s)$ and using $|\cdot|$ as our distance metric:

$$891 \quad |\mathcal{T}_{\alpha K}[V_1](i) - \mathcal{T}_{\alpha K}[V_2](i)|$$

$$892 \quad = \left| V_1(i) + \alpha \cdot \sum_{i,j \in \mathcal{S}} p_{ij} \cdot \mathcal{X}^K \cdot (r_{i,j} + \gamma V_1(j) - V_1(i)) - V_2(i) + \alpha \right.$$

$$893 \quad \left. \cdot \sum_{i,j \in \mathcal{S}} p_{ij} \cdot \mathcal{X}^K \cdot (r_{i,j} + \gamma V_2(j) - V_2(i)) \right|$$

$$\begin{aligned}
 894 \quad & |\mathcal{T}_{\alpha K}[V_1](i) - \mathcal{T}_{\alpha K}[V_2](i)| \\
 895 \quad & = \left| V_1(i) - V_2(i) \right. \\
 896 \quad & \left. + \alpha \sum_{i,j \in S} p_{ij} \cdot \left(\mathcal{X}^\tau \cdot (r_{i,j} + \gamma V_1(j) - V_1(i)) - \mathcal{X}^\tau \cdot (r_{i,j} + \gamma V_2(j) - V_2(i)) \right) \right|
 \end{aligned}$$

897 Using the relation defined above $\mathcal{E}_{a,b,K} \cdot (a - b) = \mathcal{X}^K(a) - \mathcal{X}^K(b)$

$$\begin{aligned}
 898 \quad & |\mathcal{T}_{\alpha K}[V_1](i) - \mathcal{T}_{\alpha K}[V_2](i)| \\
 899 \quad & = \left| V_1(i) - V_2(i) + \alpha \sum_{i,j \in S} p_{ij} \cdot \mathcal{E}_{V_1, V_2, K} \cdot (\gamma(V_1(j) - V_2(i)) - (V_1(j) - V_2(i))) \right|
 \end{aligned}$$

$$\begin{aligned}
 900 \quad & |\mathcal{T}_{\alpha K}[V_1](i) - \mathcal{T}_{\alpha K}[V_2](i)| \\
 901 \quad & = \left| \left(1 - \alpha \sum_{i,j \in S} p_{ij} \cdot \mathcal{E}_{V_1, V_2, K} \right) \cdot (V_1(j) - V_2(i)) \right. \\
 902 \quad & \left. + \sum_{i,j \in S} p_{ij} \cdot \mathcal{E}_{V_1, V_2, K} \cdot (\gamma(V_1(j) - V_2(i))) \right|
 \end{aligned}$$

903

904 Given that $\mathcal{E}_{a,b,K} \in [1 - |K|, 1 + |K|]$ and assuming $\alpha \in (0, (1 + |K|)^{-1})$:

$$905 \quad 1 - \alpha \cdot \sum_{i,j \in S} p_{ij} \cdot \mathcal{E}_{V_1, V_2, K} > 0$$

906 Taking this term outside the $|\cdot|$ and rearranging:

$$907 \quad |\mathcal{T}_{\alpha K}[V_1](i) - \mathcal{T}_{\alpha K}[V_2](i)| = \left(1 - \alpha \cdot (1 - \gamma) \cdot \sum_{i,j \in S} p_{ij} \cdot \mathcal{E}_{V_1, V_2, K} \right) |(V_1(j) - V_2(i))|$$

$$908 \quad |\mathcal{T}_{\alpha K}[V_1](i) - \mathcal{T}_{\alpha K}[V_2](i)| = c \cdot |(V_1(j) - V_2(i))|$$

909 Where the term:

$$910 \quad c = \left(1 - \alpha \cdot (1 - \gamma) \cdot \sum_{i,j \in S} p_{ij} \cdot \mathcal{E}_{V_1, V_2, K} \right)$$

911 To get the upper boundary of c we use the minimum value for the sum, where $\mathcal{E}_{V_1 V_2 K} = 1 - |K| \quad \forall i, j \in$
 912 S . And use the assumption that $\alpha \in (0, (1 + |K|)^{-1})$:

$$913 \quad c \leq (1 - \alpha \cdot (1 - \gamma) \cdot (1 - |K|))$$

$$914 \quad \lim_{\alpha \rightarrow 0} c = (1 - \alpha \cdot (1 - \gamma) \cdot (1 - |K|)) = 1$$

915 To get the lower boundary of c we use the maximum value for the sum, where $\mathcal{E}_{V_1 V_2 K} = 1 + |K| \quad \forall i, j \in$
 916 S . And use the assumption that $\alpha \in (0, (1 + |K|)^{-1})$:

$$917 \quad c \geq (1 - \alpha \cdot (1 - \gamma) \cdot (1 + |K|))$$

$$918 \quad \lim_{\alpha \rightarrow (1+|K|)^{-1}} c = (1 - (1 + |K|)^{-1} \cdot (1 - \gamma) \cdot (1 + |K|)) = \gamma$$

919 Therefore: $\gamma < c < 1$

$$920 \quad |\mathcal{T}_{\alpha K}[V_1](i) - \mathcal{T}_{\alpha K}[V_2](i)| \leq c \cdot |(V_1(j) - V_2(i))|$$

921 And the operator $\mathcal{T}_{\alpha \tau}[V](s)$ is a contraction mapping, under the condition $\alpha \in (0, (1 + |K|)^{-1})$

922 **1.3.2. Convergence of TD learning with D1 and D2 populations**

923 We define the D1-D2 TD-learning rule as:

$$924 \quad \hat{V}_n(s) \leftarrow \hat{V}_{n-1}(s) + \alpha \cdot \mathcal{X}^\tau(\delta_{s_{n-1}, s_n}) - \beta \hat{V}_{n-1}(s)$$

925 Note this update rule is analogous to the risk sensitive TD learning rule except for the last term that
 926 emerges from the decay factor in the P, N populations of our model.

927 Performing the same re-arrangement as above we reach:

$$928 \quad \hat{V}_n(s) \leftarrow (1 - \sigma/\alpha) \hat{V}_{n-1}(s) + \sigma/\alpha (\alpha \cdot \mathcal{X}^\tau(\delta_{n-1}) + \hat{V}_{n-1}(s) - \alpha \cdot \beta \hat{V}_{n-1}(s))$$

929 We define a new operator $\mathcal{T}'_{\alpha K}[V](s)$:

$$930 \quad \mathcal{T}'_{\alpha K}[V](s) := (1 - \alpha \cdot \beta) \cdot V(s) + \alpha \cdot \sum_{i,j \in S} p_{ij} \cdot \mathcal{X}^K \cdot \delta_{ij}$$

931 And the noise then is defined as:

$$932 \quad \omega_{n-1}(s) = \alpha \cdot \mathcal{X}^K(\delta_{s_{n-1}, s_n}) + \hat{V}_{n-1}(s) - \alpha \cdot \beta \cdot \hat{V}_{n-1}(s) - \mathcal{T}'_{\alpha K}[\hat{V}_{n-1}](s)$$

933 The update becomes:

$$934 \quad \hat{V}_n(s) \leftarrow (1 - \sigma/\alpha)\hat{V}_{n-1}(s) + \sigma/\alpha \left(\mathcal{T}'_{\alpha K}[\hat{V}_{n-1}](s) + \omega_{n-1}(s) \right)$$

935 The noise term reduces to the same expression as the one of TD learning, and so it fullfils the
 936 requirements for the theorem of stochastic iterative algorithms. We will now test whether the operator
 937 $\mathcal{T}'_{\alpha K}[V](s)$ also represents a contraction map.

$$938 \quad | \mathcal{T}_{\alpha K}[V_1](i) - \mathcal{T}_{\alpha K}[V_2](i) |$$

$$939 \quad = \left| V_1(i) + \alpha \cdot \sum_{i,j \in \mathcal{S}} p_{ij} \cdot \mathcal{X}^K \cdot (r_{i,j} + \gamma V_1(j) - V_1(i)) - \alpha \cdot \beta \cdot V_1(i) + V_2(i) + \alpha \right.$$

$$940 \quad \left. \cdot \sum_{i,j \in \mathcal{S}} p_{ij} \cdot \mathcal{X}^K \cdot (r_{i,j} + \gamma V_2(j) - V_2(i)) - \alpha \cdot \beta \cdot V_2(i) \right|$$

$$941 \quad | \mathcal{T}_{\alpha K}[V_1](i) - \mathcal{T}_{\alpha K}[V_2](i) |$$

$$942 \quad = \left| V_1(i) - V_2(i) - \alpha \cdot \beta (V_1(i) - V_2(i)) \right.$$

$$943 \quad \left. + \alpha \sum_{i,j \in \mathcal{S}} p_{ij} \cdot \left(\mathcal{X}^\tau \cdot (r_{i,j} + \gamma V_1(j) - V_1(i)) - \mathcal{X}^\tau \cdot (r_{i,j} + \gamma V_2(j) - V_2(i)) \right) \right|$$

944 Using the relation defined above $\mathcal{E}_{a,b,K} \cdot (a - b) = \mathcal{X}^K(a) - \mathcal{X}^K(b)$

$$945 \quad | \mathcal{T}_{\alpha K}[V_1](i) - \mathcal{T}_{\alpha K}[V_2](i) |$$

$$946 \quad = \left| (1 - \alpha\beta) \cdot (V_1(i) - V_2(i)) \right.$$

$$947 \quad \left. + \alpha \sum_{i,j \in \mathcal{S}} p_{ij} \cdot \mathcal{E}_{V_1, V_2, K} \cdot \left(\gamma (V_1(j) - V_2(i)) - (V_1(j) - V_2(i)) \right) \right|$$

$$\begin{aligned}
 & | \mathcal{T}_{\alpha K}[V_1](i) - \mathcal{T}_{\alpha K}[V_2](i) | \\
 & = \left| \left(1 - \alpha \cdot \beta - \alpha \sum_{i,j \in S} p_{ij} \cdot \mathcal{E}_{V_1, V_2, K} + \gamma \alpha \sum_{i,j \in S} p_{ij} \cdot \mathcal{E}_{V_1, V_2, K} \right) \cdot (V_1(j) - V_2(i)) \right|
 \end{aligned}$$

$$| \mathcal{T}_{\alpha K}[V_1](i) - \mathcal{T}_{\alpha K}[V_2](i) | = c \cdot |(V_1(j) - V_2(i))|$$

Where the term:

$$c = \left(1 - \alpha \cdot \beta - \alpha \cdot (1 - \gamma) \cdot \sum_{i,j \in S} p_{ij} \cdot \mathcal{E}_{V_1, V_2, K} \right)$$

To get the upper boundary of c we use the minimum value for the sum, where $\mathcal{E}_{V_1 V_2 K} = 1 - |K| \quad \forall i, j \in S$, and use the assumption that $\alpha \in (0, (1 + |K|)^{-1})$:

$$c \leq 1 - \alpha \cdot \beta - \alpha \cdot (1 - \gamma) \cdot (1 - |K|)$$

$$\lim_{\alpha \rightarrow 0} c = 1 - \alpha \cdot \beta - \alpha \cdot (1 - \gamma) \cdot (1 - |K|) = 1$$

To get the lower boundary of c we use the maximum value for the sum, where $\mathcal{E}_{V_1 V_2 K} = 1 + |K| \quad \forall i, j \in S$, and use the assumption that $\alpha \in (0, (1 + |K|)^{-1})$:

$$c \geq (1 - \alpha \cdot \beta - \alpha \cdot (1 - \gamma) \cdot (1 + |K|))$$

$$\lim_{\alpha \rightarrow (1+|K|)^{-1}} c = (1 - (1 + |K|)^{-1} \cdot \beta - (1 + |K|)^{-1} \cdot (1 - \gamma) \cdot (1 + |K|)) = \gamma - (1 + |K|)^{-1} \cdot \beta$$

Given that we want $c \geq 0$ we can find the parameter ranges to achieve this:

$$c = \gamma - (1 + |K|)^{-1} \cdot \beta \geq 0$$

Given that: $K \in (-1, 1)$, we use the minimum value of $|K| = 0$ to find the limit of c :

$$\lim_{|K| \rightarrow 0} c = \gamma - \beta$$

So the condition $\gamma \geq \beta$ needs to be present to keep: $0 \leq c < 1$.

Under these conditions, the operator $\mathcal{T}'_{\alpha K}[V](s)$ also represents a contraction map.

$$| \mathcal{T}'_{\alpha K}[V_1](i) - \mathcal{T}'_{\alpha K}[V_2](i) | \leq c \cdot |(V_1(j) - V_2(i))|$$

Stochastic fixed point for the value estimate:

969 Having shown convergence of the algorithm, we will now derive the convergent points for our algorithm
 970 using stochastic fixed points. For clarity, we estimate the stochastic fixed point dropping the dependency
 971 on time.

972 If learning between P and N is symmetric $\alpha^+ = \alpha^- = \alpha$. We derive the convergent estimate of $V(s_t)$
 973 with a fixed-point approach. First, we subtract the P and N update equations, to arrive to the update in
 974 the $\hat{V}(s_t)$ between the (n) and the $(n + 1)$ update:

$$975 \quad \hat{V}^{(n+1)}(s_i) \leftarrow \hat{V}^{(n)}(s_i) + \alpha \cdot \delta^{(n)} - \beta \cdot \hat{V}^{(n)}(s_i)$$

976 Where the superscripts indicate the iteration number. We can now derive the stochastic fixed-point
 977 for $\hat{V}(s_t)$:

$$978 \quad \mathbf{E}[\hat{V}^{(n+1)}(s_i) - \hat{V}^{(n)}(s_i)] = 0$$

$$979 \quad \mathbf{E}[\alpha \cdot \delta^{(n)} - \beta \cdot \hat{V}^{(n)}(s_i)] = 0$$

$$980 \quad \mathbf{E}[\alpha \cdot (r^{(n)} - \hat{V}(s_i)) - \beta \cdot \hat{V}^{(n)}(s_i)] = 0$$

$$981 \quad \alpha \cdot \mathbf{E}[r] - (\alpha + \beta) \cdot \mathbf{E}[\hat{V}(s_i)] = 0$$

$$982 \quad \mathbf{E}[\hat{V}(s_i)] = \frac{\alpha}{\alpha + \beta} \mathbf{E}[r]$$

$$983 \quad \mathbf{V}^* = \frac{\alpha}{\alpha + \beta} \mathbf{E}[r]$$

984 Where $\mathbf{V}^* = \mathbf{E}[\hat{V}(s_i)]$ is the value around which $\hat{V}(s_i)$ is expected to fluctuate after convergence.

985 Throughout this study, we have manipulated the learning rates between P and N to be asymmetric $\alpha^+ \neq$
 986 α^- or, equivalently, $\tau \neq 1 - \tau$. We can find the stochastic fixed point for this more general case:

$$987 \quad \mathbf{E}[\hat{V}^{(n+1)}(s_i) - \hat{V}^{(n)}(s_i)] = 0$$

$$988 \quad \mathbf{E}[\tau \cdot |\delta^{(n)}| \cdot \mathbf{1}_{\delta > 0} - (1 - \tau) \cdot |\delta^{(n)}| \cdot \mathbf{1}_{\delta < 0} - \beta \cdot \hat{V}^{(n)}(s_i)] = 0$$

989 To take the expectation we use the definition: $\mathbf{E}[X] = \sum_i p(x_i) \cdot x_i$. For a Bernoulli distribution, $p(x_i)$
 990 takes two values:

- 991 • $p(x_i) = p$ if reward is delivered and, thus $r = 1$, $\delta_t > 0$,
- 992 • $p(x_i) = (1 - p)$ if reward is not delivered and, thus $r = 0$, $\delta_t < 0$,

993 Therefore, we can resolve the expectation and expand the RPEs:

$$994 \quad \mathbf{E}[\tau \cdot |\delta^{(n)}| \cdot \mathbf{1}_{\delta > 0} - (1 - \tau) \cdot |\delta^{(n)}| \cdot \mathbf{1}_{\delta < 0} - \beta \cdot \hat{V}^{(n)}(s_i)] = 0$$

$$995 \quad \tau \cdot \mathbf{E}[|r - \hat{V}(s_i)| \cdot \mathbf{1}_{\delta > 0}] - (1 - \tau) \cdot \mathbf{E}[|-\hat{V}(s_i)| \cdot \mathbf{1}_{\delta < 0}] - \beta \cdot \mathbf{E}[\hat{V}(s_i)] = 0$$

996 Taking the absolute values:

$$997 \quad |r - \hat{V}(s_i)| = r - \hat{V}(s_i) \dots \text{ if } (r - \hat{V}(s_i)) > 0$$

$$998 \quad |-\hat{V}(s_i)| = \hat{V}(s_i) \dots \text{ if } (-\hat{V}(s_i)) < 0$$

$$999 \quad \tau \cdot \mathbf{E}[(r - \hat{V}(s_i)) \cdot \mathbf{1}_{\delta > 0}] - (1 - \tau) \cdot \mathbf{E}[\hat{V}(s_i) \cdot \mathbf{1}_{\delta < 0}] - \beta \cdot \mathbf{E}[\hat{V}(s_i)] = 0$$

1000 Replacing stochastic fixed point: $\mathbf{E}[\hat{V}(s_t)] = V^*$ and taking the expectations:

$$1001 \quad \beta \cdot V^* = \tau \cdot p \cdot (r - V^*) - (1 - \tau) \cdot (1 - p) \cdot V^*$$

1002 Rearranging and isolating V^* , we obtain:

$$1003 \quad V^* = \frac{\frac{\tau}{1 - \tau} \cdot \frac{p}{1 - p} \cdot r}{\frac{\tau}{1 - \tau} \cdot \frac{p}{1 - p} + 1 + \frac{\beta}{(1 - \tau) \cdot (1 - p)}}$$

1004 *Stochastic fixed point for P and N populations:*

1005 We have mentioned that the decay term (β) in the update equations serves to stabilize the estimates of the
 1006 P and N populations (i.e., avoid infinite increases). We can observe the influence of β by computing the
 1007 stochastic fixed points for these variables.

1008 For the P population:

$$1009 \quad \mathbf{E}[P^{(n+1)}(s_t) - P^{(n)}(s_t)] = 0$$

$$1010 \quad \mathbf{E}[\tau \cdot |r - V^{(n)}(s_t)| \cdot \mathbf{1}_{\delta > 0} - \beta \cdot P^{(n)}(s_t)] = 0$$

$$1011 \quad p \cdot \tau \cdot (r - V^*) - \beta \cdot P^* = 0$$

$$1012 \quad P^* = \frac{p \cdot \tau}{\beta} \cdot (r - V^*)$$

1013 Similarly, for the N population:

1014
$$\mathbf{E}[N^{(n+1)}(s_t) - N^{(n)}(s_t)] = 0$$

1015
$$\mathbf{E}[(1 - \tau) \cdot |-V^{(n)}| \cdot \mathbf{1}_{\delta < 0} - \beta \cdot N^{(n)}(s_t)] = 0$$

1016
$$(1 - p) \cdot (1 - \tau) \cdot V^* - \beta \cdot N^* = 0$$

1017
$$N^* = \frac{(1 - p) \cdot (1 - \tau)}{\beta} \cdot V^*$$

1018 As it can be seen in the stochastic fixed points P^*, N^* , the term $\frac{1}{\beta}$ is a proportionality constant. Therefore:

1019
$$\lim_{\beta \rightarrow 0} P^* = \lim_{\beta \rightarrow 0} \left(\frac{p \cdot \tau}{\beta} \cdot (r - V^*) \right) = \text{undefined}$$

1020
$$\lim_{\beta \rightarrow 0} N^* = \lim_{\beta \rightarrow 0} \left(\frac{(1 - p) \cdot (1 - \tau)}{\beta} \cdot V^* \right) = \text{undefined}$$

1021 So, $\beta \neq 0$ needs to be met for the stochastic fixed points P^*, N^* to exist. In Extended Data Fig. 1 we show
1022 empirically that the convergence rate is slower as β gets closer to 0, but it is always achieved.

1023 **1.3.3. Sensitivity of learned variables in D1-D2 model to parameters**

1024 The conditions for the D1-D2 model to reproduce the data from our habenula lesion experiment and some
1025 of the previous studies are that:

- 1026 1. The bias in V^* induced by the asymmetric learning rates doesn't change the monotonicity of the
1027 learned values as a function of the true expected value of the return distribution $\mathbf{E}[R(s)]$. In other
1028 words, regardless of the level of 'optimism' or 'pessimism', V^* monotonically increases with
1029 $\mathbf{E}[R(s)]$.
- 1030 2. Asymmetric learning rates change the concavity of V^* as a function of $\mathbf{E}[R(s)]$: 'Optimistic' or
1031 'pessimistic' value functions are concave or convex with respect to $\mathbf{E}[R(s)]$, respectively.

1032

1033 We will now analyze whether these conditions are met, considering the range of parameters of relevance:

1034 $0 < \tau < 1, r \neq 0$ and $0 < \beta < 1$

1035 For the condition 1 to be met, the first derivative of V^* with respect to $\mathbf{E}[R(s)]$ should always be positive.

1036 In the case of Bernoulli return distributions, the derivative of V^* with respect to $p(\text{reward})$ is

$$1037 \quad \frac{\partial V^*}{\partial p} = \frac{\partial}{\partial p} \left(\frac{\frac{\tau}{1-\tau} \cdot \frac{p}{1-p} \cdot r}{\frac{\tau}{1-\tau} \cdot \frac{p}{1-p} + 1 + \frac{\beta}{(1-\tau) \cdot (1-p)}} \right) = \frac{\tau \cdot r \cdot (\tau - \beta - 1)}{(\tau \cdot (2p - 1) + \beta - \tau + 1)^2}$$

1038 We can look at the fixed points of this expression, as they correspond to the value of p at which the
 1039 derivative changes the sign. This expression has fixed points at: $\tau = 0$, $r = 0$, and $\beta = \tau - 1$. Given our
 1040 parameters' ranges: $0 < \tau < 1$, $r \neq 0$ and $0 < \beta < 1$, none of those fixed points are present within those

1041 ranges. In addition, it can be seen that this $\left(\frac{\partial V^*}{\partial p}\right)$ is positive for the parameter values within those ranges.

1042 Therefore, knowing that the derivative won't reach any fixed point, V^* is always a growing monotonic
 1043 function with respect to p .

1044 For the condition 2 to be met, we can analyze the second derivative of V^* with respect to $\mathbf{E}[R(s)]$ as it
 1045 indicates the convexity of a function. The conditions to be met are:

$$1046 \quad - \quad V^* \text{ is convex if it is 'pessimistic': if } \tau < 0.5 \rightarrow \frac{\partial^2 V}{\partial p^2} > 0$$

$$1047 \quad - \quad V^* \text{ is concave if it is 'optimistic': if } \tau > 0.5 \rightarrow \frac{\partial^2 V}{\partial p^2} < 0$$

1048 In the case of Bernoulli return distributions, we take the second derivative of V^* with respect to
 1049 p (reward):

$$1050 \quad \frac{\partial^2 V^*}{\partial p^2} = \frac{\partial^2}{\partial p^2} \left(\frac{\frac{\tau}{1-\tau} \cdot \frac{p}{1-p} \cdot r}{\frac{\tau}{1-\tau} \cdot \frac{p}{1-p} + 1 + \frac{\beta}{(1-\tau) \cdot (1-p)}} \right) = \frac{2\tau \cdot (2\tau - 1) \cdot r \cdot (\tau - \beta - 1)}{(\tau \cdot (2p - 1) + \beta - p + 1)^3}$$

1051 We can again look at the fixed points of this expression. These happen at: $\tau = 0$, $r = 0$, and $\beta = \tau - 1$
 1052 and $\tau = 0.5$. Among them, the only fixed point within our parameters range is the latter. In addition, by
 1053 replacing τ in the expression above, it is easily shown that it is positive if $\tau < 0.5$ and negative if $\tau > 0.5$.
 1054 Thus, given that the ranges for the parameters are such that the second derivative won't reach any other
 1055 fixed point, condition 2 will always be met.

1056 **1.4. Distributional TD learning with D1 and D2 populations**

1057 The signatures of distributional RL were preserved in dopamine neurons firing rates after habenula
 1058 lesions (Extended Data Fig. 3-4). Therefore, we considered a third alternative to model 1 and 2, that
 1059 assigns different functions to each of the mechanisms for asymmetric learning rates.

1060 In this model (Extended Data Fig. 13) the single cell asymmetric scaling factors (α_i^+ , α_i^-) give rise to a
 1061 distributional expectile code for value and are implemented at the level of the scaling of RPE-evoked
 1062 responses of dopamine neurons:

1063
$$\delta_{i,t} := r_t + \gamma \cdot \bar{z}(s_{t+1}) - V_i(s_t)$$

1064
$$\hat{\delta}_{i,t} = \alpha_i^+ \cdot \delta_{i,t} \dots \text{if } \delta_{i,t} > 0$$

1065
$$\hat{\delta}_{i,t} = \alpha_i^- \cdot \delta_{i,t} \dots \text{if } \delta_{i,t} < 0$$

1066

1067 The modulation of receptor sensitivities, carried out downstream at the SPN level, gives rise to the global
 1068 rescaling of the value updates (η^+ , η^-) (Extended Data Fig. 13A):

1069
$$P_i(s_t) \leftarrow P_i(s_t) + \eta^+ \cdot |\hat{\delta}_i(t)| - \beta \cdot P_i(s_t) \dots \text{if } \delta_i(t) > 0$$

1070
$$N_i(s_t) \leftarrow N_i(s_t) + \eta^- \cdot |\hat{\delta}_i(t)| - \beta \cdot N_i(s_t) \dots \text{if } \delta_i(t) > 0$$

1071
$$\hat{V}_i(s_t) = P_i(s_t) - N_i(s_t)$$

1072 These set of update equations are equivalent to a modified version of the update equation of distributional
 1073 RL:

1074
$$[\Delta V_i(s_t)] = \frac{1}{N} \sum_j^N \eta^+ \cdot \alpha_i^+ \cdot \delta_{i,j} \cdot I_{\delta_{i,j}>0} + \eta^- \cdot \alpha_i^- \cdot \delta_{i,j} \cdot I_{\delta_{i,j}>0}$$

1075
$$V_i(s_t) \leftarrow V_i(s_t) + \mathbf{E}[\Delta V_i(s_t)]$$

1076 Thus, this model can give rise to biases in value learning (Extended Data Fig. 13), while keeping intact
 1077 information about the value distribution. By employing the results from the biophysical model (Fig. 6),
 1078 we found that this distributional TD model can parsimoniously explain all aspects of the data in the
 1079 habenula lesion study (Extended Data Fig. 12B), including the features of a distributional code and the
 1080 optimistic biases observed in behavior and dopamine cue-evoked responses (Extended Data Fig. 12).

1081 **1.5. Dependency of model on assumption: Log vs. linear scaling of receptor occupancy curves**

1082 Through this work, we have used the dose-occupancy curves of D1 and D2 receptors to derive the
1083 receptor sensitivities that result in the asymmetric scaling factors in Model 1. It is important to note that
1084 the slopes of the receptor occupancy curve (= receptor sensitivity) were obtained from the receptor
1085 occupancy curves plotted as a function of log of dopamine concentrations.

$$1086 \quad \alpha^+ = \frac{\Delta O_{ccD1}}{\Delta \log(C_{DA^+})}$$

$$1087 \quad \alpha^- = \frac{\Delta O_{ccD2}}{\Delta \log(C_{DA^-})}$$

1088

1089 To show that this assumption is not essential, we now derive the receptors sensitivities assuming linear
1090 changes in dopamine levels due to RPE-evoked responses.

$$1091 \quad \alpha^+ = \frac{\Delta O_{ccD1}}{\Delta C_{DA^+}}$$

$$1092 \quad \alpha^- = \frac{\Delta O_{ccD2}}{\Delta}$$

1093 As shown in Extended Data Fig. 9, the choice of a linear versus log scale affects the absolute magnitude
1094 of the derived receptor sensitivities, but the normalized metric $\tau = \frac{\alpha^+}{\alpha^- + \alpha^+}$ holds the same relationship to
1095 baseline dopamine levels with a small shift in the curve (Extended Data Fig.9, right panel). The
1096 normalized metric is the factor determining the update asymmetries and, thus, the stochastic fixed points
1097 at which the variables converge.

1098 **1.6. Normative motivation for two-factor learning rule**

1099 We have used in the previous models a so-called *two factor learning rule*, where the value updates
1100 depend only on the presynaptic activity (i.e., state input) and TD RPEs. Here, we motivate this choice
1101 from a normative approach based on previous work¹¹.

1102 Consider a linear approximation for value, where the value function (\hat{V}) is the output of a single linear
 1103 neuron. Here, \hat{V} is a linear function of the input feature-vector representing the state $\mathbf{x}(s) =$
 1104 $(x_1(s), \dots, x_n(s))$, parametrized with a weight vector $\mathbf{w} = (w_1, \dots, w_n) . . :$

$$1105 \quad \hat{V}(s, \mathbf{w}) = \mathbf{w}^T \mathbf{x}(s)$$

1106 To put it into neural terms, we can think of $x_i(s)$ as the presynaptic activity onto the value neuron \hat{V} ,
 1107 with a synaptic efficacy w_i .

1108 As before, the agent computes the TD error based on this linear approximation for value and the sampled
 1109 reward:

$$1110 \quad \delta_t = r_t + \gamma \cdot \hat{V}(s_{t+1}, \mathbf{w}) - \hat{V}(s_t, \mathbf{w})$$

1111 In the problem of *value prediction*, the agent aims to achieve the highest accuracy of prediction. One way
 1112 to achieve this is to perform *stochastic gradient descent* (SGD) with respect to the parameters (\mathbf{w}) of the
 1113 value function to minimize the objective function such as the squared error (δ_t^2). We can define this
 1114 optimization problem as: $\operatorname{argmin}_{\mathbf{w}} \left(\frac{1}{2} \delta^2 \right)$ where we have deliberately chosen the constant $\frac{1}{2}$ for clarity,
 1115 but it doesn't change the end results.

1116 To perform SGD in this minimization problem, the parameters (\mathbf{w}) should be updated in the opposite
 1117 direction of the gradient of the loss with respect to the parameters (i.e., opposite to $\nabla_{\mathbf{w}} (\frac{1}{2} \delta^2)$):

$$1118 \quad \mathbf{w} \leftarrow \mathbf{w} - \alpha \cdot \nabla_{\mathbf{w}} \left(\frac{1}{2} \delta^2 \right)$$

1119 Where α is the learning rate. To compute the gradient, we use the chain rule:

$$1120 \quad \nabla_{\mathbf{w}} \left(\frac{1}{2} \delta^2 \right) = \frac{\partial (1/2 \delta^2)}{\partial \hat{V}} \cdot \frac{\partial \hat{V}}{\partial \mathbf{w}} = \frac{\partial (1/2 (r - \hat{V})^2)}{\partial \hat{V}} \cdot \frac{\partial \mathbf{w}^T \mathbf{x}(s_t)}{\partial \mathbf{w}} = \frac{-2(r - V)}{2} \cdot \mathbf{x}(s_t) = -\delta \cdot \mathbf{x}(s_t)$$

1121 Therefore, the update for the parameters of the value function is:

$$1122 \quad \mathbf{w} \leftarrow \mathbf{w} + \alpha \cdot \delta \cdot \mathbf{x}(s_t)$$

1123 The term $\delta \cdot \mathbf{x}(s_t)$ in the equation above is what we call a *two-factor learning rule*, dependent only on the
 1124 presynaptic activity and not contingent on the post-synaptic activity.

1125 The development of the TD learning model with D1 and D2 populations (section 1.3) has respected this
 1126 learning rule, complying with what is required for SGD in the value prediction problem. Note that we

1127 have implicitly developed our models with a complete serial compound representation (CSC) of the
1128 states¹¹, where $x(s_t) = 1$ in a single element $x_i(s_t)$ representing the current state and 0 otherwise. It can
1129 be shown that with this representation, the update equation above is equivalent to:

$$1130 \quad \hat{V} \leftarrow \hat{V} + \alpha \cdot \delta$$

1131

1132 **2. Computational model of dopamine release and receptor occupancy**

1133 To predict changes in dopamine concentrations and receptor occupancies (Fig. 6), we employed a
1134 biophysical model developed elsewhere⁵⁹. It presents two interacting dynamical systems. The first
1135 system models the change in receptor occupancies while the second the change in dopamine levels per
1136 unit time.

1137 In the first system, the occupancy of receptors is modelled as a binding reaction between dopamine (*DA*)
1138 and D1 or D2 receptors (*R*), using the constants for forward and backward reactions (k_{on}, k_{off}).



1140 This formulation results in the following equation for the change in receptor occupancy $Occ(t)$ per unit
1141 time:

$$1142 \quad \frac{dOcc(t)}{dt} = (1 - Occ(t)) \times k_{on} \times C_{DA}(t) - Occ(t) \times k_{off}$$

1143 The values used for the association and dissociation constants for each receptor type (k_{on} and k_{off} ,
1144 respectively) are detailed in Table 1.

1145 In the second system, the change in dopamine concentration ($C_{DA}(t)$) is a function of both dopamine
1146 release and uptake.

$$1147 \quad \frac{dC_{DA}(t)}{dt} = DA_{\text{release}}(t) - DA_{\text{uptake}}(t)$$

1148 Dopamine release is a product of firing rate ($v(t)$) and release capacity ($\gamma(t)$)

$$1150 \quad DA_{\text{release}}(t) = \gamma(t) \cdot v(t)$$

1149 Where:

- 1151 1. $v(t)$ is the firing rate of dopamine neurons, provided by the neural data.
- 1152 2. $\gamma(t) = \gamma_{pr_n} \cdot P_r \cdot G_{D2}(t)$ is defined as the increase in $C_{DA}(t)$ by a single synchronized action
 1153 potential:
- 1154 a. $P_r = 1$ release probability in the absence of presynaptic D2-autorreceptors,
- 1155 b. $\gamma_{pr_n} = 2$ release capacity in the absence of presynaptic D2-autorreceptors. This value was set
 1156 to be deliberately high and anticipates a ~50% reduction by terminal feedback.
- 1157 c. $G_{D2}(t)$ is a multiplicative gain that represents the modulation of dopamine release by D2-
 1158 autorreceptors. This is a decaying function of the occupancy of D2-autorreceptors
 1159 ($Occ_{D2_a}(t)$) which is modelled by the same binding reaction explained above. The gain is
 1160 parametrized by the autoreceptor efficacy, $\alpha = 3$. The smaller the α the less the decay in
 1161 release with receptor occupancy.

$$G_{D2}(t) = \frac{1}{1 + \alpha \cdot Occ_{D2_a}(t)}$$

1163 Dopamine uptake is a function of the uptake of dopamine by the dopamine transporter (DAT) and other
 1164 non-DAT sources

$$1165 \quad DA_{\text{uptake}}(t) = dt \cdot \left(\frac{V_{max}^{pr_n} \cdot C_{DA}(t)}{K_m + C_{DA}(t)} - K_{nonDAT} \right)$$

1166 Where:

- 1167 - $V_{max}^{pr_n} = 1500 \frac{nM}{sec}$ the maximal uptake capacity assuming approximately 100 terminals in the
 1168 near surroundings.
- 1169 - $K_m = 160 nM$, is the Michaelis-Menten parameter for uptake mediated by DAT
- 1170 - $K_{nonDAT} = 0.04 nM$ is a constant for the dopamine removal not mediated by DAT. For
 1171 example, monoamine oxidase (MAO) and noepinephrine transporter (NET) mediated uptake.

1172 The variables of the model reported in Fig. 6 correspond to: $Occ_{D1}(t)$, $Occ_{D2}(t)$, $C_{DA}(t)$. We used as
 1173 input to the model the firing rates derived from the electrophysiological recording of optogenetically
 1174 identified dopamine neurons conducted in Tian and Uchida (2015)⁵⁵. This modeling, while considering
 1175 major processes, does not take into account all of the complexity of the biological environment in the

1176 brain, yet we used this model to obtain an approximate estimate of the order of changes in dopamine
1177 concentrations and receptor occupancies.

1178 3. Simulation details of habenula lesion data

1179 3.1. Biophysical model simulations

1180 We used the computational model described previously (methods section 2)⁵⁹ to calculate the
1181 extracellular dopamine levels and estimate the occupancy of postsynaptic receptors from the habenula
1182 lesion dataset. The model was driven by the average spike rate of dopamine neurons recorded from
1183 control or lesioned animals. For each recorded dopamine neuron, the simulations were carried on a trial
1184 by trial basis that consisted of a time window [-15, 20] sec with respect to cue onset. A relatively large
1185 window was used to allow for the relevant variables to stabilize in its baseline, as the simulations were
1186 initialized at zero.

1187 For each trial, spikes were first binned with 10-ms windows and then smoothed by a Gaussian kernel
1188 ($\sigma = 0.3 \times (ISI_{mean})$). All trials were then averaged across trials, to determine the mean single-cell
1189 response for dopamine release and D1 and D2 receptor activation. Final average dopamine concentrations
1190 and receptor occupancies were obtained from the average of all mean single-cell responses.

1191 *Computation of receptors sensitivities from the model results*

1192 We computed the receptor sensitivity from the occupancies Occ_{D1} , Occ_{D2} and their theoretical dose-
1193 occupancy curves. Starting from the occupancy at baseline, we derived the change in occupancy as a
1194 function of the transients in dopamine concentration C_{DA} elicited by RPE-evoked dopamine responses, at
1195 the level of the population average.

1196 The ratio between these quantities corresponds to the receptors' sensitivities. These are transferred as α^+
1197 and α^- to our reinforcement learning model (model 1):

$$1198 \quad \alpha^+ = \frac{\Delta Occ_{D1}}{\Delta C_{DA}} \dots \text{ if } \Delta C_{DA} > 0$$

$$1199 \quad \alpha^- = \frac{\Delta Occ_{D2}}{\Delta C_{DA}} \dots \text{ if } \Delta C_{DA} < 0$$

1200 Where ΔC_{DA} , ΔOcc_{D1} , ΔOcc_{D2} are the changes computed with respect to baseline, as: $\Delta x =$
1201 $\bar{x}_{outcome} - \bar{x}_b$, for each variable $x = \{C_{DA}, Occ_{D1}, Occ_{D2}\}$. Where \bar{x} denotes the population average

1202 response for each group. The outcome responses were taken as the average from [0,1] sec after outcome
1203 onset, while the baseline was taken as the average from [-1, 0] sec with respect to cue onset.

1204 3.2. Model 1 simulations

1205 The simulations for Model 1 were carried out with a TD learning model with D1 and D2 populations
1206 (methods section 1.3). We ran the simulations using the resultant receptor sensitivities from the
1207 biophysical model as the population-level asymmetric learning rates in Model 1 (i.e., the learning rates
1208 α^+ , α^- for P and N updates). The simulations were run for 3,000 trials on the Pavlovian conditioning task
1209 used in the study⁵⁵. We assumed a uniform distribution of trial types across the session. Each trial
1210 consisted of 4 states (baseline, cue, delay, reward), assuming Markovian dynamics between them. All
1211 variables were initialized at zero. The model had as hyper-parameters a discounting factor of $\gamma = 0.99$
1212 and a decay term $\beta = 0.002$. We report in Fig. 4, Model 1 results assuming a uniform scaling of TD
1213 RPEs across the neuronal population. In Extended Data Fig. 12 we show that this model reproduces key
1214 signatures of the data irrespective of the choice of the decay factor β .

1215 The results are not dependent on a uniform scaling of TD RPEs. Given that distributional RL signatures
1216 were preserved in the data even after habenula lesions, we also considered Model 1 under the
1217 distributional TD learning framework (Extended Data Fig. 13). For this, we used the distribution of single
1218 cell asymmetric scaling factors (α_i^+ , α_i^-) derived from the dopamine neurons firing rates. This model also
1219 reproduced key signatures of the data irrespective of the choice of the decay factor β (Extended Data Fig.
1220 12).

1221 3.3. Model 2 simulations

1222 The simulations for Model 2 were carried out with a TD learning model. As with Model 1, simulations
1223 were run for 3,000 trials on the Pavlovian conditioning task⁵⁵. We assumed a uniform distribution of trial
1224 types across the session. Each trial consisted of 4 states (baseline, cue, delay, reward), assuming
1225 Markovian dynamics between them. All variables were initialized at zero. The model had as parameters a
1226 discounting factor of $\gamma = 0.99$.

1227 We used the distribution of single cell asymmetric scaling factors derived from the firing rates of
1228 dopamine neurons as α_i^+ , α_i^- . In section 1.2 we emphasized that in order to accurately compute the TD
1229 RPE in distributional TD, we require taking samples from the estimated return distribution
1230 $\tilde{z}_i(s_{t+1}) \sim Z(s_{t+1})$. We did this by running an optimization process where we minimize for the expectile

1231 loss between the taken samples $\tilde{z}_i(s_{t+1})$, $V_i(s_{t+1})$ from the model, and τ_i as estimated from the data.

1232 The problem was defined as $\text{argmin}_{s_1 \dots s_m} \mathcal{L}(s, V, \tau)$ where:

1233
$$\mathcal{L}(s, V, \tau) = \frac{1}{M} \sum_{m=1}^M \sum_{i=1}^N |\tau_i - \mathbf{I}_{s_m < V_n}| (\hat{z}_m - V_i)^2$$
, for N neurons and M samples

1234 In the simulations, we took M samples where M equals the number of neurons (N) and performed an
 1235 update taking the expectation across all samples as described in the methods section 1.3.

1236 4. Simulations details for replications of previous experimental results

1237 4.1. Cools et al. (2009)

1238 We simulate the results from Cools et al. (2009) (Fig. 7, Extended Data Fig. 6-7) in which they tested the
 1239 effects of bromocriptine in altering learning rate asymmetry²⁴. In their study, they performed a reversal
 1240 learning task and reported a parameter called ‘relative reversal learning (RRL)’, equivalent to the
 1241 difference between the positive and negative learning rates in our model. We computed this as: $\alpha^+ \alpha^+ +$
 1242 $+ \alpha^- - \alpha^- - \alpha^+ + \alpha^- = \tau - 1 - \tau = 2\tau - 1$, reported in Fig. 7 E,F, where the parameters α^+ , α^- were
 1243 computed from the slopes of the D2l (postsynaptic D2 receptors) and D1 occupancy curves $(2\tau - 1)_{occ}$
 1244 or activation curves $(2\tau - 1)_{act}$. The change in relative reversal learning in Fig. 7 H-I was calculated as
 1245 taking the difference between drug and the ‘control’ condition as:

1246
$$\Delta(2\tau - 1) = (2\tau - 1)_{drug} - (2\tau - 1)_{control}.$$

1247 We simulated the effect of bromocriptine using the biophysical model for dopamine release and receptor
 1248 occupancy (Section 2, Methods). We added an additional ligand for D2 receptors to the update equations
 1249 for occupancy:

1250
$$\frac{dOcc_{DA,r_j}(t)}{dt} = \left(1 - Occ_{DA,r_j}(t)\right) \times k_{on}^{DA,r_j} \times C_{DA}(t) - k_{off}^{DA,r_j}$$

1251
$$\frac{dOcc_{Drug,r_j}(t)}{dt} = \left(1 - Occ_{Drug,r_j}(t)\right) \times k_{on}^{Drug,r_j} \times C_{Drug}(t) - k_{off}^{Drug,r_j}$$

1252 Where $r_j: \{D1, D2s, D2l\}$, and $k_{on}^{Drug,D2s} = 0.02083$, $k_{off}^{Drug,D2s} = 0.1$, $k_{on}^{Drug,D2l} = 0.04$, $k_{off}^{Drug,D2l} = 0.1$
 1253 are reported in Table 1⁹⁷.

1254 To calculate the effects of efficiency of the drug, we calculated the activation of D2l and D2s receptors in
 1255 the following way:

1256
$$Act_{r_j}(t) = E_{DA,r_j} \cdot Occ_{DA,r_j}(t) + E_{Drug,r_j} \cdot Occ_{Drug,r_j}(t)$$

1257 Where $E_{DA,r_j} = 1$ is the efficiency of dopamine on the receptors activation, and $E_{Drug,r_j} < 1$ the
 1258 efficiency of the drug, for $r_j: \{D1, D2s, D2l\}$. The parameter for D1 receptors was kept at $E_{Drug,D1} = 0$
 1259 for all simulations.

1260 To simulate the effects of D2s activation by the drug in D2l occupancy in Fig. 7b,e,h we report the effects
 1261 of $E_{Drug,D2s} = 0$ (solid lines) and $E_{Drug,D2s} = 0.6$ (dashed lines). To simulate the effect of the drug in
 1262 D2s and D2l activation in Fig. 7c,f,i we report the effects of $E_{Drug,D2s} = 0.6, E_{Drug,D2l} = 0.6$.

1263 We show how the qualitative nature of the effects of the drug in relative reversal learning still hold
 1264 regardless of whether the parameter τ is computed from the occupancy curves (Extended Data Fig. 7, Fig.
 1265 7n,e,h) or the activation curves (Extended Data Fig. 8, Fig. 7c,f,i). In addition, in Supplementary Figure
 1266 8-9 we show that the qualitative results still hold regardless of the choice of the efficiency parameters
 1267 $E_{Drug,D2s}$ and $E_{Drug,D2l}$.

1268 **4.2. Timmer et al. (2018)**

1269 In this study²⁵ they reported a ‘loss aversion’ parameter (λ in their results).

1270
$$SUG = (1 - \lambda) \cdot p_{gain} \cdot Gain + \lambda \cdot p_{loss} \cdot Loss$$

1271 Where SUG is the ‘subjective utility’ for a given option, and $p_{gain} = p_{loss}$.

1272 In our formulation, we assume that the task in the study is performed under steady state conditions after
 1273 having learned with a learning rate (τ). With this assumption, the SUG at task performance is equivalent to
 1274 the convergent V estimate after learning. We will show that at these steady state conditions $(1 - \tau)$ is
 1275 equivalent to (λ) .

1276 Starting with the solution for V :

1277
$$SUG = V = \frac{\tau \cdot p_{gain} \cdot r_{gain} + (1 - \tau) \cdot (1 - p_{gain}) \cdot r_{loss}}{\tau \cdot p_{gain} + (1 - \tau) \cdot (1 - p_{gain})}$$

1278 Replacing for $p_{gain} = 0.5$:

1279
$$SUG = \frac{\tau \cdot r_{gain} + (1 - \tau) \cdot r_{loss}}{\tau + (1 - \tau)}$$

1280 Given that: $\tau + (1 - \tau) = 1$

1281
$$\text{SUG} = \tau \cdot r_{\text{gain}} + (1 - \tau) \cdot r_{\text{loss}}$$

1282 Therefore, our model, applied to their task, gives rise to the same SUG computation, with λ equivalent to
1283 $(1 - \tau)$.

1284 To generate Fig. 8F, we performed the following steps:

- 1285 1. We first estimated the theoretical change in baseline DA elicited by the medication. For this, we
1286 computed the equivalent τ for the λ they report in the OFF and ON medication conditions
1287 ($\lambda_{\text{OFF}} = 1.51$, $\lambda_{\text{ON}} = 1.19$), using the relationship: $(1 - \tau) = \lambda$. We then computed the baseline
1288 DA levels that would give rise to the τ_{ON} and τ_{OFF} . With this, we computed the change in
1289 baseline DA (ΔDA) equivalent to the change $\Delta\tau = \tau_{\text{ON}} - \tau_{\text{OFF}}$. This ΔDA is the theoretical
1290 change in baseline DA elicited by the medication (Fig. 8F).
- 1291 2. To generate Fig. 8F, we sampled a set of λ from a Gaussian distribution centered at a mean of
1292 $\mu_{\lambda} = 1.51$ and a standard deviation of $\sigma_{\lambda}^2 = 3$, to emulate the distribution of λ_{OFF} they report in
1293 the OFF condition. We then computed the equivalent τ for that set of λ with the relationship
1294 above. We will call this the distribution of τ'_{OFF} .
- 1295 3. We used the derived τ'_{OFF} distribution to compute the equivalent dopamine levels. We imposed a
1296 change in baseline DA equal to the ΔDA computed in the first step and computed the new set of τ
1297 for that set of new baseline DA levels (τ'_{ON}). The ‘drug effect in loss aversion’ reported in Fig.
1298 8F is the $\tau'_{\text{ON}} - \tau'_{\text{OFF}}$ for each sample.

1299 **5. Details on habenula lesion data**

1300 **5.1 Animals, surgery and lesions.**

1301 The rodent data we re-analyzed here were first reported in Tian and Uchida (2015)⁵⁵. Below we provide a
1302 brief description of the methods. Further methodological details can be found in the original paper. !2
1303 mice were used. Bilateral habenula lesions were performed in five animals. Seven animals were in the
1304 control group including two with sham-lesion operation, one with only small contra-lateral side lesion of
1305 the medial habenula, and four animals without operations in the habenula. During surgery, a head plate
1306 was implanted on the skull, and adeno-associated virus (AAV) that express channelrhodopsin-2 (ChR2)
1307 in a Cre-dependent manner was injected into the VTA (from bregma: 3.1 mm posterior, 0.7 mm lateral,
1308 4–4.2 mm ventral). After recovery from surgery, mice were trained on the conditioning task, after which
1309 mice were randomly selected to be in lesion or sham-lesion group. Electrolytic lesions were made
1310 bilaterally using a stainless-steel electrode (15 kU, MicroProbes, MS301G) with a cathodal current of 150

1311 mA. Each side of the brain was lesioned at two locations (from bregma: 1.6 mm/1.9 mm posterior, 1.15
1312 mm lateral, 2.93 mm depth, with a 14 angle). For sham-lesion operations, no current was applied. In the
1313 same surgery, a microdrive containing electrodes and an optical fiber was implanted in the VTA (from
1314 bregma: 3.1 mm posterior, 0.7 mm lateral, 3.8–4.0 mm ventral)⁹⁸.

1315 **5.2 Behavioral task**

1316 Twelve mice were trained on a probabilistic Pavlovian task. Each trial the animal experienced one of four
1317 odor cues for 1 s, followed by a 1-s pause, followed by a reward (3.75 μ l water), an aversive air puff or
1318 nothing. Odor 1 to 3 signaled a 90%, 50% and 10% probability of reward, respectively. Odor 4 signaled a
1319 90% probability of air puff. Odor identities were randomized across trials and included: isoamyl acetate,
1320 eugenol, 1-hexanol, p-cymene, ethyl butyrate, 1-butanol, and carvone (1/10 dilution in paraffin oil). Inter-
1321 trial intervals were exponentially distributed. An infrared beam was positioned in front of the water
1322 delivery spout and each beam break was recorded as one lick event. We report the average lick rate over
1323 the interval 500–2,000 ms after cue onset.

1324 **5.3 Electrophysiology**

1325 Recordings were made using a custom-built microdrive equipped with 200- μ m-fiber optic-coupled with
1326 eight tetrodes. DA neurons were identified optogenetically⁹⁸. A stimulus-associated spike latency test
1327 (SALT) algorithm⁹⁹ was used to determine whether light pulses significantly changed a neuron's spike
1328 timing.

1329 **5.4 Neural data analysis**

1330 Data analyses were performed using MATLAB R2021b (Mathworks). To measure firing rates,
1331 peristimulus time histograms (PSTHs) were constructed using 1-ms bins. These histograms were then
1332 smoothed by convolving with the function $f(t) = (1 - e^{-t}) \cdot e^{-\frac{t}{\tau}}$ where τ was a time constant set to 20
1333 ms as in¹⁸. 44 dopamine neurons were recorded from lesioned animals (5 animals, 30 sessions), and 45
1334 dopamine neurons were recorded from control animals (7 animals, 35 sessions). We pooled all the cells
1335 across animals in each group for analysis. Cue-evoked responses were defined as the average activity
1336 from 0 to 400 ms after cue onset. Outcome-evoked responses were defined as the average activity from
1337 2000 to 2600 ms after cue onset.

1338 The normalization of cue response shown in Fig. 4 was carried out following a previous work³⁶ on a per-

1339 cell basis as: $c_{50}^{norm} = \frac{c_{50} - \bar{c}_{10}}{\bar{c}_{90} - \bar{c}_{10}}$, where \bar{c}_{90} , \bar{c}_{10} correspond to the mean across trials within a cell for the

1340 90% and 10% probability cure responses. To derive the t-statistics in Fig. 4d, we performed a two-tailed t-
1341 test of the cell's normalized responses to the 50% cue against the average midway point between
1342 responses to the 10% cue and responses to the 90% cue.

1343 The derivation of asymmetric scaling factors from outcome responses (τ_i), was carried out following³⁶,
1344 with some modifications to adapt it to the task. The procedure is illustrated in Extended Data Fig. 3.

1345 • To compute the reversal points, outcome responses were first aligned to the RPE for each trial
1346 type, computed with the true expected value of each reward distribution. Assuming a fixed reward
1347 value of 1 (arbitrary units), the expected value for the 90%, 50%, 10% reward probability trials
1348 corresponded to 0.1, 0.5, 0.9, respectively. Given this, omission responses from the 90%, 50%,
1349 10% reward probability trials correspond to RPEs of -0.9, -0.5 and -0.1. The rewarded responses
1350 from the 90%, 50%, 10% reward probability trials correspond to RPEs of 0.1, 0.5 and 0.9. The
1351 reward value is arbitrary and doesn't have an effect in this computation as it only shifts the RPE
1352 axis by a fixed amount. The reversal point for each cell (Z_i) was defined as the RPE that
1353 maximized the number of positive responses to RPEs greater than Z_i plus the number of negative
1354 responses to RPEs less than Z_i . The distribution of reversal points is reported in Extended Data
1355 Fig. 4. To obtain statistics for reliability of the computed reversal points, we partitioned the data
1356 into random halves and estimated the reversal point for each cell separately in each half. We
1357 repeated this procedure 1000 times with different random partitions, and we report the
1358 distribution of Pearson's correlation across these 1000 folds (Extended Data Fig. 4).

1359 • After measuring reversal points, we fit linear functions separately to the positive and negative
1360 domains. Given that dopamine's responses are non-linear in the reward space but present a
1361 putative utility function¹⁰⁰, we approximated the underlying utility function from the dopamine
1362 responses to RPEs of varying magnitudes. We used these empirical utilities instead of raw RPEs
1363 for computing the slopes that correspond to α_i^+ , α_i^- . We then computed the asymmetric scaling
1364 factors as $\tau_i = \frac{\alpha_i^+}{\alpha_i^+ + \alpha_i^-}$. We performed the same cross-validation procedure used for the reversal
1365 points. The distribution of R value across the 1000 folds are reported in Extended Data Fig. 4.

1366 A key prediction of distributional RL³⁶ is the presence of a correlation (across cells) between reversal
1367 points Z_i and asymmetric scaling factors τ_i . To elucidate whether signatures of distributional RL were
1368 still present after lesions, we followed the procedure given by Dabney et al. (2020)³⁶ to compute this
1369 correlation. We first randomly split the data into two disjoint halves of trials. In one half, we first

1370 calculated reversal points Z_i^1 and used them to calculate α_i^+ , α_i^- . In the other half, we again calculated the
1371 reversal points Z_i^2 . The correlation we report in Extended Data Fig. 4 is between Z_i^2 and $\tau_i = \frac{\alpha_i^+}{\alpha_i^+ + \alpha_i^-}$.

1372 **5.5 Model fitting to the anticipatory licking responses**

1373 For each trial we computed the average lick rate over the interval 500–2,000 ms after cue onset. For each
1374 model, we fit the free parameters to the lick rates using maximum likelihood estimation. The optimization
1375 was performed using the SciPy optimization toolbox (Python) that minimized the difference between the
1376 predicted lick rates and the ground truth ones, with a uniform prior distribution over the parameters. The
1377 fits were done considering three RL models that had between 2 and 3 parameters. The models, parameters
1378 and bounds used for each of them are detailed in table 2.

1379

1380

1381 **Tables**

1382 **Table 1 – Biophysical model parameters**

1383

Parameter	Abbreviation	Value
DA association constant to D2 autorreceptors	k_{on}^{D2term}	$0.3 M^{-1}s^{-1}$
DA dissociation constant to D2 autorreceptors	k_{off}^{D2term}	$0.003 s^{-1}$
DA association constant to D1 receptors	k_{on}^{D1}	$0.01 M^{-1}s^{-1}$
DA dissociation constant to D1 receptors	k_{off}^{D1}	$10 s^{-1}$
DA association constant to D2 receptors	k_{on}^{D2}	$0.2 M^{-1}s^{-1}$
DA dissociation constant to D2 receptors	k_{off}^{D2}	$2 s^{-1}$
Release probability from terminals	P_r	$1 a. u.$
Release capacity from terminals	γ_{pr_n}	$2 a. u.$
D2 autorreceptor efficacy	α	$3 a. u.$
DAT maximal uptake capacity	$V_{max}^{pr_n}$	$1500 nMs^{-1}$
Michaelis-Menten parameter DAT-mediated DA uptake	K_m	$160 nM$
Constant for dopamine removal not mediated by DAT's	K_{nonDAT}	$0.04 nM$
Bromocriptine association constant to D2 autorreceptors	$k_{on}^{Drug,D2s}$	$0.02083 M^{-1}s^{-1}$
Bromocriptine dissociation constant to D2 autorreceptors	$k_{off}^{Drug,D2s}$	$0.1 s^{-1}$
Bromocriptine association constant to D2 receptors	$k_{on}^{Drug,D2l}$	$0.04 M^{-1}s^{-1}$
Bromocriptine dissociation constant to D2 receptors	$k_{off}^{Drug,D2l}$	$0.1 s^{-1}$

1384 a.u. = arbitrary units

1385 M = mols

1386 s = seconds

1387

1388

1389 Table 2- Reinforcement learning models fit to the behavioral data from Tian & Uchida

Model	Formulation	Parameters	Parameter bounds
TD learning	$\delta = r - V$ $V \leftarrow V + \alpha \cdot \delta$ $Licking = \beta \cdot V$	α, β	$\alpha \in [.001, 1]$ $\beta \in [.1, 10]$
TD learning with reward sensitivity	$\delta = \rho \cdot r - V$ $V \leftarrow V + \alpha \cdot \delta$ $Licking = \beta \cdot V$	α, ρ, β	$\alpha \in [.001, 1]$ $\rho \in [.001, 10]$ $\beta \in [.1, 10]$
Risk sensitive TD learning	$\delta = r - V$ $V \leftarrow V + \alpha^+ \cdot \delta \quad \text{if } \delta > 0$ $V \leftarrow V + \alpha^- \cdot \delta \quad \text{if } \delta < 0$ $Licking = \beta \cdot V$	$\alpha^+, \alpha^-, \beta$	$\alpha^+ \in [.001, 1]$ $\alpha^- \in [.001, 1]$ $\beta \in [.1, 10]$

1390

1391

1392 Data availability

1393 The neural data and simulation results reported in this article have been shared in a public
 1394 deposit source in: https://osf.io/cr5mv/?view_only=bd13a2d2de1947699b56ce70610b0e9b

1395

1396 Code availability

1397 The accession codes for the data as well as the code for analysis and simulations are available at:
 1398 https://github.com/sandraromerop/D1D2_Dopamine

1399

1400

1401 **References**

- 1402 1. Brown, V. M., Zhu L., Solway A., Wang M., McCurry K., King-Casas B. & Chiu P.
1403 Reinforcement Learning Disruptions in Individuals With Depression and Sensitivity to
1404 Symptom Change Following Cognitive Behavioral Therapy. *JAMA Psychiatry* **78**, 1113–
1405 1122 (2021).
- 1406 2. Groman, S. M., Thompson, S. L., Lee, D. & Taylor, J. R. Reinforcement learning detuned in
1407 addiction: integrative and translational approaches. *Trends Neurosci.* **45**, 96–105 (2022).
- 1408 3. Ligneul, R., Sescousse, G., Barbalat, G., Domenech, P. & Dreher, J.-C. Shifted risk
1409 preferences in pathological gambling. *Psychol. Med.* **43**, 1059–1068 (2013).
- 1410 4. Mason, L., O’Sullivan, N., Bentall, R. P. & El-Dereby, W. Better than I thought: positive
1411 evaluation bias in hypomania. *PLoS One* **7**, e47754 (2012).
- 1412 5. Pizzagalli, D. A., Iosifescu, D., Hallett, L. A., Ratner, K. G. & Fava, M. Reduced hedonic
1413 capacity in major depressive disorder: evidence from a probabilistic reward task. *J.*
1414 *Psychiatr. Res.* **43**, 76–87 (2008).
- 1415 6. Verdejo-Garcia, A., Chong, T. T.-J., Stout, J. C., Yücel, M. & London, E. D. Stages of
1416 dysfunctional decision-making in addiction. *Pharmacol. Biochem. Behav.* **164**, 99–105
1417 (2018).
- 1418 7. Lim, T. V., Cardinal, R. N., Bullmore, E. T., Robbins, T. W. & Ersche, K. D. Impaired
1419 Learning From Negative Feedback in Stimulant Use Disorder: Dopaminergic Modulation.
1420 *Int. J. Neuropsychopharmacol.* **24**, 867–878 (2021).
- 1421 8. Schönfelder, S., Langer, J., Schneider, E. E. & Wessa, M. Mania risk is characterized by an
1422 aberrant optimistic update bias for positive life events. *J. Affect. Disord.* **218**, 313–321
1423 (2017).
- 1424 9. Dayan, P. & Daw, N. D. Decision theory, reinforcement learning, and the brain. *Cogn. Affect.*
1425 *Behav. Neurosci.* **8**, 429–453 (2008).
- 1426 10. Katahira, K. The relation between reinforcement learning parameters and the influence of
1427 reinforcement history on choice behavior. *J. Math. Psychol.* **66**, 59–69 (2015).

- 1428 11. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*. (Bradford Books,
1429 2018).
- 1430 12. Maia, T. V. & Frank, M. J. From reinforcement learning models to psychiatric and
1431 neurological disorders. *Nat. Neurosci.* **14**, 154–162 (2011).
- 1432 13. Eldar, E., Rutledge, R. B., Dolan, R. J. & Niv, Y. Mood as Representation of Momentum.
1433 *Trends Cogn. Sci.* **20**, 15–24 (2016).
- 1434 14. Rutledge, R. B., Skandali, N., Dayan, P. & Dolan, R. J. A computational and neural model of
1435 momentary subjective well-being. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 12252–12257 (2014).
- 1436 15. Floresco, S. B., West, A. R., Ash, B., Moore, H. & Grace, A. A. Afferent modulation of
1437 dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. *Nat.*
1438 *Neurosci.* **6**, 968–973 (2003).
- 1439 16. Wang, Y., Toyoshima, O., Kunimatsu, J., Yamada, H. & Matsumoto, M. Tonic firing mode
1440 of midbrain dopamine neurons continuously tracks reward values changing moment-by-
1441 moment. *Elife* **10**, (2021).
- 1442 17. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward.
1443 *Science* **275**, 1593–1599 (1997).
- 1444 18. Eshel, N., Tian, J., Bukwich, M. & Uchida, N. Dopamine neurons share common response
1445 function for reward prediction error. *Nat. Neurosci.* **19**, 479–486 (2016).
- 1446 19. Steinberg, E. Keiflin, R., Boivin, J., Witten I., Deisseroth K. & Janak P. A causal link
1447 between prediction errors , dopamine neurons and learning. *Nat. Neurosci.* **16**, 966–973
1448 (2013).
- 1449 20. Waelti, P., Dickinson, A. & Schultz, W. Dopamine responses comply with basic assumptions
1450 of formal learning theory. *Nature* **412**, 43–48 (2001).
- 1451 21. Korn, C. W., Sharot, T., Walter, H., Heekeren, H. R. & Dolan, R. J. Depression is related to
1452 an absence of optimistically biased belief updating about future life events. *Psychol. Med.* **44**,
1453 579–592 (2014).
- 1454 22. Rutledge, R. B., Lazzaro, S. C., Lau, B., Myers, C. E., Gluck, M. A., & Glimcher, P. W.
1455 Dopaminergic drugs modulate learning rates and perseveration in Parkinson’s patients in a
1456 dynamic foraging task. *J. Neurosci.* **29**, 15104–15114 (2009).

- 1457 23. Frank, M. J., Seeberger, L. C. & O'Reilly, R. C. By carrot or by stick: Cognitive
1458 reinforcement learning in Parkinsonism. *Science* **306**, 1940–1943 (2004).
- 1459 24. Cools, R., Frank M.J., Gibbs S., Miyakawa A., Jagust W. & D'Esposito M. Striatal
1460 dopamine predicts outcome-specific reversal learning and its sensitivity to dopaminergic
1461 drug administration. *J. Neurosci.* **29**, 1538–1543 (2009).
- 1462 25. Timmer, M. H. M., Sescousse, G., van der Schaaf, M. E., Esselink, R. A. J. & Cools, R.
1463 Reward learning deficits in Parkinson's disease depend on depression. *Psychol. Med.* **47**,
1464 2302–2311 (2017).
- 1465 26. Gradin, V. B., Kumar, P., Waiter, G., Ahearn, T., Stickle, C., Milders, M., Reid, I., Hall, J.,
1466 & Steele, J. D. Expected value and prediction error abnormalities in depression and
1467 schizophrenia. *Brain* **134**, 1751–1764 (2011).
- 1468 27. Kumar, P., Waiter, G., Ahearn, T., Milders, M., Reid, I., & Steele, J. D. Abnormal temporal
1469 difference reward-learning signals in major depression. *Brain* **131**, 2084–2093 (2008).
- 1470 28. Pizzagalli, D. A., Holmes, A. J., Dillon, D. G., Goetz, E. L., Birk, J. L., Bogdan, R.,
1471 Dougherty, D. D., Iosifescu, D. V., Rauch, S. L., & Fava, M. Reduced caudate and nucleus
1472 accumbens response to rewards in unmedicated individuals with major depressive disorder.
1473 *Am. J. Psychiatry* **166**, 702–710 (2009).
- 1474 29. Robinson, O. J., Cools, R., Carlisi, C. O., Sahakian, B. J. & Drevets, W. C. Ventral striatum
1475 response during reward and punishment reversal learning in unmedicated major depressive
1476 disorder. *Am. J. Psychiatry* **169**, 152–159 (2012).
- 1477 30. Collins, A. G. E. & Frank, M. J. Opponent actor learning (OpAL): modeling interactive
1478 effects of striatal dopamine on reinforcement learning and choice incentive. *Psychol. Rev.*
1479 **121**, 337–366 (2014).
- 1480 31. Frank, M. J. Dynamic Dopamine Modulation in the Basal Ganglia: A Neurocomputational
1481 Account of Cognitive Decits in Medicated and Non-medicated Parkinsonism. *J. Cogn.*
1482 *Neuroci.* **17**, 51-72 (2005).
- 1483 32. Mikhael, J. G. & Bogacz, R. Learning Reward Uncertainty in the Basal Ganglia. *PLoS*
1484 *Comput. Biol.* **12**, 1–28 (2016).

- 1485 33. Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G.C.R., Urakubo, H., Ishii, S. & Kasai, H.
1486 A critical time window for dopamine actions on the structural plasticity of dendritic spines.
1487 *Science* **345**, 1616–1620 (2014).
- 1488 34. Iino, Y. Sawada, T., Yamaguchi, K. *et al.* Dopamine D2 receptors in discrimination learning
1489 and spine enlargement. *Nature* **579**, 555–560 (2020).
- 1490 35. Lee, S. J., Lodder, B., Chen, Y., Patriarchi, T., Tian, L & Sabatini, B. Cell-type-specific
1491 asynchronous modulation of PKA by dopamine in learning. *Nature* **590**, 451–456 (2021).
- 1492 36. Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C., Hassabis, D., Munos, R. &
1493 Botvinick, M. A distributional code for value in dopamine-based reinforcement learning.
1494 *Nature* **1**, (2019).
- 1495 37. Bellemare, M. G., Dabney, W. & Munos, R. A distributional perspective on reinforcement
1496 learning. *34th International Conference on Machine Learning, ICML 2017* **1**, 693–711
1497 (2017).
- 1498 38. Lowet, A. S., Zheng, Q., Matias, S., Drugowitsch, J. & Uchida, N. Distributional
1499 Reinforcement Learning in the Brain. *Trends Neurosci.* **43**, 980–997 (2020).
- 1500 39. Rowland, M., Dadashi, R., Kumar, S., Munos, R., Bellemare, M. G. & Dabney, W. Statistics
1501 and samples in distributional reinforcement learning. *36th International Conference on*
1502 *Machine Learning, ICML 2019* **6**, 9727–9750 (2019).
- 1503 40. Mihatsch, O. & Neuneier, R. Risk-sensitive reinforcement learning. *Mach. Learn.* **49**, 267–
1504 290 (2002).
- 1505 41. Bellemare, M. G. & Dabney, W. *Distributional reinforcement learning*. (MIT Press, 2023).
- 1506 42. Jones, M. C. Expectiles and M-quantiles are quantiles. *Stat. Probab. Lett.* **20**, 149–153
1507 (1994).
- 1508 43. Houk, J., Davis, J., & Beiser, D. *Models of information processing in the basal ganglia*.
1509 (Bradford Books, 2019).
- 1510 44. Gerfen, C. The neostriatal mosaic: Multiple levels of compartmental organization in the basal
1511 ganglia. *Annu. Rev. Neurosci.* **15**, 285–320 (1992).

- 1512 45. Smith, Y., Bevan, M. D., Shink, E. & Bolam, J. P. Microcircuitry of the direct and indirect
1513 pathways of the basal ganglia. *Neuroscience* **86**, 353–387 (1998).
- 1514 46. Kravitz, A. V., Tye, L. D. & Kreitzer, A. C. Distinct roles for direct and indirect pathway
1515 striatal neurons in reinforcement. *Nat. Neurosci.* **15**, 816–818 (2012).
- 1516 47. Reynolds, J. N. J., Hyland, B. I. & Wickens, J. R. A cellular mechanism of reward-related
1517 learning. *Nature* **413**, 67–70 (2001).
- 1518 48. Gerfen, C. R. & Surmeier, D. J. Modulation of striatal projection systems by dopamine.
1519 *Annu. Rev. Neurosci.* **34**, 441–466 (2011).
- 1520 49. Richfield, E. K., Penney, J. B. & Young, A. B. Anatomical and affinity state comparisons
1521 between dopamine D1 and D2 receptors in the rat central nervous system. *Neuroscience* **30**,
1522 767–777 (1989).
- 1523 50. Rice, M. E. & Cragg, S. J. Dopamine spillover after quantal release: Rethinking dopamine
1524 transmission in the nigrostriatal pathway. *Brain Res. Rev.* **58**, 303–313 (2008).
- 1525 51. Gonon, F. G. & Buda, M. J. Regulation of dopamine release by impulse flow and by
1526 autoreceptors as studied by in vivo voltammetry in the rat striatum. *Neuroscience* **14**, 765–
1527 774 (1985).
- 1528 52. Dodson, P. D., Dreyer, J. K., Jennings, K. A., Syed, E. C. J., Wade-Martins, R., Cragg, S. J.,
1529 Bolam, J. P. & Magill, P. J. Representation of spontaneous movement by dopaminergic
1530 neurons is cell-type selective and disrupted in parkinsonism. *Proc. Natl. Acad. Sci. U. S. A.*
1531 **113**, E2180–E2188 (2016).
- 1532 53. Marcott, P. F., Mamaligas, A. A. & Ford, C. P. Phasic dopamine release drives rapid
1533 activation of striatal D2-receptors. *Neuron* **84**, 164–176 (2014).
- 1534 54. Jaskir, A. & Frank, M. J. On the normative advantages of dopamine and striatal opponency
1535 for learning and choice. *Elife* (2023).
- 1536 55. Tian, J. & Uchida, N. Habenula Lesions Reveal that Multiple Mechanisms Underlie
1537 Dopamine Prediction Errors. *Neuron* **87**, 1304–1316 (2015).
- 1538 56. Cui, Y., Yang, Y., Ni, Z., Dong, Y., Cai, G., Foncelle, A., Ma, S., Sang, K., Tang, S., Li, Y.,
1539 Shen, Y., Berry, H., Wu, S. & Hu, H. Astroglial Kir4.1 in the lateral habenula drives
1540 neuronal bursts in depression. *Nature* **554**, 323–327 (2018).

- 1541 57. Li, B., Piriz, J., Mirrione, M., Chung, C., D. Proulx, C., Schulz, D., Henn, F. & Malinow, R.
1542 Synaptic potentiation onto habenula neurons in the learned helplessness model of depression.
1543 *Nature* **470**, 535–541 (2011).
- 1544 58. Yang, Y., Cui, Y., Sang, K., Dong, Y., Ni, Z., Ma, S. & Hu, H. Ketamine blocks bursting in
1545 the lateral habenula to rapidly relieve depression. *Nature* **554**, 317–322 (2018).
- 1546 59. Dreyer, J. K., Herrik, K. F., Berg, R. W. & Hounsgaard, J. D. Influence of phasic and tonic
1547 dopamine release on receptor activation. *J. of Neurosci.* **30**, 14273–14283 (2010).
- 1548 60. Mikhael, J. G., Lai, L. & Gershman, S. J. *Rational inattention and tonic dopamine*. *PLoS*
1549 *Comput. Biol.* **17** e1008659 (2021).
- 1550 61. Vingerhoets, F. J., Snow, B. J., Schulzer, M., Morrison, S., Ruth, T. J., Holden, J. E., Cooper,
1551 S. & Calne, D. B. Reproducibility of fluorine-18-6-fluorodopa positron emission tomography
1552 in normal human subjects. *J. Nucl. Med.* **35**, 18–24 (1994).
- 1553 62. Cools, R., Gibbs, S. E., Miyakawa, A., Jagust, W. & D’Esposito, M. Working memory
1554 capacity predicts dopamine synthesis capacity in the human striatum. *J. Neurosci.* **28**, 1208–
1555 1212 (2008).
- 1556 63. Hoffmann, I. S. & Cubeddu, L. X. Differential effects of bromocriptine on dopamine and
1557 acetylcholine release modulatory receptors. *J. Neurochem.* **42**, 278–282 (1984).
- 1558 64. Tissari, A. H., Rossetti, Z. L., Meloni, M., Frau, M. I. & Gessa, G. L. Autoreceptors mediate
1559 the inhibition of dopamine synthesis by bromocriptine and lisuride in rats. *Eur. J.*
1560 *Pharmacol.* **91**, 463–468 (1983).
- 1561 65. Lieberman, A. Depression in Parkinson’s disease -- a review. *Acta Neurol. Scand.* **113**, 1–8
1562 (2006).
- 1563 66. Leentjens, A. F. G., Van den Akker, M., Metsemakers, J. F. M., Lousberg, R. & Verhey, F.
1564 R. J. Higher incidence of depression preceding the onset of Parkinson’s disease: a register
1565 study. *Mov. Disord.* **18**, 414–418 (2003).
- 1566 67. Nilsson, F. M., Kessing, L. V., Sørensen, T. M., Andersen, P. K. & Bolwig, T. G. Major
1567 depressive disorder in Parkinson’s disease: a register-based study. *Acta Psychiatr. Scand.*
1568 **106**, 202–211 (2002).

- 1569 68. Remy, P., Doder, M., Lees, A., Turjanski, N. & Brooks, D. Depression in Parkinson's
1570 disease: loss of dopamine and noradrenaline innervation in the limbic system. *Brain* **128**,
1571 1314–1322 (2005).
- 1572 69. Weintraub, D., Newberg, A. B., Cary, M. S., Siderowf, A. D., Moberg, P. J., Kleiner-Fisman,
1573 G., Duda, J. E., Stern, M. B., Mozley, D. & Katz, I. R. Striatal dopamine transporter imaging
1574 correlates with anxiety and depression symptoms in Parkinson's disease. *J. Nucl. Med.* **46**,
1575 227–232 (2005).
- 1576 70. Kish, S. J., Shannak, K. & Hornykiewicz, O. Uneven pattern of dopamine loss in the striatum
1577 of patients with idiopathic Parkinson's disease. Pathophysiologic and clinical implications.
1578 *N. Engl. J. Med.* **318**, 876–880 (1988).
- 1579 71. Timmer, M. H. M., Sescousse, G., Esselink, R. A. J., Piray, P. & Cools, R. Mechanisms
1580 Underlying Dopamine-Induced Risky Choice in Parkinson's Disease With and Without
1581 Depression (History). *Comput Psychiatr* **2**, 11–27 (2018).
- 1582 72. Hikida, T., Kimura, K., Wada, N., Funabiki, K. & Nakanishi Shigetada, S. Distinct Roles of
1583 Synaptic Transmission in Direct and Indirect Striatal Pathways to Reward and Aversive
1584 Behavior. *Neuron* **66**, 896–907 (2010).
- 1585 73. Hikida, T., Yawata, S., Yamaguchi, T., Danjo, T., Sasaoka, T., Wang, Y. & Nakanishi S.
1586 Pathway-specific modulation of nucleus accumbens in reward and aversive behavior via
1587 selective transmitter receptors. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 342–347 (2013).
- 1588 74. Danjo, T., Yoshimi, K., Funabiki, K., Yawata, S. & Nakanishi, S. Aversive behavior induced
1589 by optogenetic inactivation of ventral tegmental area dopamine neurons is mediated by
1590 dopamine D2 receptors in the nucleus accumbens. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 6455–
1591 6460 (2014).
- 1592 75. Yamaguchi, T., Goto, A., Nakahara, I., Yawata, S., Hikida, T., Matsuda, M., Funabiki, K. &
1593 Nakanishi, S. Role of PKA signaling in D2 receptor-expressing neurons in the core of the
1594 nucleus accumbens in aversive learning. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11383–11388
1595 (2015).
- 1596 76. Bayer, H. M. & Glimcher, P. W. Midbrain dopamine neurons encode a quantitative reward
1597 prediction error signal. *Neuron* **47**, 129–141 (2005).

- 1598 77. Hart, A. S., Rutledge, R. B., Glimcher, P. W. & Phillips, P. E. M. Phasic dopamine release in
1599 the rat nucleus accumbens symmetrically encodes a reward prediction error term. *J.*
1600 *Neurosci.* **34**, 698–704 (2014).
- 1601 78. Grace, A. A. Dysregulation of the dopamine system in the pathophysiology of schizophrenia
1602 and depression. *Nat. Rev. Neurosci.* **17**, 524–532 (2016).
- 1603 79. Anstrom, K. K., Miczek, K. A. & Budygin, E. A. Increased phasic dopamine signaling in the
1604 mesolimbic pathway during social defeat in rats. *Neuroscience* **161**, 3–12 (2009).
- 1605 80. Razzoli, M., Andreoli, M., Michielin, F., Quarta, D. & Sokal, D. M. Increased phasic activity
1606 of VTA dopamine neurons in mice 3 weeks after repeated social defeat. *Behav. Brain Res.*
1607 **218**, 253–257 (2011).
- 1608 81. Markovic, T., Pederson, C. E., Massaly, N., Vachez, Y., Ruyle, B., Murphy, C. A.,
1609 Abiraman, K., Hoon Shin, J., Garcia, J. H., Jean Yoon, H., Alvarez, V. A., Bruchas, M. R.,
1610 Creed, M. C. & Moron, J. A. Pain induces adaptations in ventral tegmental area dopamine
1611 neurons to drive anhedonia-like behavior. *Nat. Neurosci.* **24**, 1601–1613 (2021).
- 1612 82. Guo, Z., Li, S., Wu, J., Zhu, X. & Zhang, Y. Maternal deprivation increased vulnerability to
1613 depression in adult rats through DRD2 promoter methylation in the ventral tegmental area.
1614 *Front. Psychiatry* **13**, 827667 (2022).
- 1615 83. Peng, B., Hu, Q., Liu, J., Guo, S., Borgland, S. L. & Liu, S. Corticosterone attenuates
1616 reward-seeking behavior and increases anxiety via D2 receptor signaling in ventral tegmental
1617 area dopamine neurons. *J. Neurosci.* **41**, 1566–1581 (2021).
- 1618 84. Tye, K. M., Mirzabekov, J. J., Warden, M. R., Ferenczi, E. A., Tsai, H-C., Finkelstein, J.,
1619 Kim, S-Y., Adhikari, A., Thompson, K. R., Andalman A. S., Gunaydin, L. A., Witten I. &
1620 Deisseroth K. Dopamine neurons modulate neural encoding and expression of depression-
1621 related behaviour. *Nature* **493**, 537–541 (2013).
- 1622 85. Baek, K., Kwon, J., Chae, J. H., Chung, Y. A., Kralik, J. D., Min, J. A., Huh, H., Choi, K.
1623 M., Jang, K. I., Lee, N. B., Kim, S., Peterson, B. S., & Jeong, J. Heightened aversion to risk
1624 and loss in depressed patients with a suicide attempt history. *Sci. Rep.* **7**, 11228 (2017).

- 1625 86. Smoski, M. J., Lynch, T. R., Rosenthal, M. Z., Cheavens, J. S., Chapman, A. L., & Krishnan,
1626 R. R. Decision-making and risk aversion among depressive adults. *J. Behav. Ther. Exp.*
1627 *Psychiatry* **39**, 567–576 (2008).
- 1628 87. van Holst, R. J., Sescousse, G., Janssen, L. K., Janssen, M., Berry, A. S., Jagust, W. J., &
1629 Cools, R. Increased Striatal Dopamine Synthesis Capacity in Gambling Addiction. *Biol.*
1630 *Psychiatry* **83**, 1036–1043 (2018).
- 1631 88. Cools, R., Altamirano, L. & D’Esposito, M. Reversal learning in Parkinson’s disease
1632 depends on medication status and outcome valence. *Neuropsychologia* **44**, 1663–1673
1633 (2006).
- 1634 89. Cools, R., Barker, R. A., Sahakian, B. J. & Robbins, T. W. Enhanced or impaired cognitive
1635 function in Parkinson’s disease as a function of dopaminergic medication and task demands.
1636 *Cereb. Cortex* **11**, 1136–1143 (2001).
- 1637 90. Cools, R., Barker, R. A., Sahakian, B. J. & Robbins, T. W. L-Dopa medication remediates
1638 cognitive inflexibility, but increases impulsivity in patients with Parkinson’s disease.
1639 *Neuropsychologia* **41**, 1431–1441 (2003).
- 1640 91. Swainson, R., Rogers, R. D., Sahakian, B. J., Summers, B. A., Polkey, C. E., & Robbins, T.
1641 W. Probabilistic learning and reversal deficits in patients with Parkinson’s disease or frontal
1642 or temporal lobe lesions: possible adverse effects of dopaminergic medication.
1643 *Neuropsychologia* **38**, 596–612 (2000).
- 1644 92. Cox, S. M., Frank, M. J., Larcher, K., Fellows, L. K., Clark, C. A., Leyton, M., & Dagher, A.
1645 Striatal D1 and D2 signaling differentially predict learning from positive and negative
1646 outcomes. *Neuroimage* **109**, 95–101 (2015).
- 1647 93. Savitz, J. B. & Drevets, W. C. Neuroreceptor imaging in depression. *Neurobiol. Dis.* **52**, 49–
1648 65 (2013).
- 1649 94. Rescorla, R. A. & Wagner, A. R. *A theory of Pavlovian conditioning: Variations in the*
1650 *effectiveness of reinforcement and nonreinforcement, Classical Conditioning II.* 64–99
1651 (1972).

- 1652 95. Kim, H. R., Malik, A. N., Mikhael, J. G., Bech, P., Tsutsui-Kimura, I., Sun, F., Zhang, Y., Li,
1653 Y., Watabe-Uchida, M., Gershman, S. J. & Uchida, N. A unified framework for dopamine
1654 signals across timescales. *Cell* **183**, 1600-1616.e25 (2020).
- 1655 96. Bertsekas, D. P. & Tsitsiklis, J. *Neuro-Dynamic Programming*. (Athena Scientific, 1996).
- 1656 97. Mierau, J., Schneider, F. J., Ensinger, H. A., Chio, C. L., Lajiness, M. E., & Huff, R. M.
1657 Pramipexole binding and activation of cloned and expressed dopamine D2, D3 and D4
1658 receptors. *Eur. J of Pharmac.* **290**, 29-36 (1995).
- 1659 98. Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B. & Uchida, N. Neuron-type-specific signals
1660 for reward and punishment in the ventral tegmental area. *Nature* **482**, 85–88 (2012).
- 1661 99. Kvitsiani, D., Ranade, S., Hangya, B., Taniguchi, H., Huang, J. Z. & Kepec, A. Distinct
1662 behavioural and network correlates of two interneuron types in prefrontal cortex. *Nature* **498**,
1663 363–366 (2013).
- 1664 100. Stauffer, W. R., Lak, A. & Schultz, W. Dopamine reward prediction error responses reflect
1665 marginal utility. *Curr. Biol.* **24**, 2491–2500 (2014).
- 1666
- 1667
- 1668
- 1669
- 1670

1671 **Acknowledgements**

1672 We thank Ju Tian for providing the data used in this study; Jakob Dreyer for making the
1673 biophysical model available for open-source use; Roshan Cools, Peter Dayan and Sam Gershman
1674 for valuable feedback on the manuscript; Dan Polley, Mark Andermann and Elizabeth Phelps for
1675 mentorship and helpful discussions. We thank Bernardo Sabatini for discussions, and Isobel
1676 Green, Paul Masset, Mitsuko Watabe-Uchida and all the rest of the Uchida lab members for
1677 helpful comments at the different stages of this work. This work is supported by NIH BRAIN
1678 Initiative grants (R01NS116753, U19NS113201), the Simons Collaboration on Global Brain,
1679 and the Bipolar Disorder Seed Grant from the Harvard Brain Science Institute.

1680

1681 **Author information**

1682 **Contributions**

1683 S.R.P. and N.U. conceived the project. S.R.P performed the modeling work. S.R.P. wrote the
1684 first draft and S.R.P and N.U. edited the paper.

1685

1686 **Corresponding author**

1687 Correspondence to Naoshige Uchida (uchida@mcb.harvard.edu) and Sandra Romero Pinto
1688 (sromeropinto@g.harvard.edu).

1689

1690

1691

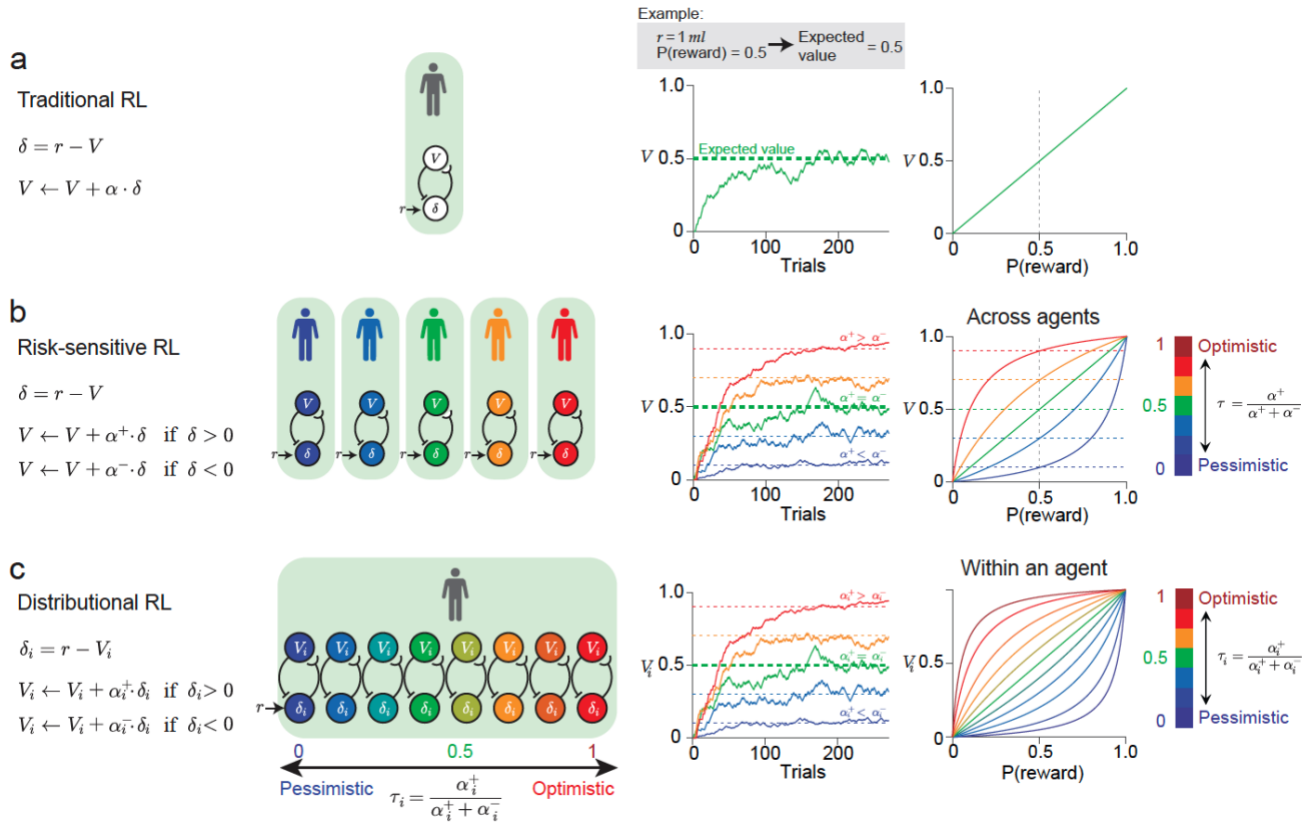
1692 **Competing interest statement**

1693 The authors declare no competing interests.

1694

1695

1696 **Figures**



1697

1698 **Figure 1 | Reinforcement learning models.**

1699 **a.** Traditional reinforcement learning with a single learning rate (α) for both positive and
1700 negative RPEs (δ) for the value updates (left). This update rule makes value estimate (V)
1701 converge on the expected value of the reward distribution (middle). When the reward probability
1702 is varied (i.e., for Bernoulli distributions), the V at convergence scales linearly with the reward
1703 probability (right).

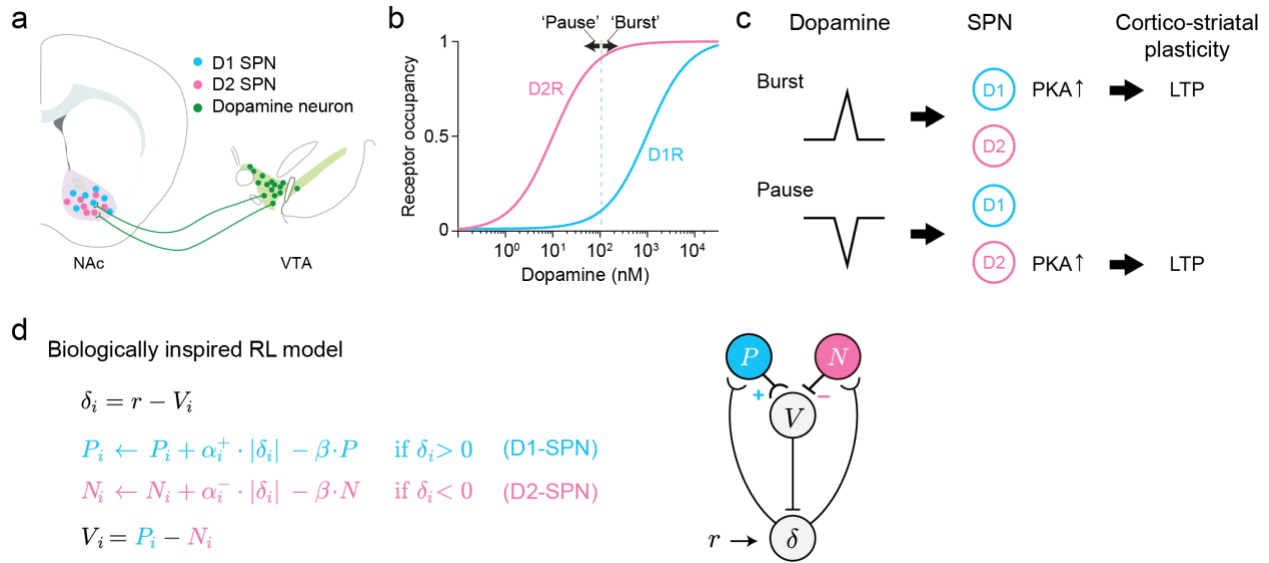
1704 **b.** Risk-sensitive reinforcement learning with different learning rates (α^+ , α^-) for positive and
1705 negative RPEs, respectively (left). This update rule makes value estimate (V) converge on the
1706 quantities that are higher or lower than the expected value of the reward distribution (middle). As
1707 the reward probabilities are varied, the convexity of the convergent value V changes depending
1708 on the asymmetry between α^+ and α^- (Methods 1.3.3). The level of the bias is determined by
1709 the asymmetric learning rate parameter τ (right).

1710 **c.** Distributional reinforcement learning contains a set of value predictors (V_i) each with a given
1711 learning rate for positive and negative RPEs (α_i^+ , α_i^- , respectively) (left). This makes each value

1712 predictor converge on the quantity equal to the τ_i -th expectile of the reward distribution. Thus,
1713 each value V_i represents an expectile, and together the set of V_i represents the entire distribution
1714 (Methods 1.2) (right).

1715

1716



1717

1718 **Figure 2 | Biologically inspired reinforcement learning model.**

1719 **a.** Schematic of the basal ganglia circuitry. Dopaminergic neurons in the VTA modulate
1720 plasticity at the level of the cortico-striatal synapses on SPNs in the NAc. The SPNs are
1721 subdivided depending on the dopamine receptor type they express (D1R or D2R).

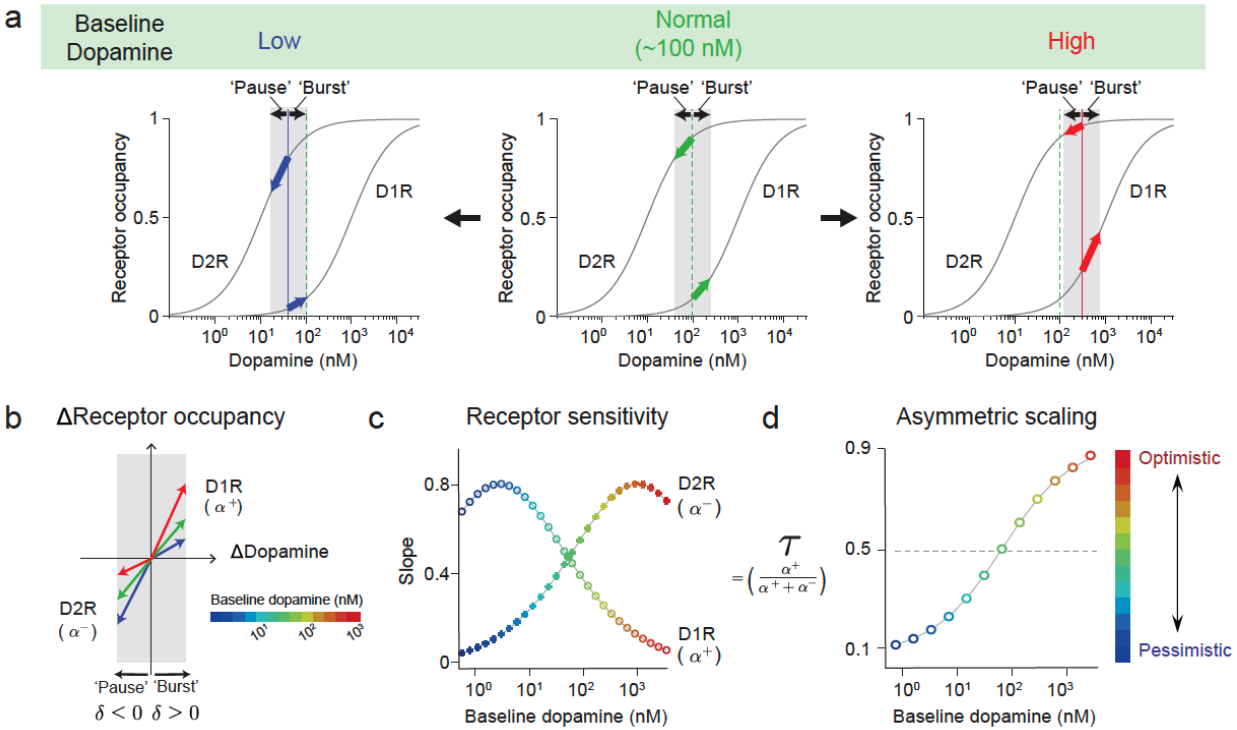
1722 **b.** Dose-occupancy curves for the D1R and D2R describing receptor occupancies as a function of
1723 dopamine concentrations. The curves are shifted between each other due to the different
1724 affinities of the receptors. The arrows represent 3-fold increase (“burst”) and decrease (“pause”)
1725 in dopamine concentrations, which causes left-ward or right-ward shifts of the same magnitudes
1726 in the log-scale.

1727 **c.** Schematic of the plasticity rules of VTA-NAc circuitry^{33–35}. Transient increases in dopamine,
1728 caused by bursts in firing rate of dopamine neurons, generates increases in PKA activity in D1R-
1729 expressing SPNs, leading to LTP in the cortico-striatal synapses. Transient decreases in
1730 dopamine, caused by pauses in firing rate of dopamine neurons, generates increases in PKA
1731 activity in D2R-expressing SPNs, leading to LTP in the cortico-striatal synapses.

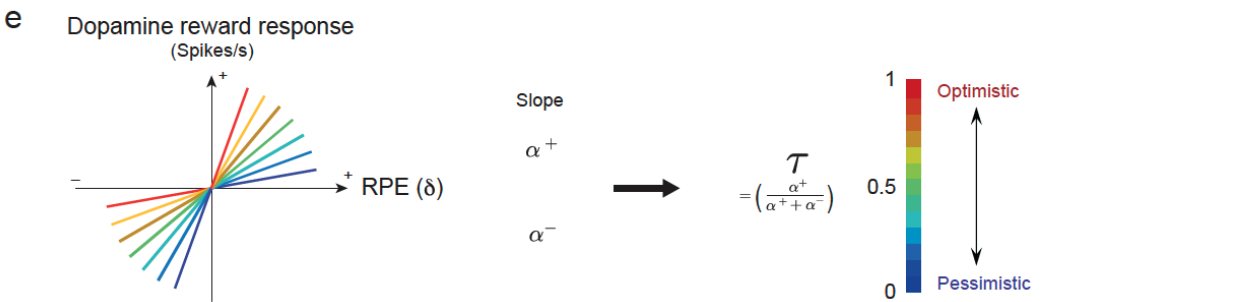
1732 **d.** Schematic and equations of biologically inspired reinforcement learning model³²

1733 VTA, ventral tegmental area; NAc, nucleus accumbens; SPN, spiny projection neurons; D1R,
1734 D1-type dopamine receptor; D2R, D2-type dopamine receptor; PKA, protein kinase A; LTP,
1735 long-term potentiation.

Model 1 Baseline dopamine → Asymmetric learning rates (τ)



Model 2 Different scaling of phasic dopamine → Asymmetric learning rates (τ)



1736

1737 **Figure 3 | Potential mechanisms for asymmetric learning.**

1738 **a.** Schematic of the mechanism by which increases or decreases in baseline dopamine modulates
 1739 the degree to which bursts and pauses in dopamine causes changes in D1R and D2R occupancy.

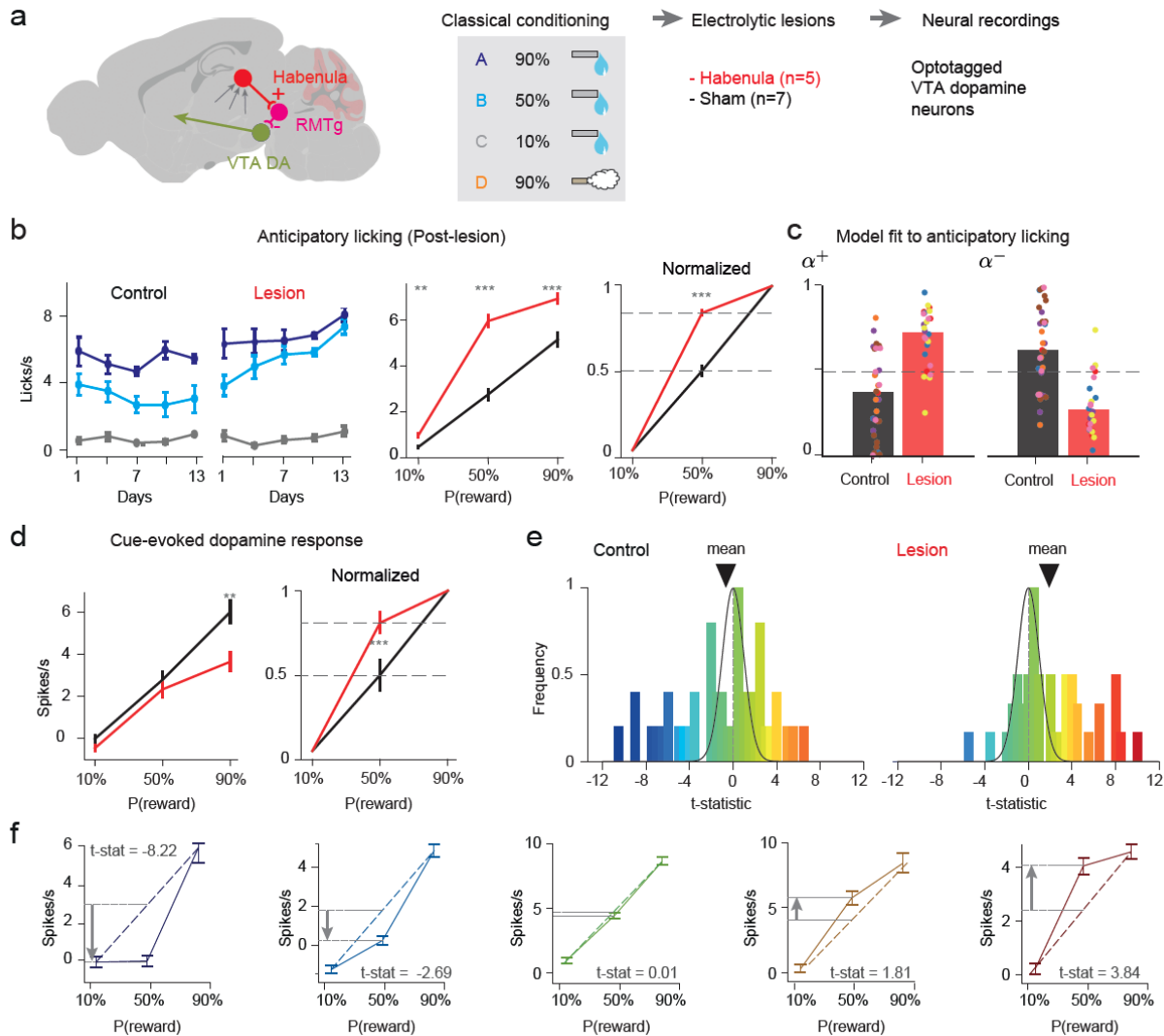
1740 Increases in baseline dopamine makes dopamine pauses to cause greater decreases in D2R

1741 occupancy than the increases in D1R occupancy caused by dopamine bursts. Conversely,

1742 decreases in dopamine, makes dopamine bursts to cause smaller increases in D1R occupancy

1743 than the decreases in D2R occupancy caused by dopamine pauses.

- 1744 **b.** Schematic of the change in receptor occupancies in D1R and D2R, for a given transient
1745 increase (‘burst’) or decrease (‘pause’) in dopamine, receptively. A pause and a burst in
1746 dopamine correspond to $\delta < 0$ and $\delta > 0$ in the model. The slope is modulated by the baseline
1747 dopamine (colormap) and corresponds to the receptor’s sensitivity to dopamine transients.
- 1748 **c.** Receptor sensitivity for D1R and D2R as a function of baseline dopamine. In Model 1, we
1749 assume that the receptor sensitivity acts as a scaling factor on the PKA activity induced by burst
1750 and pauses. That is, $PKA_{D1} \propto \alpha^+ \cdot \delta \cdot \mathbf{1}_{\delta>0}$ and $PKA_{D2} \propto \alpha^- \cdot \delta \cdot \mathbf{1}_{\delta<0}$.
- 1751 **d.** Asymmetric scaling factor (τ) as a function of baseline dopamine. Colors depict how
1752 ‘optimistic’ or ‘pessimistic’ the convergent value estimate will be when learning with a given τ .
- 1753 **e.** Model 2. Left, the relationship between dopamine reward responses (spikes/s) and RPEs. The
1754 slopes of these response functions correspond to the asymmetric learning rates (α^+ , α^-) for
1755 positive and negative RPEs, respectively. Colors depict how ‘optimistic’ or ‘pessimistic’ the
1756 convergent value estimate will be when learning with a given asymmetric scaling factor.
- 1757
- 1758



1759

1760 **Figure 4 | Habenula lesions leads to optimistic reward-seeking behavior and cue-evoked**
 1761 **responses in dopamine neurons.**

1762 **a.** Schematic of the experiment performed by Tian and Uchida (2015)⁵⁵. Animals were trained in
 1763 a classical conditioning task in which 3 odor cues predicted rewards of different probabilities
 1764 (10%, 50%, 90%) and one odor cue predicted 80% probability of an air puff. Animals then
 1765 underwent habenula ($n = 5$) or sham ($n = 7$) lesions and trained on the task again. The neural
 1766 recordings were performed from optotagged VTA dopamine neurons once behavior stabilized.

1767 **b.** Anticipatory licking across sessions after lesions (left.). There was a significant increase in
 1768 anticipatory licking to the 10% (U-statistic = -2.895 , $P = 0.003792$, two-sided Mann-Whitney U-
 1769 test), 50% (U-statistic = -5.579 , $P < 1 \times 10^{-9}$, two-sided Mann-Whitney U-test) and 90% (U-

1770 statistic = -3.487 , $P = 0.00048$, two-sided Mann-Whitney U-test) cues ($n = 31$ for control $n = 30$
1771 for lesion) that results from progressive changes across sessions. The anticipatory licking curves
1772 show a linear scaling with reward probability in the control group, and a convex curve for the
1773 lesion group (mean \pm s.e.m across animals, U-statistic = -6.444 , $P < 1 \times 10^{-7}$, two-sided Mann-
1774 Whitney U-test for the 50% cue normalized response). These curves are predicted by RL agents
1775 with symmetric and asymmetric ($\alpha^+ > \alpha^-$) learning rates for the control and lesion groups,
1776 respectively, assuming a linear mapping between anticipatory licking and value prediction.

1777 **c.** RL model fits to the anticipatory licking on a trial-by-trial basis using a risk-sensitive RL
1778 models that allows for separate learning rates of positive and negative RPEs. Each dot represents
1779 a session ($n = 35$ control, $n = 30$ lesion) and each color a mouse ($n = 7$ control, $n = 5$ lesion). The
1780 fits show a significant difference in the learning rates between control and lesion groups (U-
1781 statistic = -4.679 , $P < 1.0 \times 10^{-5}$, pooling sessions across mice in each group).

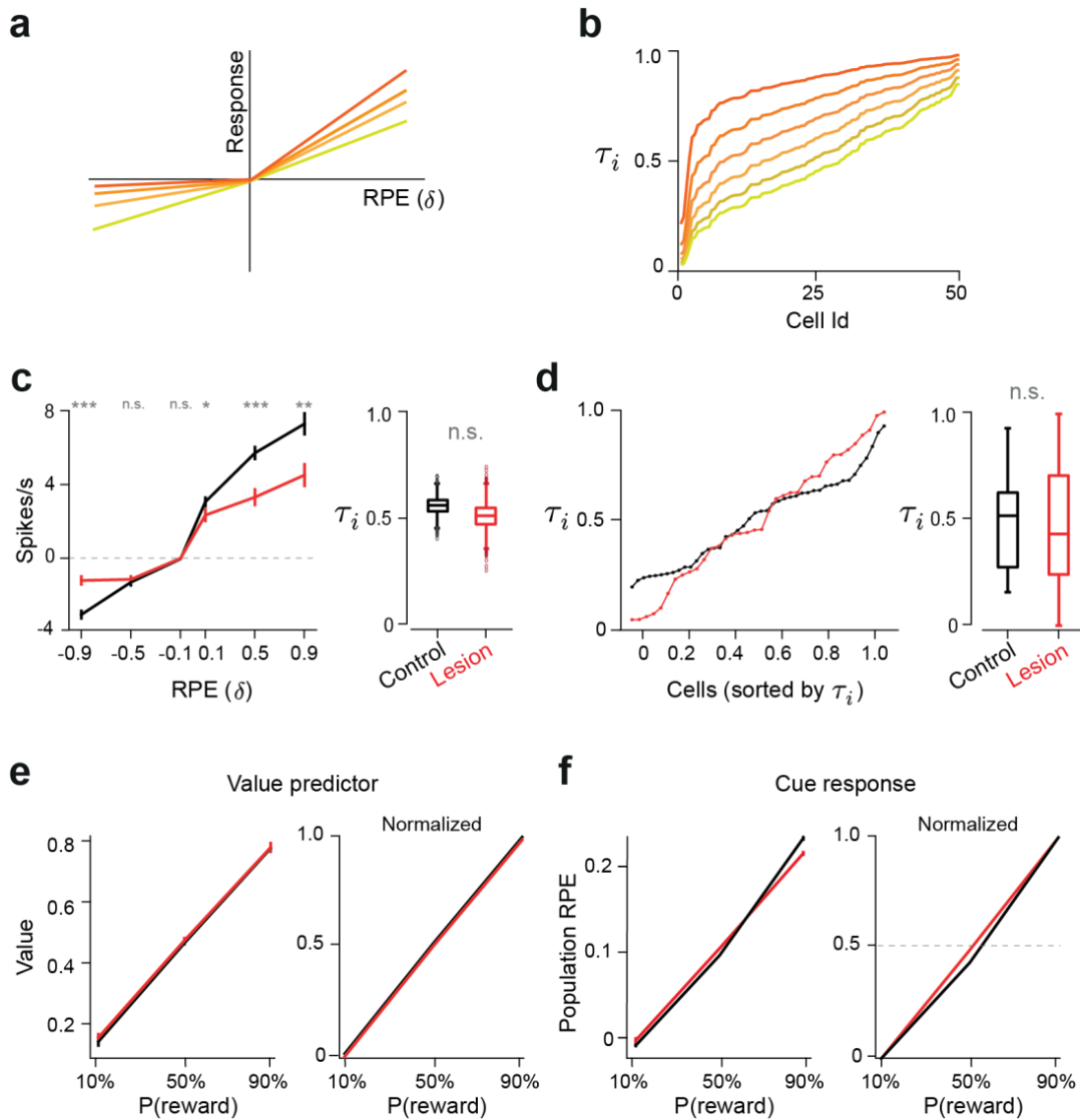
1782 **d.** Cue-evoked dopamine responses from opto-tagged VTA dopamine neurons (mean \pm s.e.m
1783 across neurons, $n = 45$ control group, $n = 44$ lesion group). There was a decrease in the absolute
1784 magnitude of responses to the 90% cue (U-statistic = 3.249 , $P = 0.0011$, two-sided Mann-
1785 Whitney U-test) after habenula lesions (left). The normalized cue-evoked responses show the
1786 similar pattern as the normalized anticipatory-licking with a linear and convex function for the
1787 control and lesion groups, respectively, with a significant increase in normalized response to the
1788 50% cue after lesions (U-statistic = -3.824 , $P = 0.000131$, two-sided Mann-Whitney U-test)
1789 These curves are predicted by agents with symmetric and asymmetric learning rates for control
1790 and lesion groups, respectively.

1791 **e.** Distribution of t -statistics comparing the cue-evoked response to the linear interpolation point
1792 between the 90% and 10% cue-evoked responses for each dopamine neuron. The distribution of
1793 t -statistics for the control and lesion cases was wider than what is expected from random noise
1794 (Monte Carlo test for standard deviation different from zero: $P = 0.0222$ control, $P = 0.0217$
1795 lesion, 1000 batches). The distribution was shifted to values larger than 0 in the lesion case
1796 (Monte Carlo test for mean larger than zero: $P = 1$ control, $P = 0.022$ lesion, 1000 batches)
1797 indicative of an optimistic bias in the distribution. The lesion group distribution was also
1798 significantly shifted to higher values with respect to the control group distribution (U-statistic =
1799 -2.815 , $P = 0.0024$, single-sided Mann-Whitney U-test). Arrow heads: the mean of the t -
1800 statistics.

1801 **f.** Example of t -statistics calculations for dopamine neurons taken from the control group (mean
1802 \pm s.e.m across trials). A t -statistic value close to 0 indicates linear scaling of cue-evoked
1803 responses with reward probability; a t -statistics value lower or greater than 0 indicates a concave
1804 or convex function of cue-evoked responses against reward probability, indicative of a
1805 pessimistic or an optimistic bias, respectively.

1806

1807



1808

1809 **Figure 5 | Model 2 cannot explain optimistic biases in behavior and cue-evoked dopamine**
 1810 **responses of habenula lesioned animals.**

1811 **a.** Possible changes in habenula lesion mice that could explain optimistic biases based on Model
 1812 2. At the level of the population dopamine responses, an optimistic bias can be caused by an
 1813 increase in the slope of the average reward responses to positive RPEs and/or a decrease in the
 1814 slope of the average reward responses to negative RPEs.

1815 **b.** At the level of the distribution of individual dopamine neuron responses, an optimistic bias
 1816 can be caused by an overall increase in the mean of the distribution of asymmetric scaling factors
 1817 (τ_i), computed from each individual neuron response function.

1818 **c.** Observed reward responses as a function of RPEs, averaged across the population of dopamine
1819 neurons for the control and lesion groups (left, mean \pm s.e.m across neurons, $n = 45$ control
1820 group, $n = 44$ lesion group). There was a significant decrease in the reward responses for the
1821 50% cue (U-statistic = 3.726, $P = 0.000195$, two-sided Mann-Whitney U-test) and 90% cue (U-
1822 statistic = 2.987, $P = 0.00281$, two-sided Mann-Whitney U-test), and for the omission responses
1823 for the 90% cue (U-statistic = -4.940, $P < 10^{-4}$, two-sided Mann-Whitney U-test). Distribution of
1824 asymmetric scaling factors (τ), computed from the average response function over the recorded
1825 neurons for the control and lesion groups (right). The distributions are the result of bootstrapping
1826 by randomly sampling neurons in 5,000 iterations. The distribution of differences between the
1827 obtained asymmetric scaling factors ($\tau_{lesion} - \tau_{control}$) was not significantly larger than zero (5th
1828 percentile = -0.1605).

1829 **d.** Distribution of asymmetric scaling factors (τ_i), computed from each individual neuron
1830 response function for the control and lesion groups. Each dot represents a single neuron ($n = 45$
1831 control group, $n = 44$ lesion group), and the neurons were sorted by asymmetric scaling factors
1832 (τ_i). The means were not significantly different (right) (t -statistic = 0.3277, $P = 0.627$, t -test).

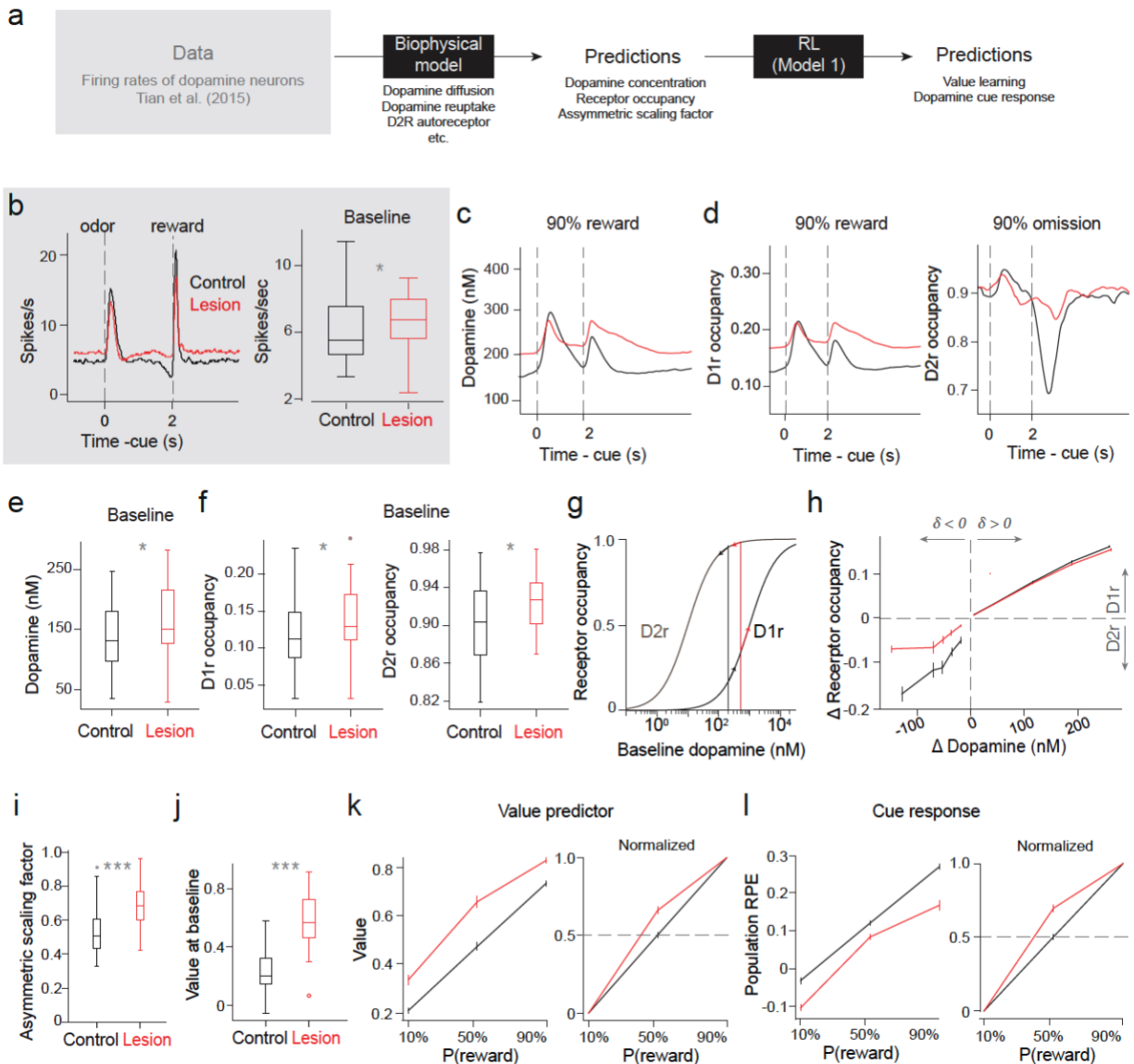
1833 **e.** Value predictions based on a TD learning model trained using the assumptions of Model 2 and
1834 the asymmetric scaling factors derived from the data. The model did not show any optimistic
1835 bias in the value predictors of the model trained with the lesion-derived asymmetric scaling
1836 factors.

1837 **f.** TD errors at cue show no signs of an optimistic bias in the model trained with the lesion-
1838 derived asymmetric scaling factors.

1839 Centre of box plot shows the median; edges are 25th and 75th percentiles; and whiskers are the
1840 most extreme data points not considered as outliers.

1841

1842



1843

1844 **Figure 6 | Model 1 can account for optimistic biases in reward-seeking behavior and cue-**
 1845 **evoked dopamine responses.**

1846 **a.** Schematic of the analysis. A biophysical model was used to predict dopamine concentrations,
 1847 receptor occupancies, and value learning based on firing rates of dopamine neurons recorded in
 1848 Tian et al. (2015).

1849 **b.** Average firing rates of dopamine neurons across the population for the control and lesion
 1850 groups (left, $n = 45$ control group, $n = 44$ lesion group). Baseline firing rates were significantly
 1851 greater in the lesion compared to the control group (right) (U-statistic = -2.429 , $P = 0.0151$,
 1852 single-sided Mann-Whitney U-test).

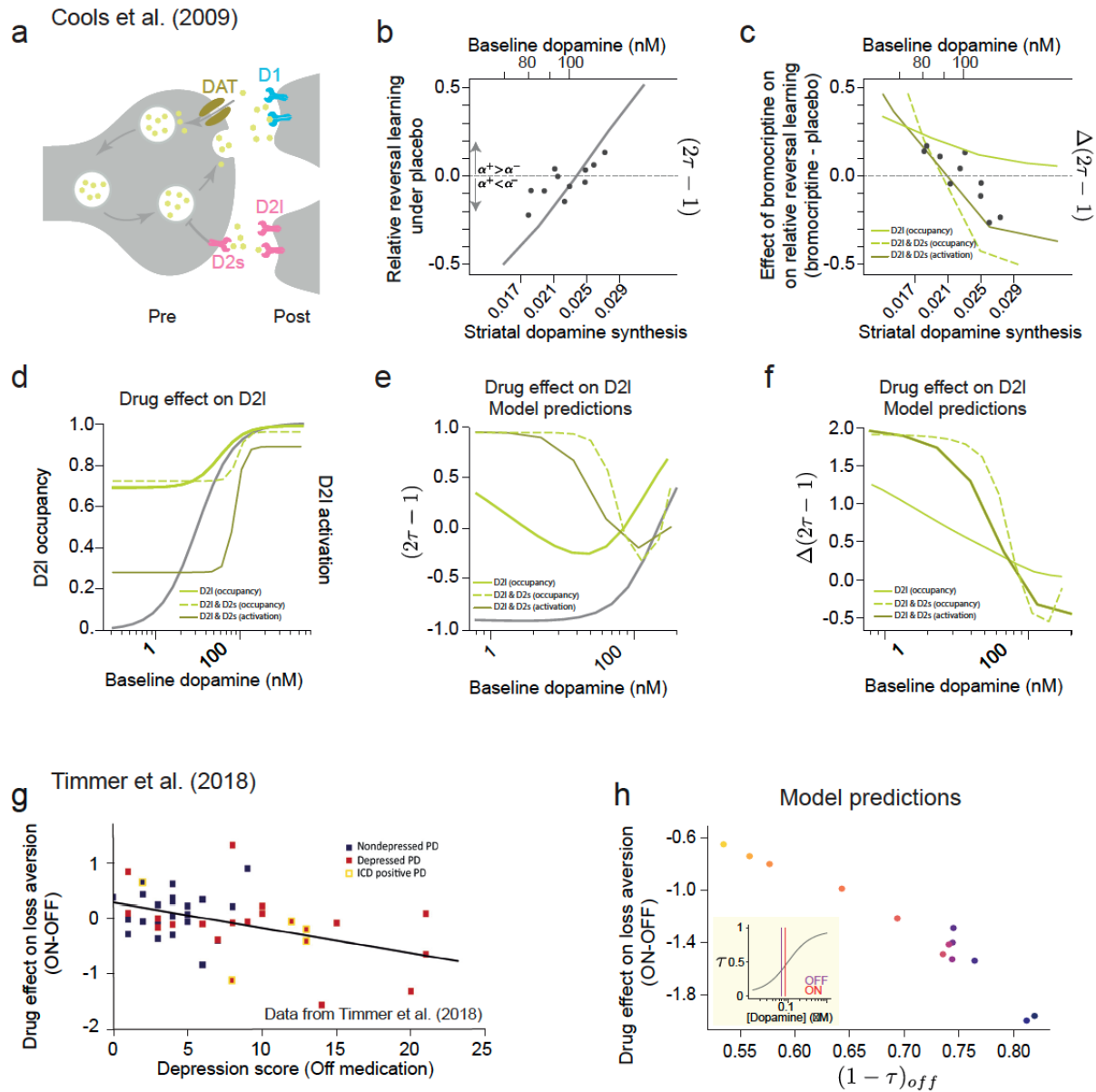
- 1853 **c.** Dopamine concentrations predicted from the firing rates of dopamine neurons based on the
1854 biophysical model of dopamine. Predictions for 90% reward trials are shown.
- 1855 **d.** Receptor occupancies predicted by the same biophysical model. Predictions for rewarded
1856 (left) and reward omission (right) trials in 90% reward trials are shown separately for D1R (left)
1857 and D2R (right), respectively ($n = 45$ control group, $n = 44$ lesion group).
- 1858 **e.** Mean dopamine concentrations at baseline predicted by the model (U-statistic = -2.109 , $P =$
1859 0.0175 , single-sided Mann-Whitney U-test).
- 1860 **f.** Mean receptor occupancies at baseline predicted by the model ($n = 45$ control group, $n = 44$
1861 lesion group). There is a significant increase in occupancies for both the D1R and D2R in the
1862 lesion compared to the control group (U-statistic = -2.1664 , $P = 0.0151$, U-statistic = -2.1328 , $P =$
1863 0.0165 for D1R and D2R respectively, single-sided Mann-Whitney U-test).
- 1864 **g.** Schematic showing the model predicted changes in dopamine concentrations and receptor
1865 occupancies for the control (black) and lesion (red) groups. The arrows depict the increase or
1866 decrease in occupancy for a positive or negative dopamine transient of a fixed magnitude.
- 1867 **h.** Changes in receptor occupancy as a function of dopamine transients predicted by the model.
1868 The slope for the positive and negative domains correspond to the receptor sensitivities of D1R
1869 and D2R (α^+ , α^-), respectively.
- 1870 **i.** Asymmetric scaling factors derived from the receptors' sensitivities for the control and lesion
1871 groups (i.e., τ in model 1, $n = 45$ control group, $n = 44$ lesion group). There was a significant
1872 increase in the lesion group with respect to controls (U-statistic = -12.205 , $P < 1.0 \times 10^{-6}$, single-
1873 sided Mann-Whitney U-test). Note that the increase in the asymmetry was driven mainly due to
1874 decreases in D2R sensitivity (panel h).
- 1875 **j.** Value predictions at baseline in a TD learning model trained with the receptor sensitivities
1876 derived from the biophysical. There was a significant increase in the value predictors at baseline
1877 in the model using the lesion group's derived parameters with respect to control. controls (t -
1878 statistic = -6.417 , $P < 1.0 \times 10^{-6}$, t -test).
- 1879 **k.** Value predictions at convergence of a TD learning model trained using the assumptions of
1880 Model 1 and the asymmetric scaling factors derived from receptors' sensitivities predicted by the
1881 biophysical model. The model led to a significant increase in the value predictions for all cues
1882 (U-statistic = -4.690 , $P < 1.0 \times 10^{-4}$, U-statistic = -4.734 , $P < 1.0 \times 10^{-4}$, U-statistic = -4.602 , $P <$

1883 1.0×10^{-4} , single-sided Mann-Whitney U-test, for the 10%, 50% and 90% reward probability
1884 cues) and an optimistic bias in the normalized value prediction to the 50% reward probability cue
1885 (t -statistic = -5.576 , $P < 1.0 \times 10^{-4}$, t -test) in accordance with the anticipatory licking observed in
1886 the data.

1887 **I. Predicted cue responses.** There is an overall decrease in RPEs in lesioned animals (left) due to
1888 an increase in the baseline (pre-cue) value prediction (U-statistic = 4.932 , $P < 1.0 \times 10^{-5}$, U-
1889 statistic = -3.658 , $P = 0.00025$, U-statistic = 4.734 , $P < 1.0 \times 10^{-4}$, single-sided Mann-Whitney
1890 U-test for the 10%, 50% and 90% reward probability cues), which is consistent with the
1891 decreases in the absolute magnitudes of dopamine cue-evoked responses in the lesion group (Fig.
1892 4c). The normalized TD errors at for the 50% reward probability cue show signs of an optimistic
1893 bias (U-statistic = -4.624 , $P < 1.0 \times 10^{-4}$, single-sided Mann-Whitney U-test).

1894 Centre of box plot shows the median; edges are 25th and 75th percentiles; and whiskers are the
1895 most extreme data points not considered as outliers.

1896



1897

1898 **Figure 7 | Model 1 predicts asymmetric learning rates in healthy humans given inter-**
 1899 **individual differences in baseline dopamine, and in Parkinson's disease patients given**
 1900 **inter-individual differences in depressive-like symptoms.**

1901 **a.** Schematic of the events occurring at dopaminergic axon terminal. Pre- and post-synaptic sites
 1902 predominantly express D2s (short) and D2l (long) subtypes, respectively.

1903 **b.** “Relative reversal learning (RRL)” under placebo conditions as a function of dopamine striatal
 1904 synthesis capacity measured with PET radio imaging (black dots, left y-axis, bottom x-axis).

1905 Figure taken from Cools et al. (2009)²⁴. Positive values of RRL indicate a bias favoring learning
1906 from gains relative to losses, and vice versa for negative values of RRL. There was a positive
1907 relationship between RRL and dopamine synthesis capacity. Model 1 predictions of RRL ($2\tau -$
1908 1 in Model 1) as a function of baseline dopamine using the receptors occupancy curve,
1909 recapitulate the positive relationship shown in the results from Cools et al. (2009)²⁴ (gray line,
1910 right y-axis, top x-axis).

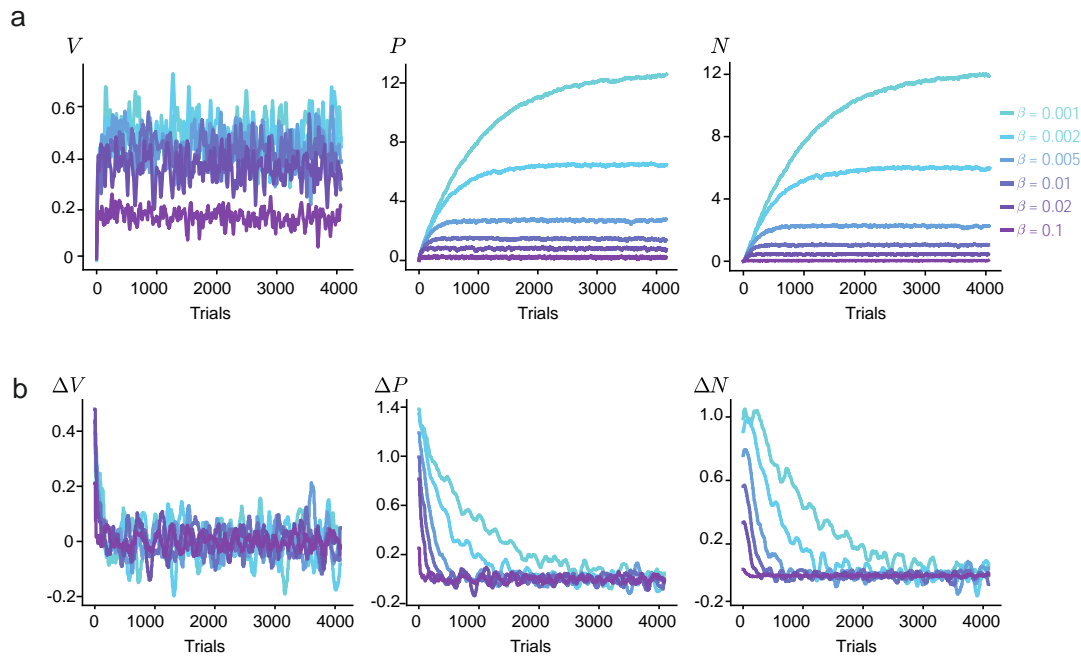
1911 **c.** The change in RRL induced by bromocriptine was negatively correlated with striatal
1912 dopamine synthesis capacity. Figure from Cools et al. (2009)²⁴ (black dots, left y-axis, bottom x-
1913 axis). Model 1 recapitulates qualitatively the effect of bromocriptine in RRL, equivalent to
1914 $\Delta(2\tau - 1)$. The solid light green line represents the $\Delta(2\tau - 1)$ when considering bromocriptine's
1915 effect on *D2l occupancy only*; the dashed line represents the $\Delta(2\tau - 1)$ when both *D2l and D2s*
1916 *occupancy* was considered; and the dark green line represents the $\Delta(2\tau - 1)$ when both *D2l and*
1917 *D2s activation* was considered (this includes the fact that bromocriptine is a partial agonist for
1918 the D2l and D2s receptors). The curves were obtained by imposing a concentration of $10^{0.8}$ nM
1919 of bromocriptine in the biophysical model.

1920 **d.** Receptor occupancy curves for the D2l receptors at baseline (grey line) and in the presence of
1921 $10^{0.8}$ nM of bromocriptine: Solid light green line corresponds to considering bromocriptine's
1922 effects on D2l receptors occupancy alone; dashed line, corresponds to considering
1923 bromocriptine's effects on both D2l and D2s receptors; solid dark green line corresponds to
1924 considering bromocriptine's effect on the activation curves of both D2l and D2s receptors. The
1925 binding of the drug to the D2l receptors alone causes an increase in occupancy. This happens to a
1926 larger extent when starting from a low dopamine level at baseline than in high dopamine levels.
1927 The binding of the drug to D2s receptors in addition to D2l receptors causes a rightward shift in
1928 the curves. The activation levels are lower than 1 even at the drug levels where occupancy is
1929 close to 1, due to the lower efficiency of bromocriptine in receptor activation (Methos 4.1). See
1930 Extended Figure 6 and 7 for the effect of changing bromocriptine's concentration and efficiency
1931 of activation.

1932 **e.** Same as in panel **d.** but now reporting $(2\tau - 1)$ calculated from the D2l receptor's occupancy
1933 and activation curves. An increase in $2\tau - 1$ happens to a larger extent when starting from a low
1934 dopamine level than from high dopamine level.

1935 **f.** Same as in panel **d.** but now reporting $\Delta(2\tau - 1)$ calculated from the D2l receptor's occupancy
1936 and activation curves. Model 1 recapitulates qualitatively the effect of bromocriptine on RRL.
1937 **g.** The effect of PD medication (L-DOPA) on loss aversion is negatively correlated with their
1938 off-medication depression score. Figure from Timmer et al. (2018)⁷¹.
1939 **h.** Model 1 recapitulates qualitatively the effect of PD medication in loss aversion. We assumed
1940 that the asymmetry in favor of learning from losses relative to gains $(1 - \tau)_{off}$ scales with the
1941 baseline dopamine levels. Given this, we derived a distribution of off-medication baseline
1942 dopamine levels centered around the mean $(1 - \tau)_{off}$ derived from the data of Timmer et al.
1943 (2018)⁷¹ (see methods 0). We then imposed a fixed increase in baseline dopamine to simulate L-
1944 DOPA effects. We derived the new loss-aversion parameter $(1 - \tau)_{on}$ at the shifted baseline
1945 dopamine levels. The y-axis shows the change in loss aversion for each sample of the
1946 distribution of baseline dopamine levels. If the off-medication depression score is correlated with
1947 $(1 - \tau)_{off}$ then model would predict the result in Timmer et al. (2018)⁷¹.
1948 PD: Parkinson's disease.
1949

1950 Extended data Figures



1951

1952 **Extended Data Fig. 1 | Variables of model 1 show convergence irrespective of the value of**
1953 **the decay factor.**

1954 **a.** Value predictors V (left), P population (middle) and N population (right) across trials of
1955 training for an RL agent of model 1. Color of lines denotes the value of the decay factor (β) in
1956 the update rules for the P and N populations. Colormap is the same for all figures (left). All the
1957 model variables show convergence for every value of the decay factor β .

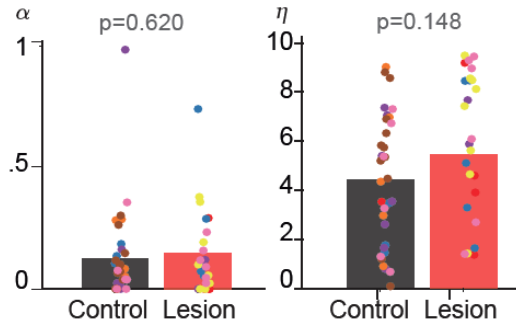
1958 **b.** Difference in the variables estimates between consecutive trials of training, for the value
1959 predictors V (left, ΔV), P population (middle, ΔP) and N population (right, ΔN). All the variables
1960 show convergence for every value of the decay factor β (shown as a ΔV , ΔP , ΔN equal to zero).

a TD learning

$$\delta = r - V$$

$$V \leftarrow V + \alpha \cdot \delta$$

$$\text{Licking} = \eta \cdot V$$

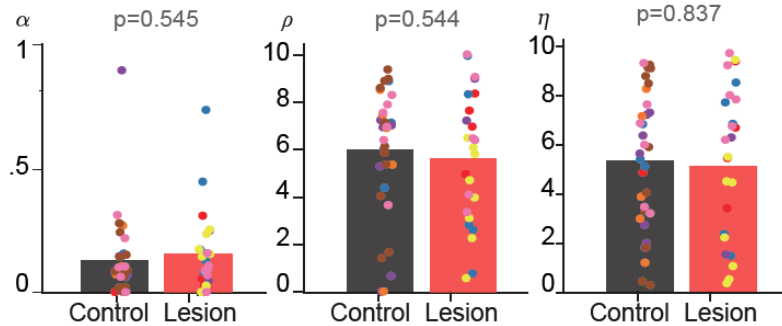


b TD learning with reward sensitivity

$$\delta = \rho \cdot r - V$$

$$V \leftarrow V + \alpha \cdot \delta$$

$$\text{Licking} = \eta \cdot V$$



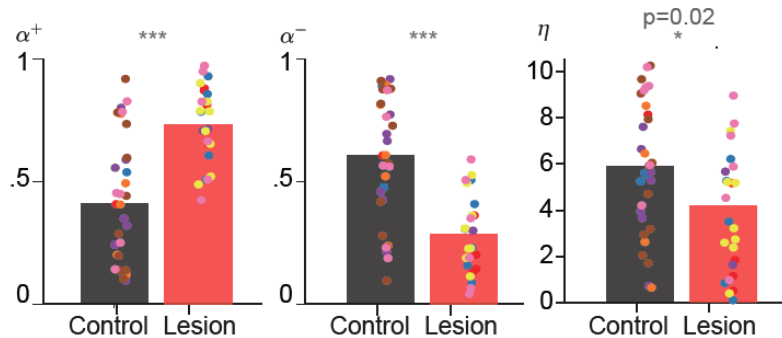
c Risk sensitive reinforcement learning

$$\delta = r - V$$

$$V \leftarrow V + \alpha^+ \cdot \delta \quad \text{if } \delta > 0$$

$$V \leftarrow V + \alpha^- \cdot \delta \quad \text{if } \delta < 0$$

$$\text{Licking} = \eta \cdot V$$



1961

1962

1963 **Extended Data Fig. 2 | RL model fits to the trial-by-trial anticipatory licking responses.**

1964 **a.** TD learning fits reveal no significant difference across groups in the learning rate (left, U-
 1965 statistic = -4.954 , $P = 0.620$, two-sided Mann-Whitney U-test) nor in the linear mapping between
 1966 value predictions and anticipatory licking responses (U-statistic = -1.445 , $P = 0.148$, two-sided
 1967 Mann-Whitney U-test).

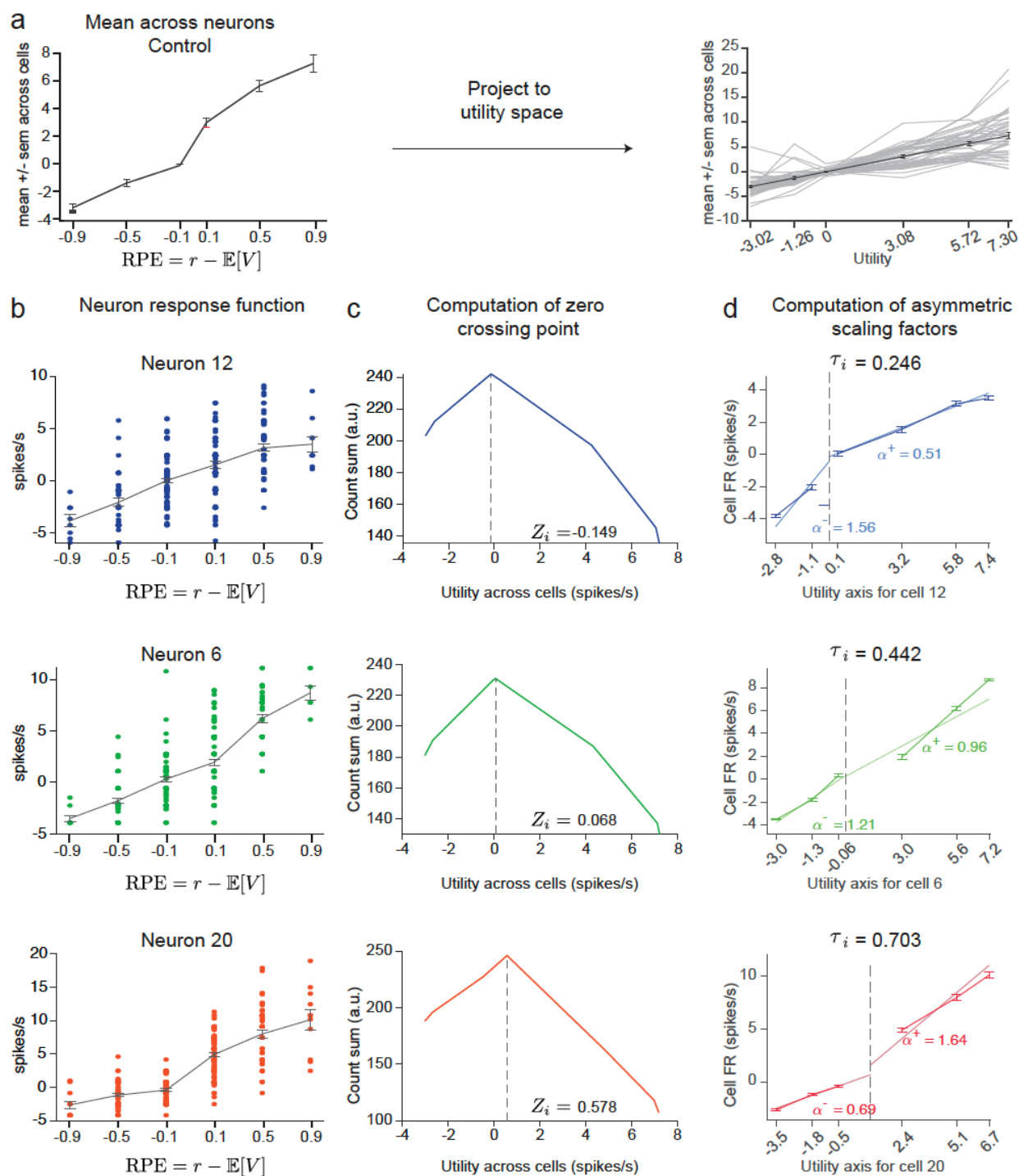
1968 **b.** Model fits of TD learning with reward sensitivity reveal no difference across groups in the
 1969 learning rate (left, U-statistic = 0.206 , $P = 0.836$, two-sided Mann-Whitney U-test) nor in the

1970 linear mapping between value predictions and anticipatory licking responses (middle, U-statistic
1971 = -0.7844 , $P = 0.4327$, two-sided Mann-Whitney U-test), nor in the reward sensitivity (right) (U-
1972 statistic 0.545 , $P = 0.605$, two-sided Mann-Whitney U-test).

1973 **c.** Model fits of TD learning with asymmetric learning rates for positive vs negative RPEs. This
1974 model reveals a significant difference across groups in the asymmetry between α^+ and α^- (U-
1975 statistic = -4.678 , $P < 1.0 \times 10^{-5}$, two-sided Mann-Whitney U-test) and a small but significant
1976 difference between the linear mapping between value predictions and anticipatory licking
1977 responses (right, U-statistic = 2.33 , $P = 0.02$, two-sided Mann-Whitney U-test).

1978

1979



1980

1981 **Extended Data Fig. 3 | Signatures of distributional reinforcement learning model are**
 1982 **preserved after habenula lesions.**

1983 **a.** RPE -evoked responses at outcome as a function of the theoretical RPE for each trial type. The
 1984 figure shows the average response function across neurons from the control group. The

1985 computation of zero-crossing points and asymmetric scaling factors is carried out in the ‘utility
1986 space’(the average response function) as in ³⁶ to account for response nonlinearities.

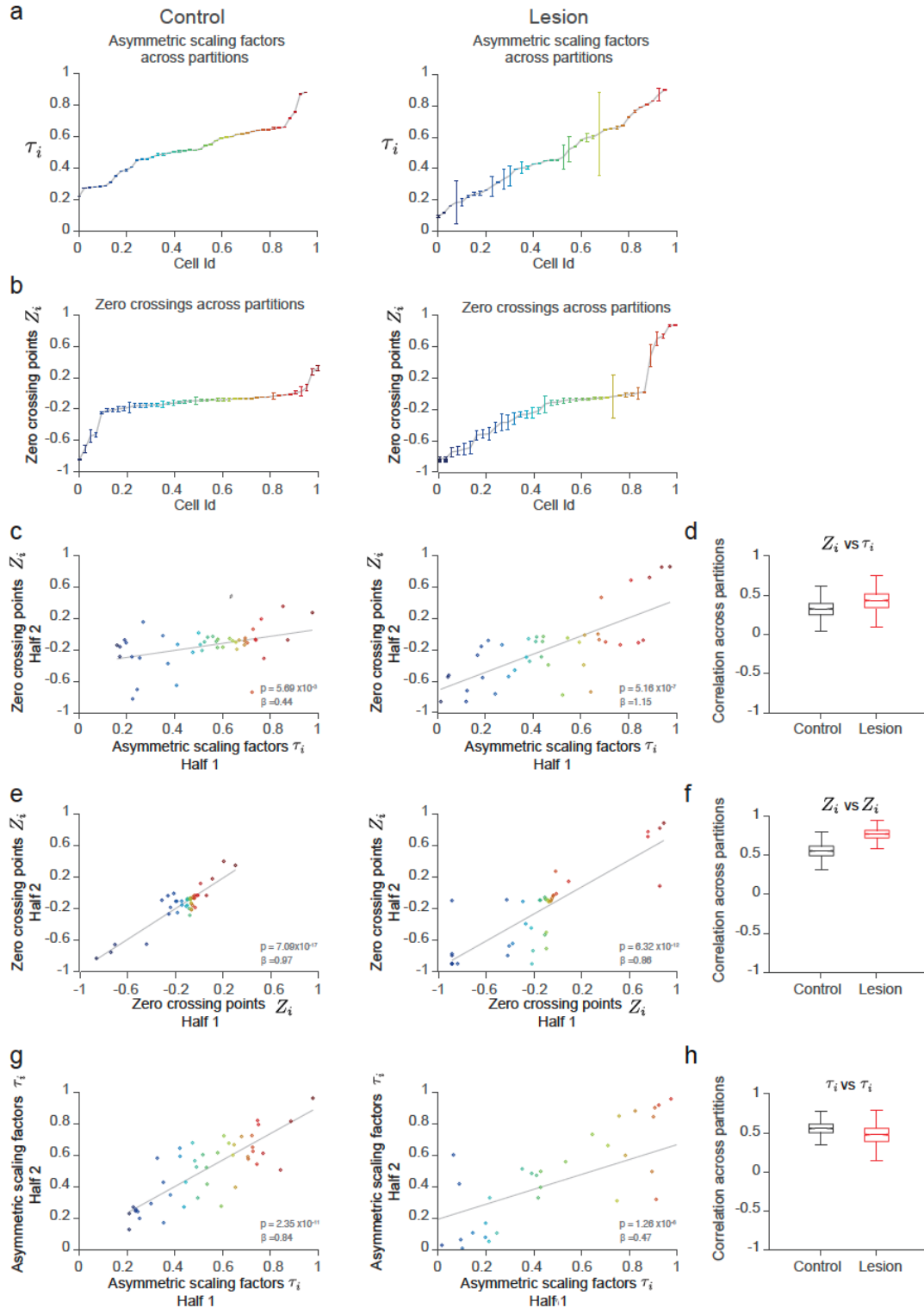
1987 **b.** Example of the response function of 3 dopamine neurons from the control group ordered by
1988 their asymmetric scaling factors: pessimistic, neutral and optimistic, from top to bottom.

1989 **c.** Computation of zero-crossing points for the neurons in B. The reversal points for each cell (Z_i)
1990 were defined as the point in utility space that maximized the number of positive responses to
1991 points greater than Z_i plus the number of negative responses to points less than Z_i . The y-axis
1992 shows the sum of responses below and above each point in the utility space. The zero-crossing
1993 point is shown as the maxima in this function.

1994 **d.** Computation of asymmetric scaling factors for the neurons in c. Here, the responses functions
1995 in b have been projected to the utility space in A and realigned according to their zero-crossing
1996 points. The asymmetric learning rates (α^+ , α^-) are taken to be the slopes of these response
1997 functions.

1998

1999



2000

2001 **Extended Data Fig. 4 | Distributional reinforcement learning variables from the Habenula**

2002 **lesion dataset.**

2003 **a.** Distribution of asymmetric scaling factors for the dopamine neurons from the control (left)
2004 and lesion (right) groups. The error bars were derived by randomly sampling trials to compute
2005 the asymmetric scaling factors for 1000 iterations.

2006 **b.** Distribution of zero crossing points for the control (left) and lesion(right) groups. The error
2007 bars were derived as in a.

2008 **c.** Correlation of asymmetric scaling factors (x-axis) and zero-crossing points (y-axis) computed
2009 on disjoint halves of trials for an example partition.

2010 **d.** Distribution of correlation coefficients between asymmetric scaling factors (x-axis) and zero-
2011 crossing points (y-axis) across disjoint halves of trials for 1000 partitions for the control and
2012 lesion groups.

2013 **e.** Correlation between zero-crossing points computed on disjoint halves of trials for an example
2014 partition.

2015 **f.** Distribution of correlation coefficients between zero-crossing points computed on disjoint
2016 halves of trials for 1000 partitions for the control and lesion groups.

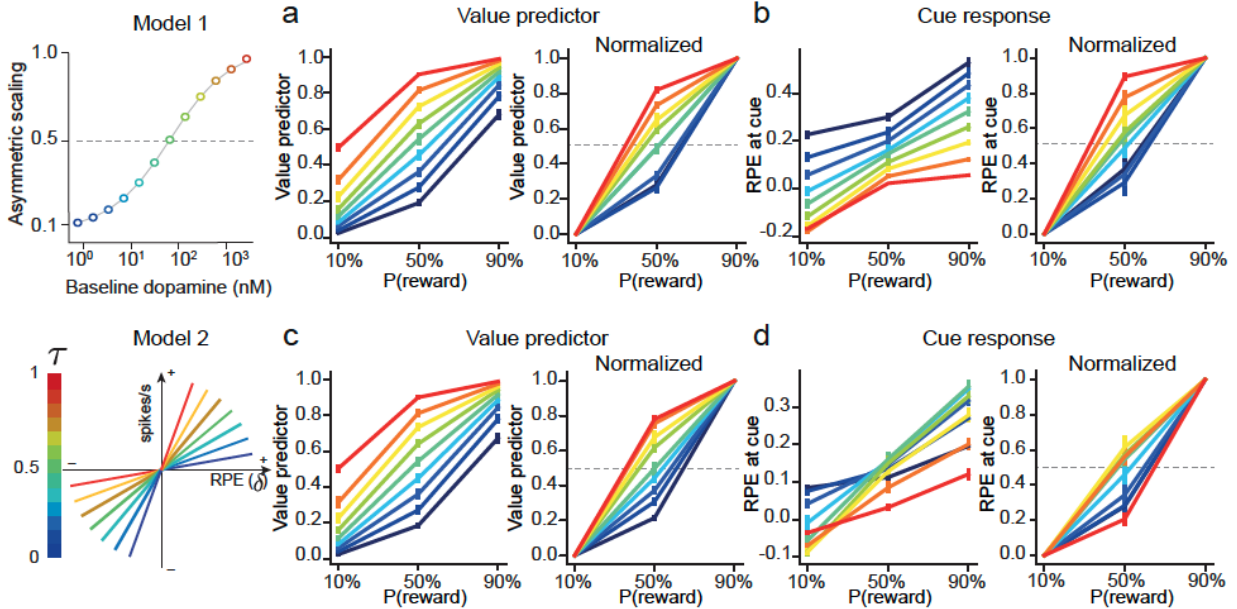
2017 **g.** Correlation between asymmetric scaling factors computed on disjoint halves of trials for an
2018 example partition.

2019 **h.** Distribution of correlation coefficients between asymmetric scaling factors computed on
2020 disjoint halves of trials for 1000 partitions for the control and lesion groups.

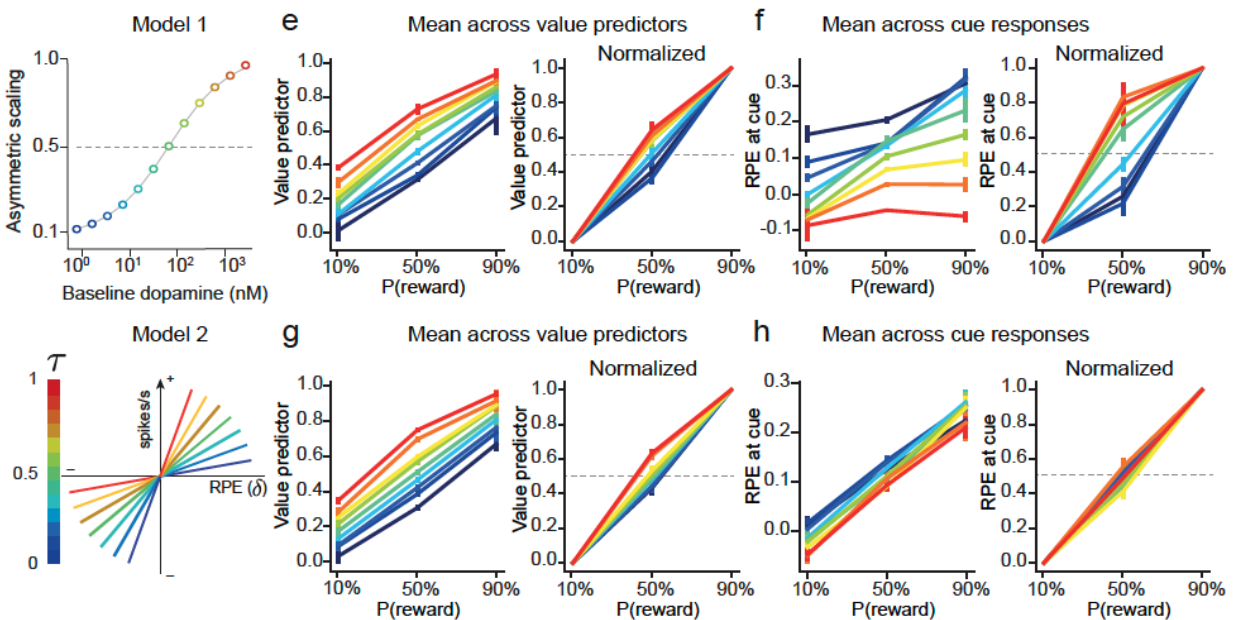
2021

2022

TD learning



Distributional TD learning

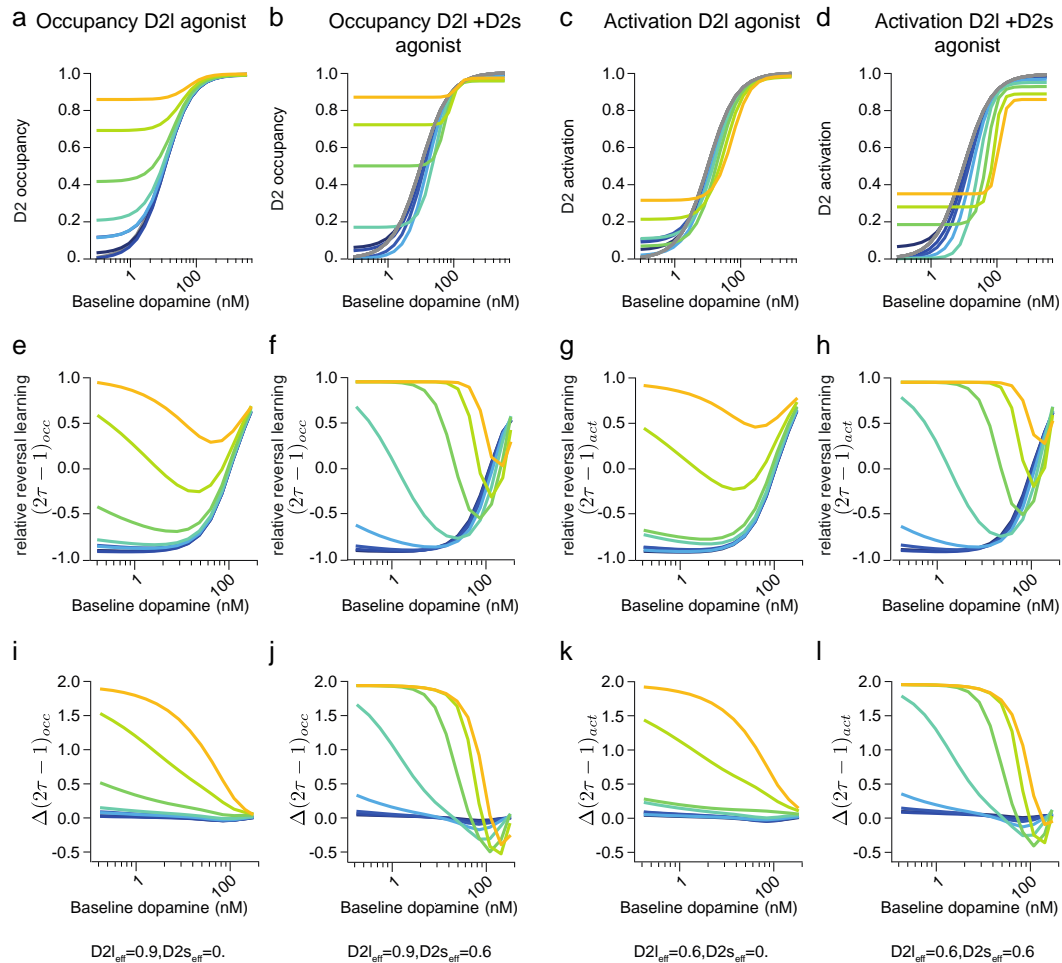


2023

2024 **Extended Data Fig. 5 | Biases in cue-evoked responses in the Habenula lesion data cannot**
 2025 **be explained by asymmetric scaling of RPEs (Model 2).**

2026 a. Value predictors derived from model 1 with TD learning for a set of baseline dopamine levels
 2027 (colormap). The optimistic and pessimistic biases are present.

- 2028 **b.** Cue responses derived from model 1 with TD learning for a set of baseline dopamine levels
2029 (colormap). The optimistic and pessimistic biases are revealed when the responses are
2030 normalized.
- 2031 **c.** Value predictors derived from model 2 with TD learning for a set of baseline dopamine levels
2032 (colormap). The optimistic and pessimistic biases are present.
- 2033 **d.** Cue responses derived from model 2 with TD learning for a set of baseline dopamine levels
2034 (colormap). The optimistic and pessimistic biases are absent in both the normalized and the raw
2035 TD errors.
- 2036 **e.** Mean across the distribution of value predictors derived from model 1 with distributional TD
2037 learning for a set of baseline dopamine levels (colormap). The optimistic and pessimistic biases
2038 are present.
- 2039 **f.** Mean across the distribution of cue responses derived from model 1 with distributional TD
2040 learning for a set of baseline dopamine levels (colormap). The optimistic and pessimistic biases
2041 are revealed when the responses are normalized.
- 2042 **g.** Mean across the distribution of value predictors derived from model 2 with distributional TD
2043 learning for a set of baseline dopamine levels (colormap). The optimistic and pessimistic biases
2044 are present.
- 2045 **h.** Mean across the distribution of cue responses derived from model 2 with distributional TD
2046 learning for a set of baseline dopamine levels (colormap). The optimistic and pessimistic biases
2047 are absent in both the normalized and the raw TD errors.
- 2048
- 2049



2050

2051 **Extended Data Fig. 6 | Model 1 predicts asymmetric learning rates and the effect of**
 2052 **bromocriptine in healthy humans given inter-individual differences in baseline dopamine.**

2053 **a.** Occupancy curves for the D2l receptors at baseline (grey line) and when considering
 2054 bromocriptine's effects in D2l receptors alone. The binding of the drug to the D2l receptors alone
 2055 causes an increase in the occupancy. This happens to a larger extent when starting from a low
 2056 dopamine level at baseline than in high dopamine levels.

2057 **b.** Occupancy curves for the D2l receptors at baseline (grey line) and when considering
 2058 bromocriptine's effects in both D2l and D2s receptors. The binding of the drug to D2s receptors
 2059 causes a rightwards shifts in the curves.

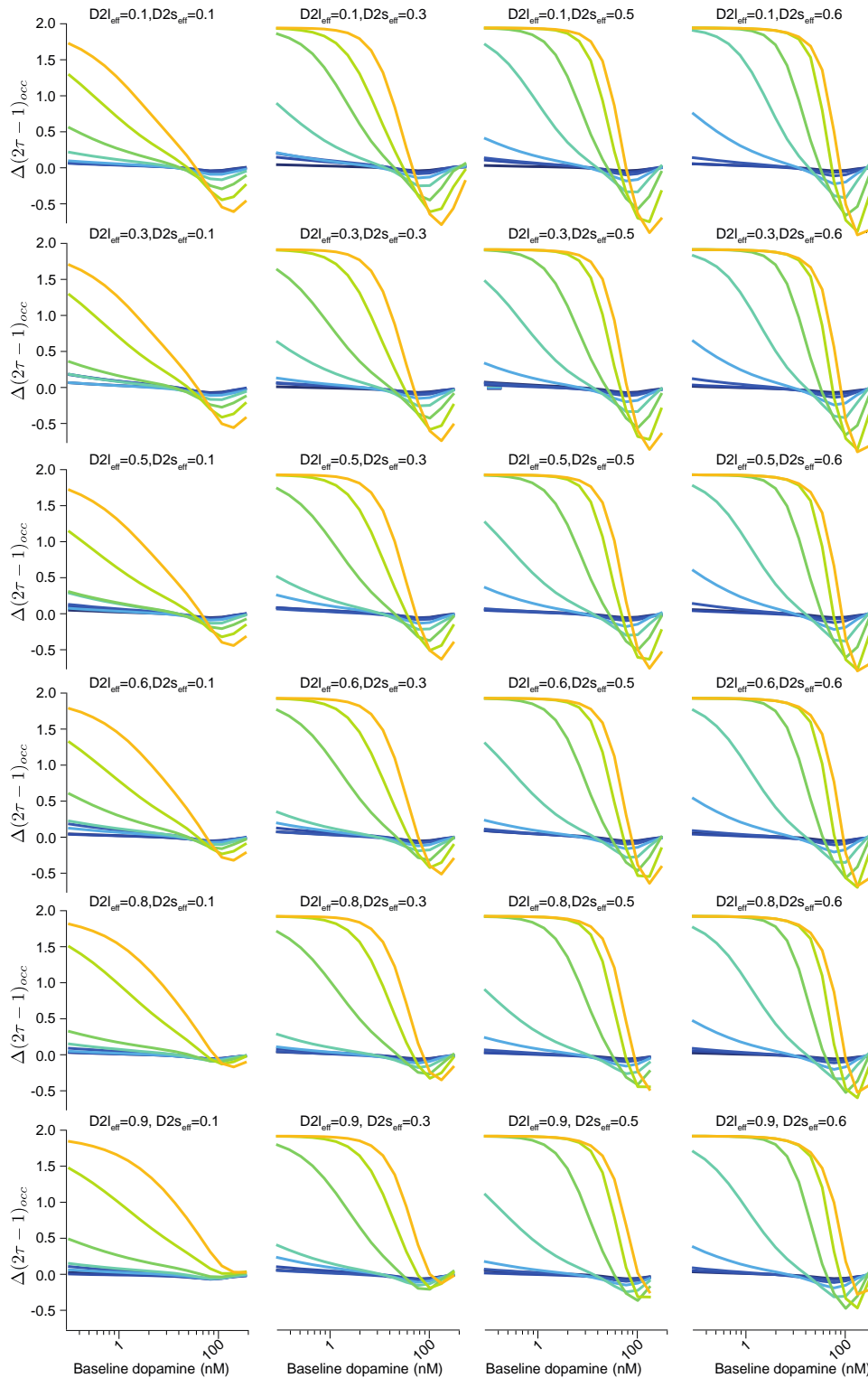
- 2060 **c.** Activation curves for the D2l receptors at baseline (grey line) and when considering
2061 bromocriptine's effects in D2l receptors alone, including the *partial* quality of the agonism of
2062 this drug on the receptor (i.e., efficiency <1).
- 2063 **d.** Activation curves for the D2l receptors at baseline (grey line) and when considering
2064 bromocriptine's effects in both D2l and D2s receptors, including the *partial* quality of the
2065 agonism.
- 2066 **e.** Relative reversal learning (RRL) calculated as $2\tau - 1$ in model 1, as a function of baseline
2067 dopamine (x-axis) and drug concentration (color) using the receptors occupancy curve,
2068 considering only bromocriptine's effect in D2l receptors.
- 2069 **f.** Relative reversal learning (RRL) calculated as $2\tau - 1$ in model 1, as a function of baseline
2070 dopamine (x-axis) and drug concentration (color) using the receptors occupancy curve,
2071 considering bromocriptine's effect in both D2l and D2s receptors.
- 2072 **g.** $2\tau - 1$ in model 1, as a function of baseline dopamine (x-axis) and drug concentration (color)
2073 using the receptors activation curve, considering only bromocriptine's effect in D2l receptors.
- 2074 **h.** $2\tau - 1$ in model 1, as a function of baseline dopamine (x-axis) and drug concentration (color)
2075 using the receptors activation curve, considering bromocriptine's effect in both D2l and D2s
2076 receptors.
- 2077 **i.** The change in $2\tau - 1$ induced by the drug at different concentrations (color) with respect to
2078 the baseline condition, as a function of baseline dopamine (x-axis). The curves represent the
2079 change when calculating $2\tau - 1$ from the occupancy curves considering only D2l binding.
- 2080 **j.** The change in $2\tau - 1$ induced by the drug at different concentrations (color) with respect to
2081 the baseline condition, as a function of baseline dopamine (x-axis). The curves represent the
2082 change when calculating $2\tau - 1$ from the occupancy curves considering both D2l and D2s
2083 binding
- 2084 **k.** The change in $2\tau - 1$ induced by the drug at different concentrations (color) with respect to
2085 the baseline condition, as a function of baseline dopamine (x-axis). The curves represent the
2086 change when calculating $2\tau - 1$ from the activation curves considering only D2l activation.
- 2087 **l.** The change in $2\tau - 1$ induced by the drug at different concentrations (color) with respect to
2088 the baseline condition, as a function of baseline dopamine (x-axis). The curves represent the

2089 change when calculating $2\tau - 1$ from the activation curves considering both D2l and D2s
2090 activation.

2091 The parameters of efficiency of activation of D2 receptors by the drug ($D2l_{eff}, D2s_{eff}$) used in
2092 each column are reported at the bottom of the figure.

2093

2094



2095

2096 **Extended Data Fig. 7 | Robustness of the effect of bromocriptine in the relative reversal**

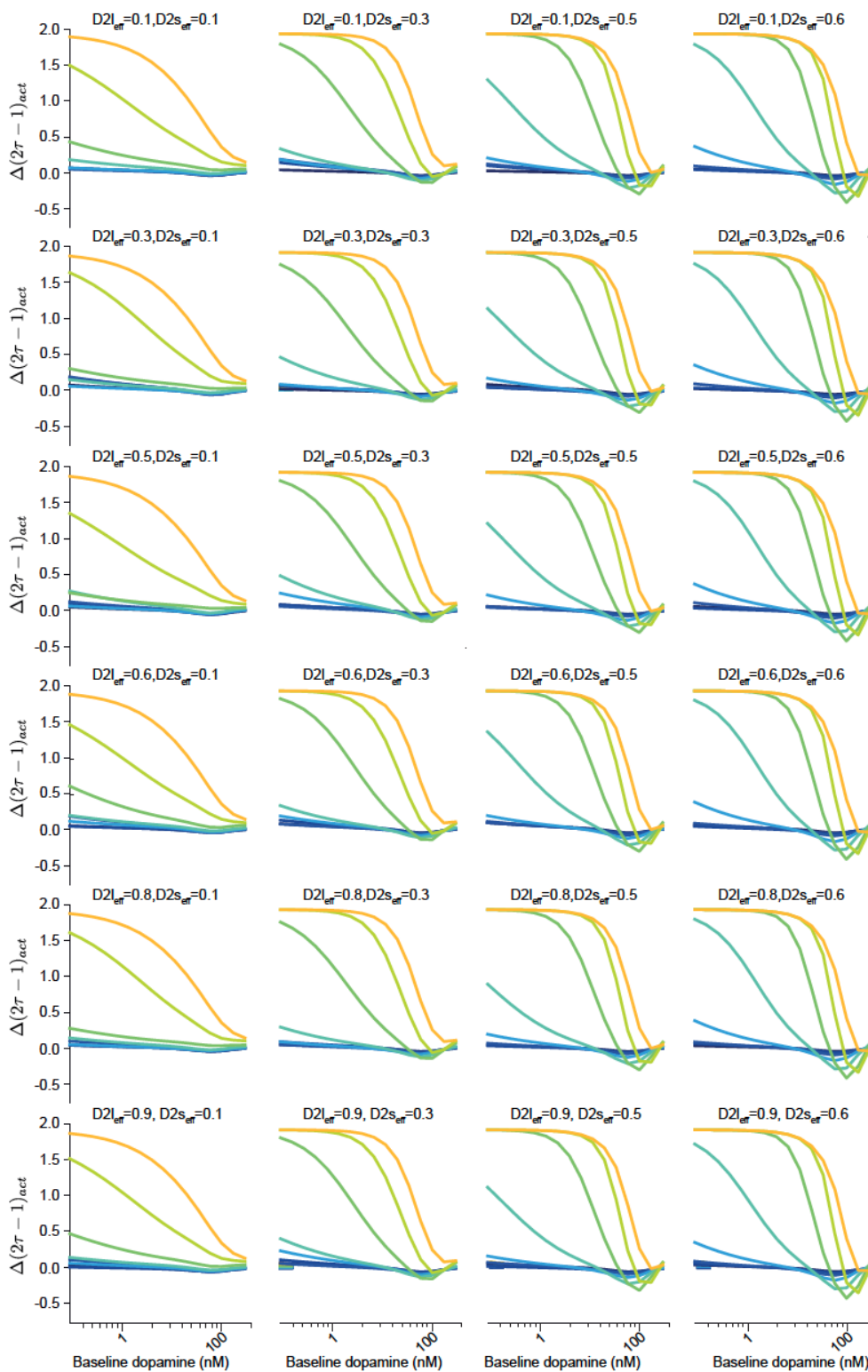
2097 **learning calculated from the occupancy curves to the choice of the drug efficiency**

2098 **parameter.**

2099 The qualitative effects on bromocriptine in the change in relative reversal learning $\Delta(2\tau - 1)$
2100 calculated from the D2 *occupancy* curves. Results hold regardless of the choice of the efficiency
2101 of the drug on D2l (rows) or D2s (columns) efficiency.

2102

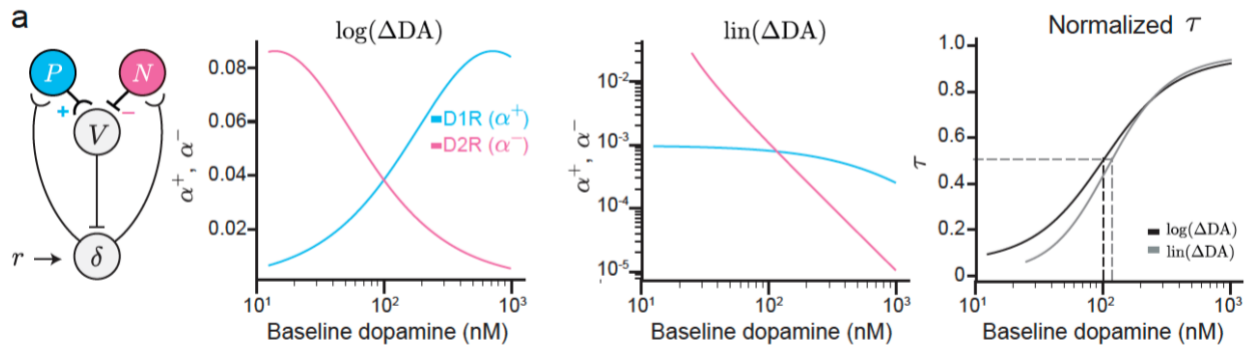
2103



2104

2105 **Extended Data Fig. 8 | Robustness of the effect of bromocriptine in the relative reversal**
2106 **learning calculated from the activation curves to the choice of the drug efficiency**
2107 **parameter.**

2108 The qualitative effects on bromocriptine in the change in relative reversal learning $\Delta(2\tau - 1)$
2109 calculated from the D2 *activation* curves. Results hold regardless of the choice of the efficiency
2110 of the drug on D2l (rows) or D2s (columns) efficiency.
2111



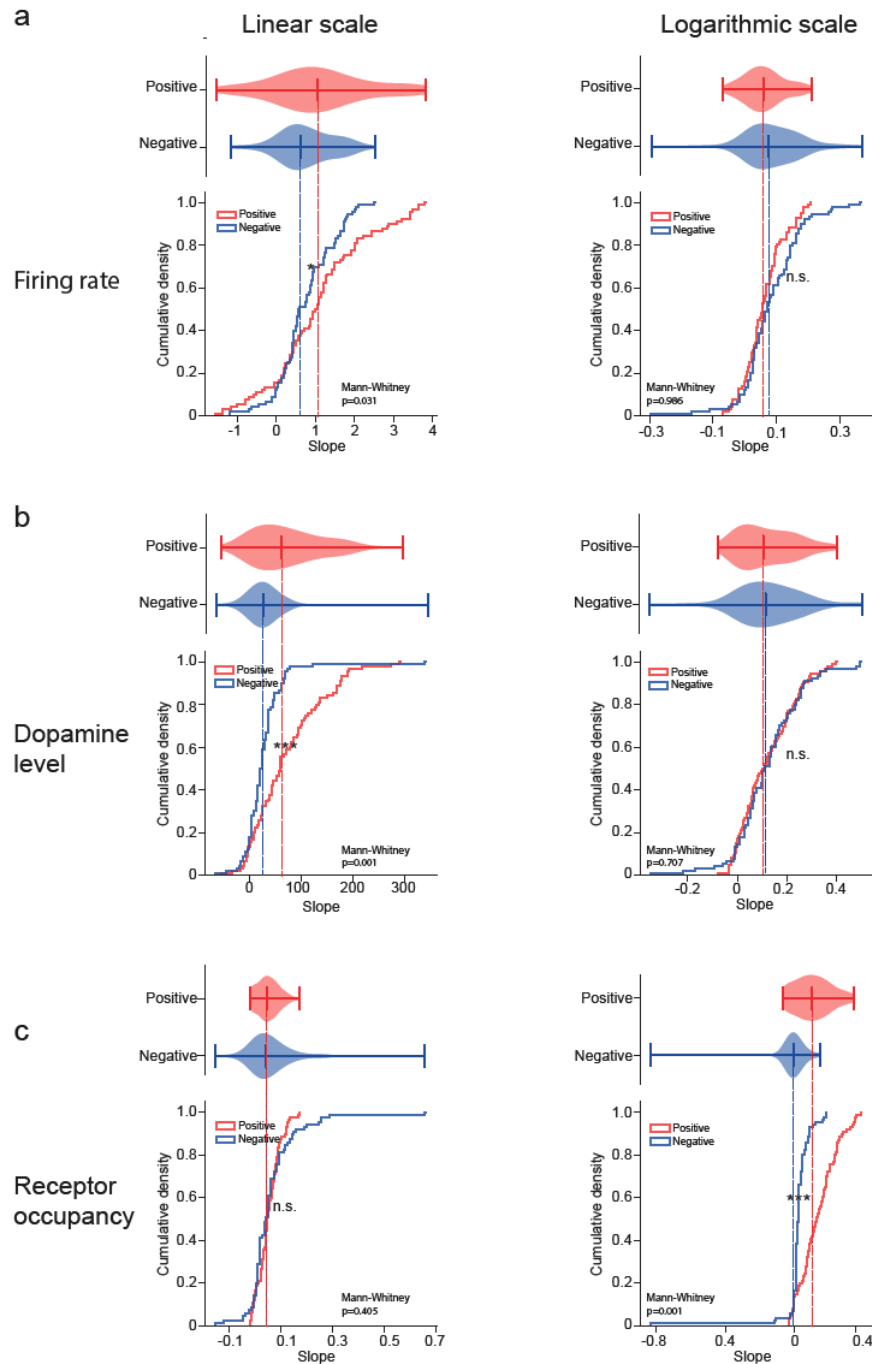
2112

2113

2114 **Extended Data Fig. 9 | The qualitative aspects of Model 1 are preserved irrespective of the**
2115 **assumption made about the changes in baseline dopamine caused by dopamine transients.**

2116 **a.** Computation of receptor sensitivities (i.e., slope of dose-occupancy curves, α^+ , α^-) assuming
2117 logarithmic (left) or linear (middle) changes in baseline dopamine induced by dopamine
2118 transients ($\log\Delta DA$, $\ln\Delta DA$, respectively). The absolute magnitude of the slopes differs
2119 depending on the assumption made about the changes in baseline dopamine (logarithmic vs
2120 linear) but the asymmetric scaling factor presents only a small shift in the curve as a function of
2121 baseline dopamine (right). The qualitative aspects of the model (i.e., non-monotonic relationship
2122 of the asymmetric scaling factor with baseline dopamine) is preserved regardless on this
2123 assumption.

2124



2125

2126 **Extended Data Fig. 10 | Change in dopamine firing rates, dopamine concentration and**
2127 **receptor occupancy as a function of RPEs in the linear scale or logarithmic scale.**

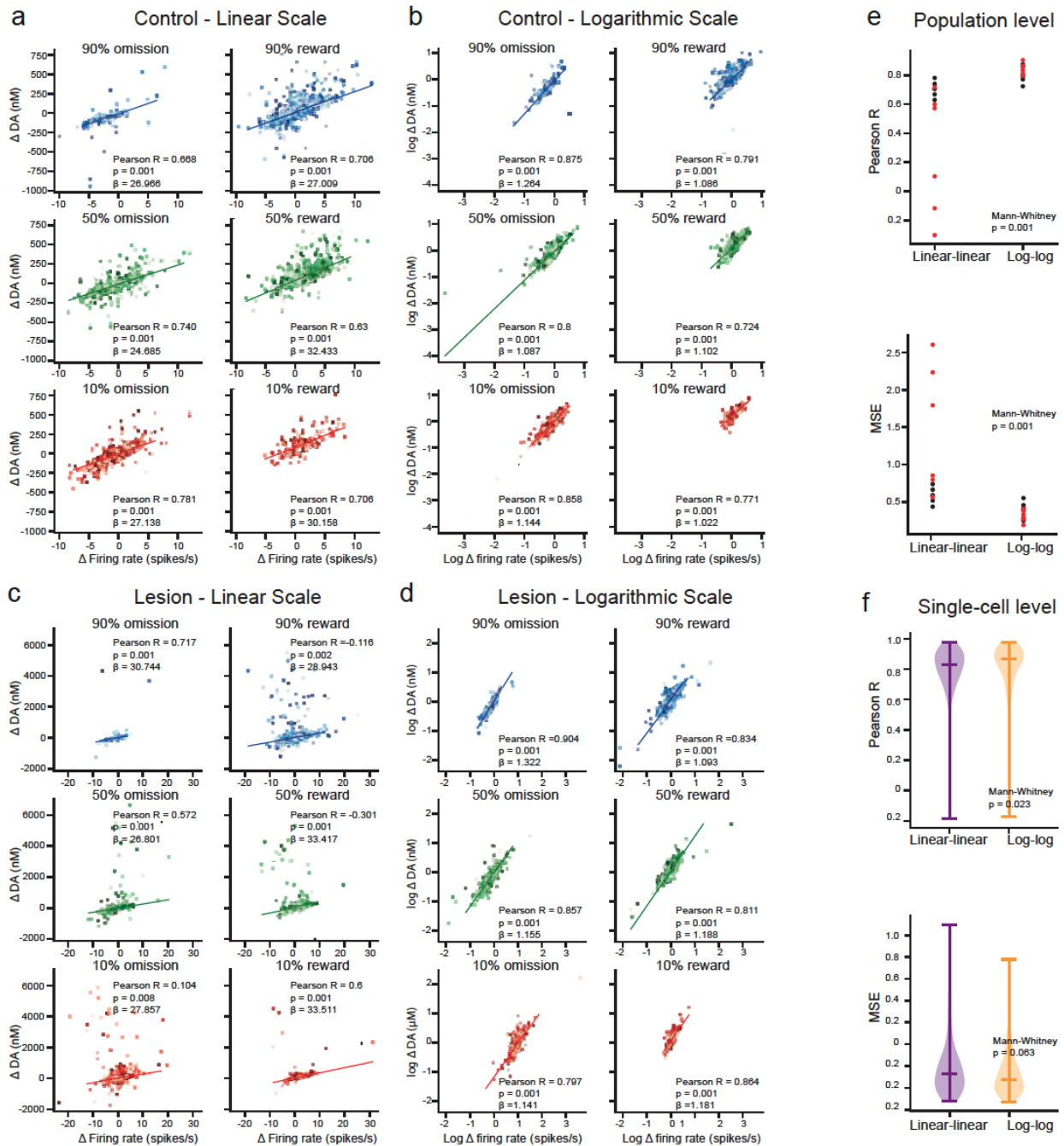
2128 **a.** Distributions of the slopes of the change in firing rate derived as a function of RPEs in the
2129 positive and negative domains, computed in the linear (left) or logarithmic (right) scale,

2130 calculated at a single neuron level. The slopes are asymmetric if considered in the linear scale,
2131 with the negative transients presenting a shallower slope than the positive ones. The slopes are
2132 symmetric if considered in the logarithmic scale.

2133 **b.** Distributions of the slopes of the change in dopamine levels derived from the biophysical
2134 model as a function of RPEs in the positive and negative domains, computed in the linear (left)
2135 or logarithmic (right) scale, calculated at a single neuron level. The slopes are again asymmetric
2136 if considered in the linear scale but symmetric if considered in the logarithmic scale.

2137 **c.** Slope of the change in receptor occupancy derived from the biophysical model for a given
2138 RPE in the positive and negative domains, computed in the linear (left) or logarithmic (right)
2139 scale, calculated at a single neuron level. The slopes are symmetric if considered in the linear
2140 scale but asymmetric if considered in the logarithmic scale.

2141



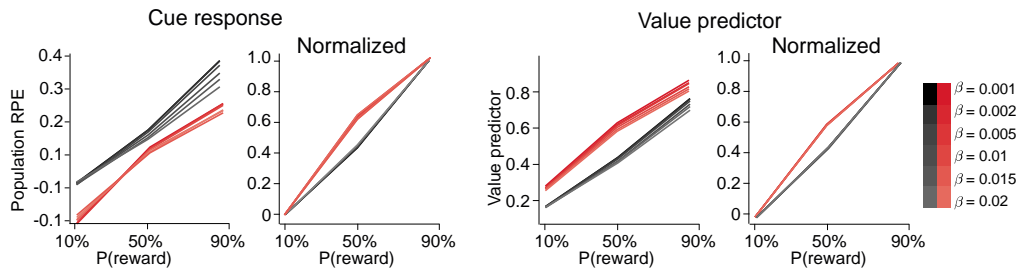
2142

2143 **Extended Data Fig. 11 | Relationship between changes in firing rates and changes in**
 2144 **dopamine concentration derived from the biophysical model.**

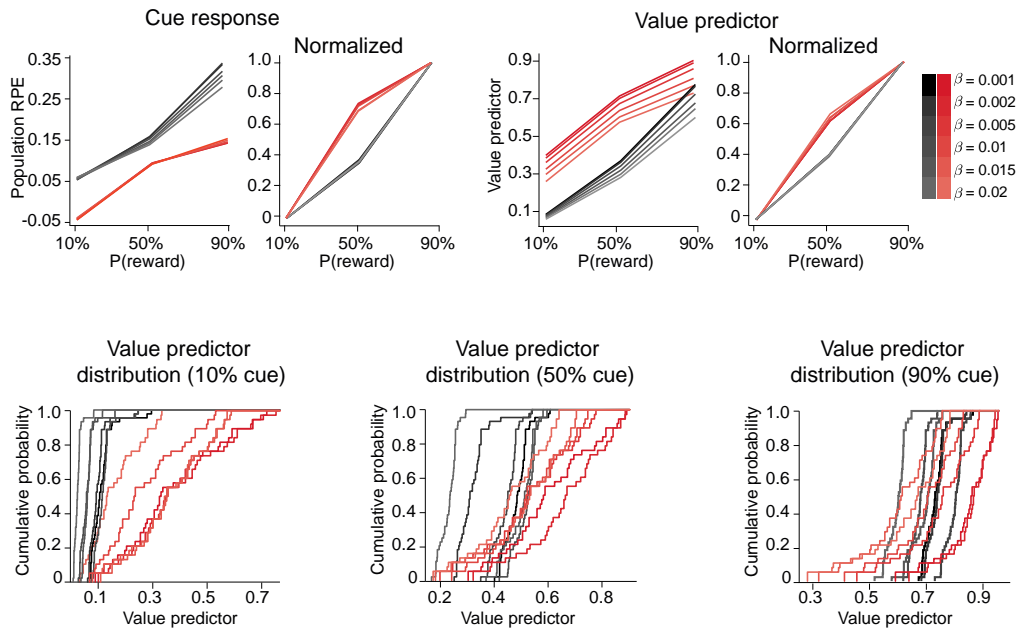
2145 **a.** Linear fits to the relationship between changes in firing rates and changes in dopamine
 2146 concentration evoked by the TD error at outcome in the linear scale for the control group. The
 2147 fits are done separately for each trial type.

- 2148 **b.** Same as a, but fits are done in the logarithmic scale.
- 2149 **c.** Same as a, but fits are done for the lesion group.
- 2150 **d.** Same as c, but fits are done in the logarithmic scale.
- 2151 **e.** Distribution of the Pearson correlation coefficients (top) and means squared error (MSE,
2152 bottom) between the predicted change in dopamine concentration by the linear regression and the
2153 ground truth derived from the biophysical model. The coefficients are derived from the fits in
2154 figures a-e, done by pooling all trials for each trial type (each point each trial type, with black for
2155 control and red for lesion group). There was a significant increase in the Pearson correlation
2156 coefficient and a near significant decrease in the MSE if the changes are considered to happen in
2157 the logarithmic scale.
- 2158 **f.** Same as E, but the linear regression fits are done for each neuron separately by pooling all
2159 trials. There was a significant increase in the single-cell distribution of Pearson correlation
2160 coefficients (top) and a significant decrease in the MSE distribution (bottom) if the changes are
2161 considered to happen in the logarithmic scale.
- 2162
- 2163

a Model 1 - No distributional TD



b Model 1 - Distributional TD



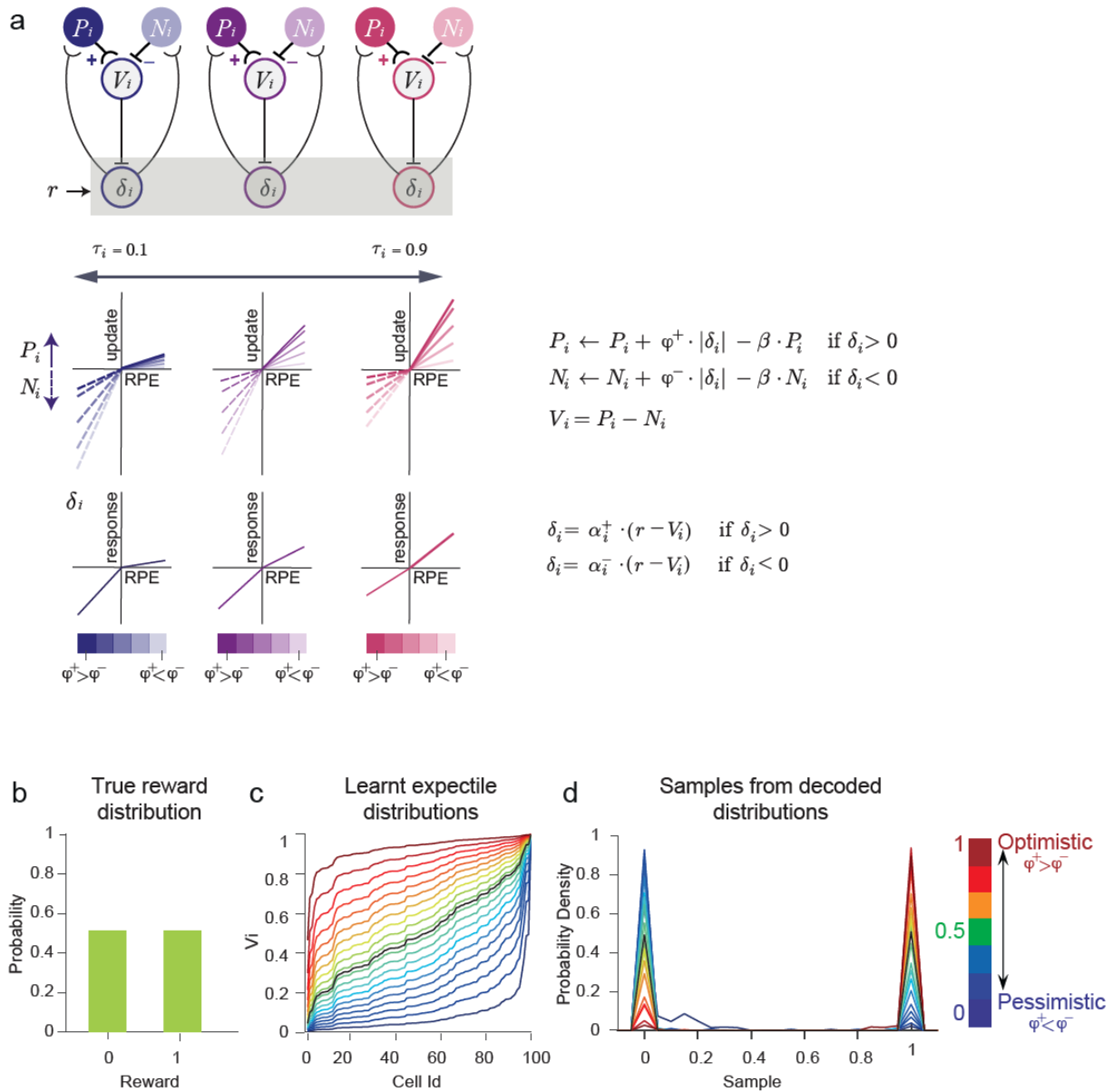
2164

2165 **Extended Data Fig. 12 | Model 1 captures signatures of the data irrespective of the choice of**
 2166 **the decay factor and is compatible with distributional RL.**

2167 **a.** Model 1 with standard TD learning. Simulations were run using the receptors sensitivities
 2168 from the biophysical models and data-derived asymmetric scaling factors (see Methods 3.3).
 2169 The model's predictions capture the signatures in cue-evoked dopamine responses (left) and
 2170 value predictions (right) irrespective of the choice of the decay factor (β).

2171 **b.** Model 1 within the distributional RL framework (see Methods 3.3). The model's predictions
 2172 also capture the signatures in cue-evoked dopamine responses (left) and value predictions (right)
 2173 irrespective of the choice of the decay factor (β). Bottom row shows the distribution of value
 2174 predictors for each reward-predictive cue.

2175



2176

2177 **Extended Data Fig. 13 | Distributional reinforcement learning with D1 and D2 populations**
 2178 **(Model 1).**

2179 **a.** Schematic of the distributional RL model with D1 and D2 populations. The schematic
 2180 represents three different value predictors (pessimistic, neutral and optimistic from left to right)
 2181 with their respective P and N neurons. The level of optimism of each individual value predictor
 2182 is determined by the scaling factors of the individual dopamine RPE-evoked responses (α_i^+ , α_i^- ,
 2183 represented by the color in the colormap from purple to pink) and allows the model to encode
 2184 information about the distribution of rewards (bottom). The global level of ‘optimism’ or

2185 ‘pessimism’ of the agent is given by the re-scaling of the RPEs by the P and N receptors
2186 sensitivities in the model (ϕ^+ , ϕ^- , represented with the color saturation).

2187 **b.** Example of a Bernoulli distribution, equivalent to the reward distribution predicted by the
2188 50% cue.

2189 **c.** Distribution of expectiles learnt by the distributional RL model with D1 and D2 population for
2190 the reward distribution in b. The expectiles are sorted based on the asymmetric scaling factor of
2191 each individual dopamine neuron. Colormap represents the level of optimistic or pessimism of
2192 each agent.

2193 **d.** Samples from the decoded distributions for the set of expectiles in c. The probability density is
2194 bimodal, in accordance with the distribution in b. As the agents goes from pessimism to
2195 optimism, the probability density modes change in their relative magnitude.

2196