

1 **Title:** Dopamine regulates the exploration-exploitation trade-off in rats

2

3 François Cinotti^{1*}, Virginie Fresno^{2,3*}, Nassim Akilil¹, Etienne Coutureau^{2,3}, Benoît Girard¹, Alain R. Marchand^{2,3§},

4 Mehdi Khamassi^{1§}

5

6 1. Institut des Systèmes Intelligents et de Robotique, Sorbonne Universités, UPMC Univ Paris 06, CNRS F-
7 75005, Paris, France

8 2. CNRS, Institut de Neurosciences Cognitives et Intégratives d'Aquitaine (INCLIA, UMR 5287), Bordeaux,
9 France,

10 3. Université de Bordeaux, INCLIA, Bordeaux, France

11 * equally contributing authors

12 § equally contributing senior authors

13

14 **Data availability statement:** The datasets generated and/or analysed during the current study are available from
15 the corresponding author on reasonable request.

16 **Corresponding author:** Mehdi Khamassi, mehdi.khamassi@upmc.fr, +33 1 44 27 28 85

17 **Keywords:** Dopamine, Exploration-exploitation dilemma, Decision-making, Reinforcement learning, Non-
18 stationary bandit task

19

20

21

22 **Abstract:**

23 In a volatile environment where rewards are uncertain, successful performance requires a delicate balance
24 between exploitation of the best option and exploration of alternative choices. It has theoretically been proposed
25 that dopamine controls this exploration-exploitation trade-off, specifically that the higher the level of tonic
26 dopamine, the more exploitation is favored. We demonstrate here that there is a formal relationship between the
27 rescaling of dopamine positive reward prediction errors and the exploration-exploitation trade-off in simple non-
28 stationary multi-armed bandit tasks. We further show in rats performing such a task that systemically
29 antagonizing dopamine receptors greatly increases the number of random choices without affecting learning

30 capacities. Simulations and comparison of a set of different computational models (an extended Q-learning
31 model, a directed exploration model, and a meta-learning model) fitted on each individual confirm that,
32 independently of the model, decreasing dopaminergic activity does not affect learning rate but is equivalent to an
33 increase in exploration rate. This study shows that dopamine could adapt the exploration-exploitation trade-off in
34 decision making when facing changing environmental contingencies.

35

36 **Introduction**

37 All organisms need to make choices for their survival while being confronted to uncertainty in their
38 environment. Animals and humans tend to exploit actions likely to provide desirable outcomes, but they must
39 also take into account the possibility that environmental contingencies and the outcome of their actions may vary
40 with time. Behavioral flexibility is thus needed in volatile environments in order to detect and learn new
41 contingencies[1]. This requires a delicate balance between exploitation of known resources and exploration of
42 alternative options that may have become advantageous. How this exploration/exploitation dilemma may be
43 resolved and regulated is still a subject of active research in the fields of Neuroscience and Machine
44 Learning[2]–[5].

45 Dopamine holds a fundamental place in contemporary studies of learning and decision-making. The role of
46 phasic dopamine signals in supporting learning now appears strongly established[6]–[8]. Dopamine reward
47 prediction error signals have been identified in a variety of instrumental and Pavlovian conditioning tasks[9]–
48 [12]. They affect plasticity and action value learning in cortico-basal networks[13]–[15] and have been directly
49 related to behavioral adaptation in a number of decision-making tasks in humans, non-human primates[16] and
50 rodents[17]–[20]. Accordingly, it is commonly assumed that manipulations of dopamine activity affect the rate of
51 learning, but this could represent a misconception.

52 Besides learning, the role of dopamine in the control of behavioral performance is still unclear. Dopamine is
53 known to modulate incentive choice (the tendency to differentially weigh costs and benefits)[21], [22] and risk-
54 taking behavior[23], as well as other motivational aspects such as effort and response vigour[24]. Because
55 dopamine is one of the key factors that may encode success or uncertainty, it might modulate decisions by
56 biasing them toward options that present the largest uncertainty[25], [26]. This would correspond to a “directed”
57 exploration strategy[5], [27], [28]. Alternatively, success and failure could affect tonic dopamine levels and
58 control random exploration of all options, as recently proposed by Humphries et al. (2012)[29]. This form of
59 undirected exploration, which is often difficult to distinguish from performance, may be viewed as “noise” in the

60 choice process[30]–[32]. Previous computational analyses of behavioral data in stochastic tasks have yielded
61 mixed results, some suggesting a promotion of random[33] or directed[26] exploration by dopamine with
62 possibly an effect on learning[26], others a reduction of random exploration[29], [34], [35], and yet others an
63 effect on learning only[36].

64 In the present study, we first show formally that under fairly general assumptions, any manipulation that reduces
65 the magnitude of dopamine positive reward prediction errors does not change learning rate but instead changes
66 the level of random exploration. We then proceed to test this idea experimentally in rats while applying a variety
67 of computational models to the behavioral data. To dissociate learning and performance components,
68 probabilistic tasks where the best option changes with time (known in Machine Learning as non-stationary
69 bandit tasks) are particularly appropriate, because they require periodical exploration and relearning phases and
70 are amenable to computational modeling using well characterized reinforcement learning methods[37]. In this
71 work, we develop such a 3-armed bandit task in rats with varying levels of uncertainty to investigate how
72 dopamine controls the exploration level within an individual. We then examine the effects of dopamine blockade
73 on learning and performance variables following injection of various doses of the D1/D2 receptor antagonist
74 flupenthixol in different sessions. We follow by replicating these data with an extended reinforcement learning
75 model (Q-learning) with forgetting and verify our conclusions on a variety of alternative models such as a
76 directed exploration model, an ϵ -greedy random exploration model, and a meta-learning model. This allows us
77 to explicitly distinguish learning from exploration variables, and to show that dopamine activity is specifically
78 involved in controlling the level of random exploration rather than the learning rate.

79

80 **Results**

81 ***Mathematical relationship between reward prediction errors and random exploration-exploitation***

82 We first give here a formal demonstration (**Supporting material**) that in a Q-model such as the one we applied
83 to our task, a reduction of the amplitude of positive reward prediction errors (RPEs) directly translates into an
84 increase in random exploration levels. In other words, it is mathematically equivalent to changing the value of
85 the inverse temperature, and has no effect on the learning rate parameter. Briefly, in this demonstration, we
86 assume that positive, but not negative prediction errors are pharmacologically reduced, which corresponds to a
87 decrease in the value of the reward when it is present. On each trial, the Q-value of the performed action is
88 revised in proportion to the RPE, so it is exactly a fraction of what it would be in the absence of pharmacological

89 manipulation. For non-performed actions, if there is forgetting, the Q-value decreases in proportion to the Q-
90 value itself, which preserves proportionality. As a result, throughout the learning process, all Q-values are
91 downscaled in the same proportion as the reward. When these values are plugged into the softmax process, the
92 result is exactly equivalent to a decrease of the inverse temperature, again in the same proportion. The learning
93 rate or the forgetting rate are not affected in any way. This result shows that under fairly general conditions the
94 effects of a pharmacological manipulation of dopamine-dependent learning should be described as changes in
95 exploration rate rather than as changes in learning rate. Indeed, manipulation of these two factors predicts
96 distinct behavioral profiles: under different learning rates, both performance and win-shift curves (see Methods)
97 should differ at early stages of blocks, but then converge to similar levels (**Supplementary Fig. 2a,b**).
98 Conversely, when only the exploration rate differs, performance curves should tend toward different asymptotes
99 while win-shift curves should shift downwards as a whole when the exploration rate increases (**Supplementary**
100 **Fig. 2c,d**).

101

102 *Dopamine blockade affects exploratory behavior*

103 We then undertook to confirm this result experimentally on a 3-armed bandit task in rats. As a first step, we
104 examined rats' behavior at different phases of the task once it was well acquired. We focused our analysis on the
105 learning phase that is required each time the target lever changes. Overall, the rats were able to identify the
106 correct lever over the course of a block, despite the stochasticity of rewards (risk).

107 Performance (**Fig. 1b**) increased within a block toward an asymptote in both low and high risk conditions
108 ($F(5,110) = 186.7, p < 0.0001$), with better performance being observed under low risk ($F(1,22) = 148.2,$
109 $p < 0.0001$). Dopaminergic blockade by flupenthixol, a D1-D2 dopamine receptor antagonist, decreased
110 performance ($F(3,66) = 5.61, p = 0.0017$) irrespective of trial phase or risk condition (largest $F = 1.83, p = 0.15$).

111 We then looked at exploration, indexed as a first approximation by a win-shift index which only includes shifts
112 from the current best lever (**Fig. 2a**). This index decreased within blocks ($F(5,85) = 27.7, p < 0.0001$) but
113 increased with risk ($F(1,17) = 25.7, p < 0.0001$). Dopaminergic blockade dose-dependently elevated win-shift at
114 all stages ($F(3,51) = 14.5, p < 0.0001$), without interacting with trial or risk (largest $F: F(3,51) = 1.40, p = 0.25$).
115 Because win-shift in the early stages of blocks may not reflect exploration, but rather a return to a previously
116 reinforced lever, we also limited our analysis of win-shift to the last 8 trials when performance has stabilized,
117 meaning that the correct lever has been identified. In this case also there was a significant dose effect
118 (**Supplementary Fig. 1**). Another index of shifting behavior, lose-shift (**Fig. 3a**), which may denote a correction

119 strategy, was not significantly affected by the pharmacological condition, possibly because of a ceiling effect.
120 We then examined whether the effects of flupenthixol on both win-shift and performance resulted from a
121 negative effect on learning rate, or from a positive effect on exploration rate[38]. In the present task, these two
122 factors predict distinct behavioral profiles, in particular when considering asymptotic behavior (**Supplementary**
123 **Fig. 2**). Behavioral data indeed point to a change in exploration rate, as individual performance levels during the
124 last six trials of blocks were significantly affected by flupenthixol dose ($F(3,66) = 8.85, p < 0.0001$)
125 (**Supplementary Fig. 3a,b**). Similarly, the win-shift rate in the last 6 trials of blocks was significantly affected
126 by flupenthixol dose ($F(3,66) = 8.76, p < 0.0001$) (**Supplementary Fig. 3c,d**). This persistent difference in the
127 curves at the end of each learning phase is strongly suggestive of a change in exploration rate.

128

129 ***Flupenthixol alters exploration parameter in simulated rats***

130 We then examined how well a Q-learning model extended with a forgetting mechanism was able to account for
131 behavioral data (see **Methods**). Average action values derived from the model, individually fitted to each rat for
132 each dose of flupenthixol and constrained to the rat's actual choices, predicted a very high proportion of the
133 variance of individual rats' choices, (67.7% to 96.4%, **Supplementary Table 1** and **Supplementary Fig. 4**).
134 Moreover, unconstrained simulated data generated using these optimized parameters were highly similar to the
135 actual behavior of the rats (**Fig. 1c, Fig. 2b, Fig. 3B, Supplementary Fig.1, and Supplementary Fig. 6**). When
136 experimental and simulated data were pooled together for repeated-measures ANOVA, there was no significant
137 main effect of simulations on within-block evolution of either performance (Fig. 1c, $F(1,22) = 4.27, p = 0.051$)
138 or win-shift (Fig. 2b, $F(1,17) = 0.006, p = 0.94$). However, a significant interaction between simulations and risk
139 did emerge in the case of performance ($F(1,22) = 4.42, p = 0.047$) and there was also a significant interaction
140 between trials and simulations for both indicators ($F(5,110) = 6.98, p < 0.0001$ for performance; $F(5,85) = 2.90, p$
141 $= 0.018$ for win-shift). Crucially however, no interaction involving simulations and flupenthixol dose could be
142 detected (smallest $p=0.10$). Moreover, when analyzed separately from experimental data, simulated data did in
143 fact replicate the effects of flupenthixol on both measures ($F(3,66) = 5.23, p = 0.0027$ and $F(3,66) = 5.69, p =$
144 0.0016 for performance and win-shift respectively).

145

146 ***Model optimization dissociates exploration from learning***

147 To disentangle the effects of flupenthixol on learning versus performance, we then examined the values of the
148 different parameters α , β and α_2 across pharmacological conditions (**Fig. 4**). Flupenthixol had no discernible

149 effect on the learning rate α (Friedman ANOVA $\chi^2(3) = 4.04$, $p = 0.26$) or on the forgetting rate α_2 ($\chi^2(3) = 1.38$,
150 $p = 0.71$), but clearly decreased the exploration parameter β ($\chi^2(3) = 15.1$, $p = 0.0018$). *Post-hoc* tests revealed
151 that β for 0.2 and 0.3 mg/kg was significantly smaller than for 0 mg/kg ($p=0.012$ and $p = 0.0024$ respectively).
152 Thus, the only parameter of the model significantly affected by dopaminergic blockade was the exploration
153 parameter which decreased (i.e. exploration increased) as dopaminergic inhibition increased.

154 We next generalized this result by testing the optimized parameters of a range of different models which include
155 a standard Q-learning model (**Supplementary Fig. 7**), an ϵ -greedy model of action selection (**Supplementary**
156 **Fig. 8**), a model of directed exploration (**Supplementary Fig. 9**) which includes an uncertainty bonus to bias
157 decision-making towards options associated with a large value uncertainty[25] and a meta-learning model which
158 sets the value of the inverse temperature based on accumulated reward prediction errors (**Supplementary Fig.**
159 **10**). In all of these cases, the only parameter that significantly varied with dose condition was the one responsible
160 for controlling random exploration which flupenthixol invariably increased.

161 Finally, because of the strong interaction between learning rate and inverse temperature[39], we verified that our
162 methodology was able to distinguish variations in exploration rate β from variations in the learning rate α , by
163 applying full parameter optimization (leaving all three parameters free) to an artificial data set generated with
164 either α or β varying between doses while the other two parameters remain constant. When only the learning rate
165 α varied (**Fig. 5a**), the full optimization did indeed find that α significantly decreased ($\chi^2(3) = 16.1$, $p = 0.0011$),
166 but not the two other parameters ($\chi^2(3) \leq 1.54$, $p \geq 0.67$). Conversely, on an artificial data set where only the
167 exploration parameter β varied (**Fig. 5b**), the subsequent optimization correctly identified β as the only varying
168 parameter ($p = 0.0007$ but $p \geq 0.12$ for the other two parameters). These results clearly show that the presented
169 computational analysis can disentangle the effects of dopamine manipulations on exploration rate from possible
170 effects on learning rate.

171 Discussion

172 This study presents a formal demonstration that in simple reinforcement learning models, a reduction of the
173 amplitude of positive reward prediction errors directly translates into changes in random exploration levels. In
174 rats tested on the probabilistic choice task, we experimentally show that systemic administrations of
175 flupenthixol, a D1-D2 antagonist, dose-dependently increases random exploration, and only indirectly affects
176 performance. Dopamine blockade increased win-shift behavior under both low and high risk conditions and
177 noticeably late in a block when the rats had acquired the correct response. We reproduce behavioral data using
178 unconstrained simulations and show under a variety of models that exploration rate is the only parameter
179 significantly affected by dopamine blockade in this task.

180 To our knowledge, this is the first study that combines a within-subject approach with rigorous modelling to
181 evaluate the role of dopamine levels in controlling the balance between exploration and exploitation without
182 affecting the learning rate. The effects of various doses of dopamine antagonist in the same individual indicate
183 that normal dopamine levels limit undirected exploration in this task and therefore bias performance toward
184 exploitation without affecting the learning rate, a result which is consistent with that of Lee et al[34] using a
185 different animal model, the macaque, and a different decision-making task. Undirected exploration may reflect
186 several factors, and flupenthixol could for instance have reduced motivation[40], [41], response vigor[42] or
187 attention. However, the observation that learning rates were unchanged argues in favor of a selective effect on
188 exploratory choices.

189 In the task, performance improves within a block as one of the levers is gradually identified as target.
190 Concurrently, win-shift decreases as learning progresses. At the beginning of each block, performance drops to
191 chance levels or below, while win-shift increases. These high levels of win-shift in the absence of drug
192 correspond to moments when the values of the various options have not been well identified and the rat's
193 uncertainty is high. Indeed, in the high risk condition, identifying the correct lever appears more difficult and is
194 associated with both lower performance and higher win-shift levels. Under dopamine blockade, the dose-
195 dependent increase in win-shift appears independent of uncertainty, since it does not interact with either trial
196 rank within a block or risk level. This is consistent with the notion that dopamine unconditionally scales action
197 values and controls noise in the last stage of decision making, where action values are converted to actual
198 choices[29]. As to the effect of flupenthixol on lose-shift, the simulations (based on 100 runs) do show a slight
199 increase corresponding to increased random exploration. This effect does not reach significance in the actual
200 behavioral data, where it is probably more difficult to detect close to the ceiling value (probability of lose-shift in
201 random choice: 0.67).

202 Several past studies[26], [33], [43] have focused on the effects on choice of inter-individual differences in
203 dopaminergic function. Our data thus stand in contrast with those of Beeler *et al.*[33], [43] who observed that
204 hyperdopaminergic mice allocate more time and energy to expensive options. On the basis of computational
205 modeling, the authors interpreted their data as an increase in undirected exploration. However, data from
206 primates and human subjects instead indicate increases in exploration with reduced dopaminergic activity, in
207 agreement with our results. In particular, in an investigation of the behavioral disparities between human subjects
208 due to genes controlling prefrontal and striatal dopamine function, Frank *et al.*[26] concluded that COMT alleles
209 associated with lower dopamine levels increased directed exploration in their timing task. Eisenegger *et al.*[35]

210 found in a between-subject design that a strong dose of sulpiride, a dopamine D2 receptor antagonist, in healthy
211 human subjects favored exploration without affecting learning in a probabilistic task. These inter-individual
212 studies supposed a fixed exploration level, and were inadequate to describe changes in exploration in the same
213 subject. However, a similar result was achieved within subjects in an oculomotor decision task by Lee *et al.*[34]
214 who demonstrated that injections of dopamine type 2 receptor antagonist in the dorsal striatum of two macaques
215 deteriorated the animals' performance in a manner best explained by an increase of noise in the decision-making
216 process rather than an effect on learning.

217 In the three-arm bandit task used in the present study, we modeled learning of the correct action using an
218 extended Q-learning model with forgetting, and we modeled choice behavior using a softmax mechanism. Our
219 model was sufficient to account for behavioral performances as shown by i) the high similarity of the simulated
220 (unconstrained) and experimental behavioral data (**Figs. 1, 2, 3**) and ii) the high correlation between the modeled
221 value (constrained by the rats' choices) of the different levers and actual choice probability, even during periods
222 of low value when the target lever was not yet identified (**Supplementary Fig. 4**).

223 We show that a forgetting mechanism is required to adequately account for the rats' behavior as a simple Q-
224 learning mechanism appears unable to cope with the multiple target shifts (reversals) involved throughout the
225 task (**Supplementary Fig. 5**). In our model, forgetting is important to reduce the value of competing actions
226 even when these actions are not chosen any more, unlike simple Q-learning which only adjusts the value of
227 actions actually performed. The observation that the forgetting rate is generally larger than the learning rate (**Fig.**
228 **4**) implies that the rats tend to persist on a choice even in the absence of reward[37]. This process stands in
229 contrast with some theories of directed exploration[44] which predict that unchosen options become attractive as
230 uncertainty about their outcome increases.

231 Our model does not include any mechanism to track uncertainty about action values, unlike several models of
232 choice behavior in humans[4], [5], [26], [28], [45], [46]. Our simulations furthermore show that the gradual
233 reduction in win-shift within a block does not reflect a dynamic adaptation of the model parameters since it is
234 reproduced in the simulations where these parameters are kept constant. Instead, this decrease is a consequence
235 of the interaction between value learning and the softmax mechanism. Choice is more variable when actions
236 values are relatively similar, and it then becomes less variable as the values of the various actions get better
237 differentiated. We did observe a significant effect of risk on performance and win-shift in the behavioral data, but
238 because the same effect was present in the simulations, it is attributable to a slower acquisition of value in the
239 high risk situation due to increased stochasticity.

240 As expected from the formal analysis, dopamine was found to specifically control the exploration parameter β
241 (inverse temperature) which represents undirected exploration or random noise in the choice process converting
242 values to actions[27], [31], [32], rather than directed exploration driven by uncertainty[5], [25], [26], [28].
243 Furthermore, this result is still valid with other models such as the standard Q-learning, the ϵ -greedy version of
244 Q-learning and even a directed exploration model. Our data agree with the theoretical proposal by Humphries et
245 al. (2012)[29] that tonic dopamine in the basal ganglia could modulate the exploration-exploitation trade-off
246 during decision-making, On the basis of a prior, biologically inspired model of the basal ganglia[47], they
247 showed that changing simulated tonic dopamine levels had similar effects as changes in the β parameter.

248 Our study highlights a common misconception that equates the well-established role of dopamine in learning[48]

249 with an effect on learning rate. To the extent that learning is based on reward prediction errors and action
250 selection on a softmax mechanism, as is typically assumed in model-free reinforcement learning, our formal
251 analysis indicates that the inverse temperature, the parameter controlling random exploration, is the only
252 parameter affected by simple manipulations of the reward prediction error signal. Notably, in our task, both
253 behavioral performance and fitted learning rate were largely unaffected by flupenthixol. In a probabilistic
254 learning task, Pessiglione *et al.* [49] showed that administration of L-DOPA, a chemical precursor of dopamine
255 known for enhancing dopaminergic functions, improved performance in accumulating gain compared to subjects
256 under a dopamine antagonist. This improvement was attributed to an increase in learning from positive reward
257 errors, but increased dopamine could also have reduced exploration. Similarly, Krugel 2009[36] reported that
258 COMT alleles increasing dopamine levels were associated with better performance in a reward-based learning
259 task with reversals, and they explained their results by a modulation of learning rate. In contrast, there are reports
260 in humans that probabilistic learning is insensitive to dopamine antagonists[50], [51]. Our results call for careful
261 modeling of the impact of dopaminergic manipulations in behavioral tasks as changes in random exploration
262 rates could easily be mistaken for changes in learning rate.

263 As fluctuations in tonic dopamine levels track the average reward rate[24], it seems natural to use such a signal
264 to regulate the exploration-exploitation trade-off[32]: high reward rates suggest that the current policy is
265 appropriate and the subject could crystallize its behavior by exploiting more. Conversely, sudden drops in reward
266 rate leading to tonic dopamine decreases may lead to increased exploration of the environment in search for
267 better options. Dopamine levels could thus contribute to dynamically regulate exploratory choices in volatile
268 environments where option values change with time. Here, we show that dopamine blockade affects undirected
269 exploration independently from the changes in uncertainty levels within blocks. Dopaminergic regulation of
270 exploration appears to occur at a longer time scale than that of a few trials, which would constitute a form of
271 meta-learning[3], [32] adapting behavior to the general characteristics of the task rather than to immediate
272 events.

273

274

275 **Methods**

276 ***Behavior***

277 Male Long Evans rats (n=24) were obtained from Janvier Labs (France) at the age of 2 months and initially
278 accustomed to the laboratory facility for two weeks before the beginning of the experiments. They were housed
279 in pairs in standard polycarbonate cages (49 × 26 × 20 cm) with sawdust bedding. The facility was maintained at
280 21±1°C, with a 12-hour light/dark cycle (7 AM/7 PM) with food and water initially available *ad libitum*. Rats
281 were tested only during the light portion of the cycle. The experiments were conducted in agreement with French
282 (council directive 2013-118, February 1, 2013) and international (directive 2010-63, September 22, 2010,
283 European Community) legislations and received approval # 5012064-A from the local Ethics Committee of
284 Université de Bordeaux.

285 Animals were trained and tested in eight identical conditioning chambers (40 cm wide × 30 cm deep × 35 cm
286 high, Imetronic, Pessac, France), each located inside a sound and light-attenuating wooden compartment (74 ×
287 46 × 50 cm). Each compartment had a ventilation fan producing a background noise of 55 dB and four light-
288 emitting diodes on the ceiling for illumination of the chamber. Each chamber had two opaque panels on the right
289 and left sides, two clear Perspex walls on the back and front sides, and a stainless-steel grid floor (rod diameter:
290 0.5 cm; inter-rod distance: 1.5 cm). Three retractable levers (4 × 1 × 2 cm) could be inserted on the left wall. In
291 the middle of the opposite wall, a magazine (6 × 4.5 × 4.5 cm) collected food pellets (45 mg, F0165, Bio_Serv,
292 NJ, USA) from a dispenser located outside the operant chamber. The magazine was equipped with infrared cells
293 to detect the animal's visits. Three LED (one above each lever) were simultaneously lit as a signal for trial onset.
294 A personal computer connected to the operant chambers via an Imetronic interface and equipped with POLY
295 software (Imetronic, Pessac, France) controlled the equipment and recorded the data.

296 During the behavioral experiments, rats were maintained at 90% of their original weight by restricting their food
297 intake to ~15 g/day. For pre-training, all rats were trained for 3 days to collect rewards during 30 min magazine
298 training sessions. Rewards were delivered in the magazine on a random time 60 sec schedule. The conditioning
299 cage was lit for the duration of each session. The rats then received training for 3 days under a continuous
300 reinforcement, fixed ratio schedule FR1 (i.e. each lever press was rewarded with one pellet) until they had
301 earned 30 pellets or 30 min had elapsed. At this stage, each lever was presented continuously for one session and
302 the magazine was placed adjacent to the lever (side counterbalanced across rats). Thereafter, all three levers were
303 on the left wall and the magazine on the right wall. The levers were kept retracted throughout the session except
304 during the choice phases. On the next two sessions, levers were successively presented 30 times in a pseudo-
305 random order (FR1-trials). One press on the presented lever produced a reward and retraction of the lever. On the
306 next eight sessions, levers were presented 30 times but each time five presses were required to obtain the reward
307 (FR5-trials). As a result, all rats readily pressed the levers as soon as they were presented. The rats then
308 underwent 24 training sessions of training in the probabilistic choice task, 20 sessions of six trial blocks each and
309 four double sessions of 12 blocks each.

310 Following 24 sessions of training in the task, rats received i.p. injections (1ml/kg) of D1-D2 receptor antagonist
311 Flupenthixol (FLU) or saline, 20 min prior to each double session of test, for a total of 16 injections. Four doses
312 of Flupenthixol (Cis-(Z)-Flupenthixol dihydrochloride, Sigma, dissolved in saline at 0, 0.1, 0.2 or 0.3 mg/ml)
313 were selected according to a pilot experiment. All rats received each dose (0, 0.1, 0.2 or 0.3 mg/kg) in separate
314 sessions according to a latin square design with at least two days of recovery between injections. After two days

315 of recovery, a second series of injections was performed under similar conditions. In addition, on the day
316 preceding each of the eight tests, a retraining session under saline was performed.

317 The experimental task (**Fig. 1a**) consisted in a three-armed bandit task where rats had to select one of three
318 levers in order to receive the reward. A trial began with a 2 sec warning light, and then the three retractable
319 levers were presented to the rat. Pressing one of the levers could immediately result in the delivery of a reward
320 with various probabilities. Two different risk levels were imposed: In the low risk condition (LR) one lever was
321 designated as the target lever and rewarded with probability 7/8 (87.5%) while the other levers were rewarded
322 with probability 1/16 (6.25%). In the high risk condition (HR), the target lever was rewarded with probability 5/8
323 (62.5%) and the other two possibilities with probability 3/16 (18.75%), making discrimination of the target lever
324 much harder. After a lever press, the levers were retracted and the trial (rewarded or not) was terminated. Inter-
325 trial interval randomly varied in range 4.5-8 sec. Trials were grouped into unsignaled blocks of fixed length (24
326 trials each) characterized by a constant combination of target lever and risk. The target lever always changed
327 between block. Therefore rats had to re-learn the target lever on each block. Blocks were ordered pseudo-
328 randomly within a session with all combinations of target and risk counterbalanced and tested twice.

329

330 ***Model fitting and simulations***

331 Behavioral data were modeled by assuming that rats continuously assign a value to each lever. This value is
332 adjusted according to a standard Q-learning algorithm, with the addition of a forgetting mechanism. On each trial
333 t where an action a_t was chosen, values are learned gradually by first calculating a reward prediction error δ
334 representing the discrepancy between the reward received r_t and what was expected, i.e. the previous estimate of
335 the value of the chosen action $Q_t(a_t)$, and then using δ to update the value of this action $Q_{t+1}(a_t)$:

$$336 \quad r_t = \begin{cases} 1 & \text{if the trial is rewarded} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$337 \quad \delta_t = r_t - Q_t(a_t) \quad (2)$$

$$338 \quad Q_{t+1}(a_t) = Q_t(a_t) + \alpha \delta_t \quad (3)$$

339

340 The learning rate parameter α determines how quickly the system learns from observed outcomes: low values
341 ensure that action values are relatively stable. In the present task where the correct action periodically changes,
342 higher learning rates should allow a rapid increase in performance across a block, at the cost of an increased

343 sensitivity to the stochastic nature of reinforcement.

344 In order to improve the intra-block dynamics of the model, a forgetting mechanism[52] was added:

345

$$346 \quad Q_{t+1}(a \neq a_t) = (1 - \alpha_2) Q_t(a \neq a_t) \quad (4)$$

347 Thus, the Q-values of non-selected actions gradually regress to 0, at which they are initialized, at a rate fixed by
348 constant α_2 . These values would otherwise only be updated when the corresponding action is selected. This
349 mechanism was found to be necessary to achieve a good fit of the dynamics of win-shift (see **Supplementary**
350 **Figs. 5 and 6** comparing the win-shift curves of forgetting and non-forgetting models to the experimental data)
351 and corresponds to a perseverance mechanism independent of reward history[37]. Additionally, once combined
352 with the action selection model described further down, we compared the log-likelihoods of the two optimized
353 models adjusted for the number of extra parameters using either the Akaike Information Criterion (AIC) or the
354 Bayesian Information Criterion (BIC) calculated for the whole population (n=23) and the four conditions (the
355 better model being the one with lower scores). The Q-learning model with forgetting parameters had lower
356 scores than the model without forgetting (AIC: 74534 vs. 82772 ; BIC: 76149 vs. 84386). We also compared
357 individual log-likelihood scores for each rat and dose using the likelihood ratio test for nested models[39]. Given
358 a model M1 nested into a more complex model M2 and their log-likelihood after optimization ll1 and ll2, $d =$
359 $2 \cdot (ll2 - ll1)$ follows a chi-squared distribution with 1 degree of freedom (1 added parameter in M2) under the null
360 hypothesis that the log-likelihood of M2 is not better than M1. In all rats except one, the forgetting model
361 brought a highly significant improvement in likelihood when compared to the simpler model.

362 Given the estimated Q-value for each action, actions are selected by sampling from a softmax probability
363 distribution:

$$364 \quad P(a_{t+1} = a_i) = \frac{e^{\beta Q_t(a_i)}}{\sum_j e^{\beta Q_t(a_j)}} \quad (5)$$

365 The key parameter of this function, the inverse temperature β (here called "exploration parameter"), determines
366 the relationship between action values and action selection, or in other terms between learning and performance.
367 Low values of β result in almost equiprobable action selection (hence exploration), independently of the learned
368 Q-values, while high values of β greatly favour the best action over all the others (i.e. exploitation). This

369 equation (4) is especially crucial for optimization because it defines the probability of a rat's action at each trial
370 given its parameter-set Θ (including learned values), which was used to calculate the likelihood of each rat's
371 entire history of choices, H , under the parameters of each model:

372

$$373 \quad P\langle H|\theta\rangle = \prod_{t=1}^{n_{\text{trials}}} P\langle a_t|\theta\rangle = \prod_{t=1}^{n_{\text{trials}}} \frac{e^{\beta Q_t(a_t)}}{\sum_j e^{\beta Q_t(a_j)}}$$

374 (6)

375 Parameter optimization was then carried out separately for each one of the four pharmacological conditions (0,
376 0.1, 0.2 and 0.3 mg/kg administrations of flupenthixol) by maximizing the log-likelihood of the model for each
377 rat so as to get individual parameter-sets for each subject. Each parameter was initialized in three different points
378 of its parameter space to avoid being trapped in local maxima.

379 To verify that conclusions drawn from the characteristics of the optimized model, we checked that the model was
380 able to properly reproduce the experimental data by running unconstrained simulations using the individual
381 parameters fitted to each rat 100 times on the full experiment. These simulations were then averaged and
382 compared to the original data. This verification procedure has recently been advocated[53] as a standard and
383 crucial requirement when modeling experimental data.

384 We used the same optimization method on four additional models. The first of these additional models is the
385 standard Q-learning model which is identical to the model just presented except it has no forgetting mechanism.
386 The second additional model is the forgetting Q-learning model presented earlier with an ϵ -greedy action
387 selection mechanism instead of the softmax rule. This mechanism simply selects the action with highest Q-value
388 with probability $1-\epsilon$ and the remaining two actions with probability $\epsilon/2$. Thus the larger the ϵ parameter, the
389 more exploratory the behavior of the animal. The third model we tested is taken from the literature[25] and
390 consists in combining Q-values, which estimate the expected payoff of a given action, with an uncertainty bonus
391 u , which estimates the variance of these payoffs. Q-values are updated in the same way as the extended Q-
392 learning model (equations (2) and (3)). Each time an action a is selected and rewarded, we calculate the
393 uncertainty prediction error $\xi_t(a)$ based on squared reward prediction errors (taken from equation 1) and the
394 previous estimate of the variance:

$$395 \quad \xi_t(a) = \delta_t^2 - v_{t-1}(a) \quad (7)$$

396 Using squared prediction errors means that it is the magnitude of reward prediction errors which matters and

397 provides us with an estimate of variance. The expected uncertainty of the same action can then be updated using
398 its own learning rate parameter α^Φ :

$$399 \quad v_t(a) = v_{t-1}(a) + \alpha^\Phi \xi_t(a) \quad (8)$$

400 The expected uncertainties of the other two actions remain the same, and all expected uncertainties were
401 initialized at 0, as were the Q-values. Finally, the expected uncertainties are combined with the Q-values through
402 a weighting parameter Φ before being plugged into the softmax equation:

$$403 \quad P(a_{t+1} = a_t) = \frac{e^{\beta(Q_t(a_t) + \Phi v_t(a_t))}}{\sum_j e^{\beta(Q_t(a_j) + \Phi v_t(a_j))}} \quad (9)$$

404 We chose this model as a representative of directed-exploration models as it can bias action selection towards
405 potentially unrewarding choices if these are highly uncertain. As explained in the main text of this article, even
406 when using this sophisticated version of exploration, it is still β , the parameter controlling random exploration,
407 which is affected by dopamine inhibition.

408 Finally, given that our reported findings suggest the possibility of a meta-learning process based on
409 dopaminergic control of the exploration rate, we also tested a forgetting Q-learning model in which β is
410 controlled by an accumulation of past reward prediction errors R_t , intended to represent tonic dopamine (under
411 the simple assumption that tonic dopamine is simply the result of past phasic activity):

$$412 \quad R_t = R_{t-1} + \alpha^R \delta_t \quad (10)$$

413 The inverse temperature is simply defined as a linear function of R_t :

$$414 \quad \beta_{t+1} = \beta_0 + (\beta_1 - \beta_0) R_t \quad (11)$$

415 This model presents three additional parameters compared to the simple forgetting model, α^R , which simply
416 determines how much impact recent prediction errors have relative to older ones, β which is the value of β when
417 R_t equals 0, and β_1 , the value of when R_t equals 1. Initially, we set R_0 to 0 and by the same logic β equal to β_0 .
418 When optimized, we found a small significant effect of flupenthixol on β_1 , comforting once again our results.

419

420 **Data analysis**

421 Trials from each 24-trial block were grouped into six bins of four trials and averaged for each of the four
422 flupenthixol doses and each of the two risk levels in each rat. Asymptotic performance and win-shift levels were
423 estimated on the last six trials of blocks, respectively. Performance was represented by the proportion of trials
424 where the target lever was selected. Exploration was indexed by win-shift, the proportion of trials where rats

425 changed lever choice after being rewarded on the target lever. One rat did not complete the task under 0.3 mg/kg
426 of flupenthixol and was therefore removed from analysis. Five other rats missed occasional parts of the blocks
427 (no win), so statistical analysis of win-shift concerned 18 rats. Experimental or simulated data were submitted to
428 repeated-measures ANOVAs with three factors (flupenthixol dose, risk and trial) with an additional factor
429 (experiment/model) when appropriate. *Post-hoc* comparisons were performed using simple t-tests with
430 Bonferroni correction. Model parameters were analyzed as a function of dose using non-parametric Friedman's
431 ANOVA as the distribution of β values violated the assumptions of normality (Shapiro-Wilk test of normality
432 $p < 0.0001$ for 0.2 mg/kg of flupenthixol), and sphericity (Mauchly test: $\chi^2(5) = 24.68$, $p = 0.0002$) required for a
433 repeated-measures ANOVA. First-order error risk was set at 0.05 (two-sided tests).

434

435 **Acknowledgements**

436 We thank Paul Apicella, Mark D. Humphries, Stefano Palminteri and Emmanuel Procyk for valuable
437 discussions, and Yoan Salafranke for animal care. This work was supported by the French Agence Nationale de
438 la Recherche (ANR) "Learning Under Uncertainty" Project under reference ANR-11-BSV4-006,
439 "Neurobehavioral assessment of a computational model of reward learning" CRCNS 2015 Project under
440 reference ANR-15-NEUC-0001, and by Labex SMART (ANR-11- LABX-65) Online Budgeted Learning Project
441 supported by French state funds managed by the ANR within the Investissements d'Avenir programme under
442 reference ANR-11-IDEX-0004-02.

443

444 **Author Contributions**

445 A.M., M.K., E.C. and B.G. designed the experiments. V.F. carried out the behavioral protocols. V.F. and A.M.
446 analyzed the behavioral data. M.K., B.G. and F.C. designed the models. F.C. and N.A. ran the models and
447 simulations. F.C., M.K., B.G. and A.M. analyzed the simulation data. F.C., A.M. and M.K. wrote the manuscript.
448 All authors discussed the results and commented on the manuscript.

449

450 **Competing Interests Statement**

451 The authors declare no competing financial or non-financial interests.

452

453

454 References

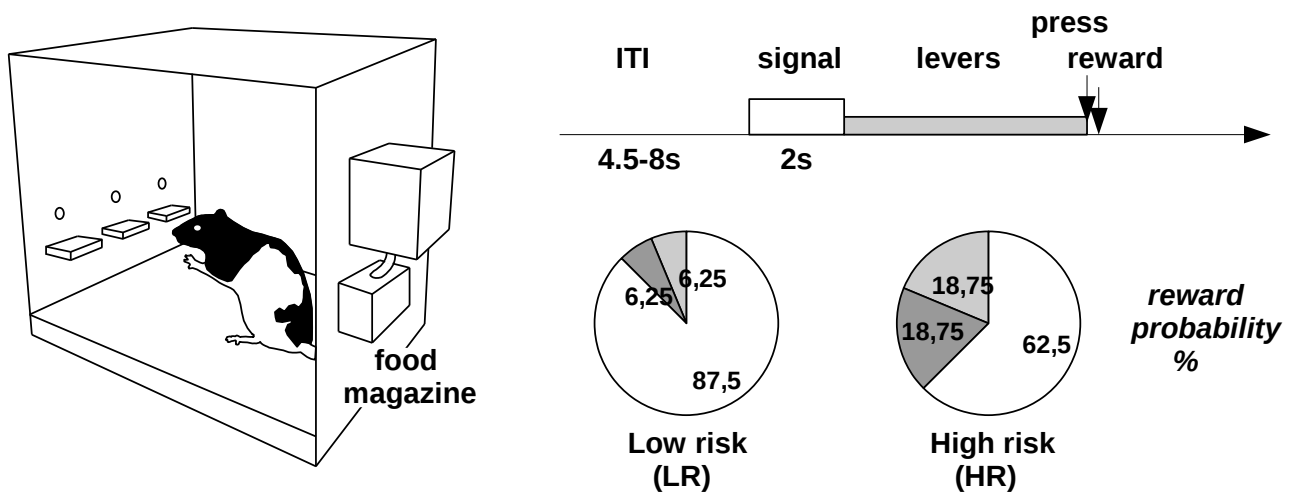
- 455 [1] M. F. S. Rushworth and T. E. J. Behrens, "Choice, uncertainty and value in prefrontal and
456 cingulate cortex," *Nature Neuroscience*, vol. 11, no. 4. pp. 389–397, 2008.
- 457 [2] N. D. Daw, J. P. O'Doherty, P. Dayan, B. Seymour, and R. J. Dolan, "Cortical substrates for
458 exploratory decisions in humans," *Nature*, vol. 441, no. 7095, pp. 876–879, 2006.
- 459 [3] N. Schweighofer and K. Doya, "Meta-learning in reinforcement learning," *Neural Networks*, vol.
460 16, no. 1, pp. 5–9, 2003.
- 461 [4] M. R. Nassar, R. C. Wilson, B. Heasley, and J. I. Gold, "An Approximately Bayesian Delta-Rule
462 Model Explains the Dynamics of Belief Updating in a Changing Environment," *J. Neurosci.*, vol.
463 30, no. 37, pp. 12366–12378, Sep. 2010.
- 464 [5] R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, and J. D. Cohen, "Humans use directed and
465 random exploration to solve the explore-exploit dilemma," *J exp Psychol Gen*, vol. 143, no. 6,
466 pp. 2074–2081, 2014.
- 467 [6] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward.,"
468 *Science (80-.)*, vol. 275, pp. 1593–1599, 1997.
- 469 [7] W. Schultz, "Updating dopamine reward signals," *Current Opinion in Neurobiology*, vol. 23, no.
470 2. pp. 229–238, 2013.
- 471 [8] M. Watabe-Uchida, N. Eshel, and N. Uchida, "Neural Circuitry of Reward Prediction Error,"
472 *Annu. Rev. Neurosci.*, vol. 40, no. 1, pp. 373–394, 2017.
- 473 [9] H. M. Bayer and P. W. Glimcher, "Midbrain Dopamine Neurons Encode a Quantitative Reward
474 Prediction Error Signal," *Neuron*, vol. 47, no. 1, pp. 129–141, 2005.
- 475 [10] G. Morris, A. Nevet, D. Arkadir, E. Vaadia, and H. Bergman, "Midbrain dopamine neurons
476 encode decisions for future action.," *Nat. Neurosci.*, vol. 9, no. 8, pp. 1057–1063, 2006.
- 477 [11] M. R. Roesch, D. J. Calu, and G. Schoenbaum, "Dopamine neurons encode the better option
478 in rats deciding between differently delayed or sized rewards.," *Nat. Neurosci.*, vol. 10, no. 12,
479 pp. 1615–24, 2007.
- 480 [12] M. Matsumoto and O. Hikosaka, "Two types of dopamine neuron distinctly convey positive and
481 negative motivational signals," *Nature*, vol. 459, no. 7248, pp. 837–841, Jun. 2009.
- 482 [13] D. Centonze, B. Picconi, P. Gubellini, G. Bernardi, and P. Calabresi, "Dopaminergic control of
483 synaptic plasticity in the dorsal striatum," *Eur. J. Neurosci.*, vol. 13, no. 6, pp. 1071–1077, 2001.
- 484 [14] J. N. Reynolds, B. I. Hyland, and J. R. Wickens, "A cellular mechanism of reward-related
485 learning.," *Nature*, vol. 413, no. 6851, pp. 67–70, Sep. 2001.
- 486 [15] E. M. Izhikevich, "Solving the distal reward problem through linkage of STDP and dopamine
487 signaling," *Cereb. Cortex*, vol. 17, no. 10, pp. 2443–2452, 2007.
- 488 [16] V. D. Costa, V. L. Tran, J. Turchi, and B. B. Averbeck, "Dopamine modulates novelty seeking
489 behavior during decision making," *Behav. Neurosci.*, vol. 128, no. 5, pp. 556–566, 2014.
- 490 [17] D. M. Haluk and S. B. Floresco, "Ventral striatal dopamine modulation of different forms of
491 behavioral flexibility.," *Neuropsychopharmacology*, vol. 34, no. 8, pp. 2041–52, Jul. 2009.
- 492 [18] S. B. Flagel *et al.*, "A selective role for dopamine in stimulus–reward learning," *Nature*, vol. 469,
493 no. 7328, pp. 53–57, Jan. 2011.

- 494 [19] G. K. Papageorgiou, M. Baudonnat, F. Cucca, and M. E. Walton, "Mesolimbic Dopamine
495 Encodes Prediction Errors in a State-Dependent Manner.," *Cell Rep.*, vol. 15, no. 2, pp. 221–8,
496 Apr. 2016.
- 497 [20] N. L. Jenni, J. D. Larkin, and S. B. Floresco, "Prefrontal Dopamine D1 and D2 Receptors
498 Regulate Dissociable Aspects of Decision Making via Distinct Ventral Striatal and Amygdalar
499 Circuits.," *J. Neurosci.*, vol. 37, no. 26, pp. 6200–6213, Jun. 2017.
- 500 [21] J. Salamone, M. Correa, S. Mingote, and S. Weber, "Beyond the reward hypothesis: alternative
501 functions of nucleus accumbens dopamine," *Curr. Opin. Pharmacol.*, vol. 5, no. 1, pp. 34–41,
502 Feb. 2005.
- 503 [22] C. W. Berridge and A. F. T. Arnsten, "Psychostimulants and motivated behavior: Arousal and
504 cognition," *Neurosci. Biobehav. Rev.*, vol. 37, no. 9, pp. 1976–1984, Nov. 2013.
- 505 [23] C. M. Stopper, M. T. L. Tse, D. R. Montes, C. R. Wiedman, and S. B. Floresco, "Overriding
506 Phasic Dopamine Signals Redirects Action Selection during Risk/Reward Decision Making,"
507 *Neuron*, vol. 84, no. 1, pp. 177–189, Oct. 2014.
- 508 [24] Y. Niv, N. D. Daw, D. Joel, and P. Dayan, "Tonic dopamine: Opportunity costs and the control of
509 response vigor," *Psychopharmacology (Berl.)*, vol. 191, no. 3, pp. 507–520, 2007.
- 510 [25] J. Naudé *et al.*, "Nicotinic receptors in the ventral tegmental area promote uncertainty-seeking,"
511 *Nat. Neurosci.*, no. October 2015, 2016.
- 512 [26] M. J. Frank, B. B. Doll, J. Oas-Terpstra, and F. Moreno, "The neurogenetics of exploration and
513 exploitation: Prefrontal and striatal dopaminergic components," in *Nature Neuroscience*, 2009,
514 vol. 12, no. 8, pp. 1062–1068.
- 515 [27] W. K. Zajkowski, M. Kossut, and R. C. Wilson, "A causal role for right frontopolar cortex in
516 directed, but not random, exploration," *Elife*, vol. 6, pp. 1–18, 2017.
- 517 [28] I. Cogliati Dezza, A. J. Yu, A. Cleeremans, and W. Alexander, "Learning the value of information
518 and reward over time when solving exploration-exploitation problems," *Sci. Rep.*, vol. 7, no. 1,
519 p. 16919, Dec. 2017.
- 520 [29] M. D. Humphries, M. Khamassi, and K. Gurney, "Dopaminergic control of the exploration-
521 exploitation trade-off via the basal ganglia," *Front. Neurosci.*, vol. 6, no. FEB, pp. 1–14, 2012.
- 522 [30] R. Sutton and A. Barto, "Reinforcement Learning: An Introduction," *MIT Press, Cambridge,*
523 *Massachusetts*, 1998. .
- 524 [31] K. Doya, "Modulators of decision making," *Nat. Neurosci.*, vol. 11, no. 4, pp. 410–416, Apr.
525 2008.
- 526 [32] M. Khamassi, P. Enel, P. F. Dominey, and E. Procyk, "Medial prefrontal cortex and the adaptive
527 regulation of reinforcement learning parameters," *Prog Brain Res*, vol. 202, pp. 441–464, 2013.
- 528 [33] J. a Beeler, N. Daw, C. R. M. Frazier, and X. Zhuang, "Tonic dopamine modulates exploitation
529 of reward learning.," *Front. Behav. Neurosci.*, vol. 4, no. November, p. 170, 2010.
- 530 [34] E. Lee, M. Seo, O. Dal Monte, and B. B. Averbeck, "Injection of a Dopamine Type 2 Receptor
531 Antagonist into the Dorsal Striatum Disrupts Choices Driven by Previous Outcomes, But Not
532 Perceptual Inference," *J. Neurosci.*, vol. 35, no. 16, pp. 6298–6306, 2015.
- 533 [35] C. Eisenegger *et al.*, "Role of dopamine D2 receptors in human reinforcement learning.,"
534 *Neuropsychopharmacology*, vol. 39, no. 10, pp. 2366–75, 2014.

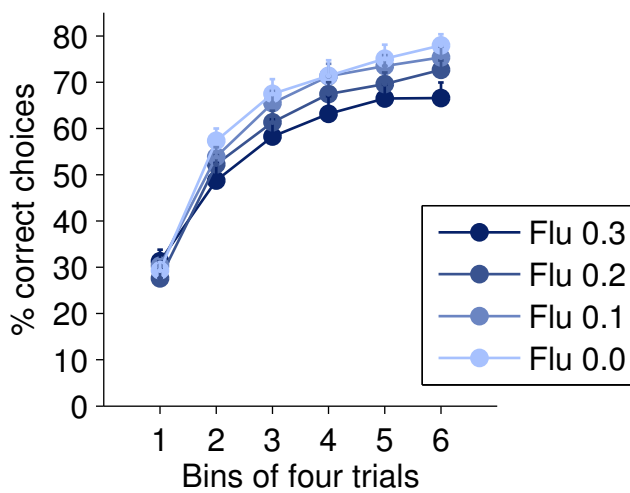
- 535 [36] L. K. Krugel, G. Biele, P. N. C. Mohr, S.-C. Li, and H. R. Heekeren, "Genetic variation in
536 dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions.,"
537 *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 42, pp. 17951–6, Oct. 2009.
- 538 [37] K. Katahira, "The relation between reinforcement learning parameters and the influence of
539 reinforcement history on choice behavior," *J. Math. Psychol.*, vol. 66, pp. 59–69, 2015.
- 540 [38] B. B. Averbeck and V. D. Costa, "Motivational neural circuits underlying reinforcement
541 learning," *Nat. Neurosci.*, vol. 20, no. 4, pp. 505–512, 2017.
- 542 [39] N. D. Daw, "Trial-by-trial data analysis using computational models," *Decis. Making, Affect.*
543 *Learn. Atten. Perform. XXIII*, pp. 1–26, 2011.
- 544 [40] A. Dickinson, J. Smith, and J. Mirenowicz, "Dissociation of Pavlovian and instrumental
545 incentive learning under dopamine antagonists.," *Behav. Neurosci.*, vol. 114, no. 3, pp. 468–83,
546 Jun. 2000.
- 547 [41] M. F. Barbano, M. Le Saux, and M. Cador, "Involvement of dopamine and opioids in the
548 motivation to eat: influence of palatability, homeostatic state, and behavioral paradigms,"
549 *Psychopharmacology (Berl.)*, vol. 203, no. 3, pp. 475–487, Apr. 2009.
- 550 [42] Y. Niv, "Cost, benefit, tonic, phasic: What do response rates tell us about dopamine and
551 motivation?," *Ann. N. Y. Acad. Sci.*, vol. 1104, pp. 357–376, 2007.
- 552 [43] J. a. Beeler, C. R. M. Frazier, and X. Zhuang, "Putting desire on a budget: dopamine and
553 energy expenditure, reconciling reward and resources.," *Front. Integr. Neurosci.*, vol. 6, no.
554 July, p. 49, 2012.
- 555 [44] B. B. Averbeck, "Theory of Choice in Bandit, Information Sampling and Foraging Tasks," *PLoS*
556 *Comput. Biol.*, vol. 11, no. 3, pp. 1–28, 2015.
- 557 [45] T. E. J. Behrens, M. W. Woolrich, M. E. Walton, and M. F. S. Rushworth, "Learning the value of
558 information in an uncertain world.," *Nat. Neurosci.*, vol. 10, no. 9, pp. 1214–21, 2007.
- 559 [46] M. Jepma, P. R. Murphy, M. R. Nassar, M. Rangel-Gomez, M. Meeter, and S. Nieuwenhuis,
560 "Catecholaminergic Regulation of Learning Rate in a Dynamic Environment," *PLOS Comput.*
561 *Biol.*, vol. 12, no. 10, p. e1005171, Oct. 2016.
- 562 [47] K. N. Gurney, M. Humphries, R. Wood, T. J. Prescott, and P. Redgrave, "Testing computational
563 hypotheses of brain systems function: a case study with the basal ganglia.," *Network*, vol. 15,
564 no. 4, pp. 263–90, Nov. 2004.
- 565 [48] W. Schultz, "Neuronal Reward and Decision Signals: From Theories to Data," *Physiol. Rev.*,
566 vol. 95, no. 3, pp. 853–951, 2015.
- 567 [49] M. Pessiglione, B. Seymour, G. Flandin, R. J. Dolan, and C. D. Frith, "Dopamine-dependent
568 prediction errors underpin reward-seeking behaviour in humans," *Nature*, vol. 442, no. 7106,
569 pp. 1042–1045, 2006.
- 570 [50] T. Shiner, M. Symmonds, M. Guitart-Masip, S. M. Fleming, K. J. Friston, and R. J. Dolan,
571 "Dopamine, salience, and response set shifting in prefrontal cortex," *Cereb. Cortex*, vol. 25, no.
572 10, pp. 3629–3639, 2015.
- 573 [51] P. Smittenaar, H. W. Chase, E. Aarts, B. Nusslein, B. R. Bloem, and R. Cools, "Decomposing
574 effects of dopaminergic medication in Parkinson's disease on probabilistic action selection -
575 learning or performance?," *Eur. J. Neurosci.*, vol. 35, no. 7, pp. 1144–1151, 2012.
- 576 [52] M. Ito and K. Doya, "Validation of decision-making models and analysis of decision variables in

- 577 the rat basal ganglia,” *J. Neurosci.*, vol. 29, no. 31, pp. 9861–9874, 2009.
- 578 [53] S. Palminteri, V. Wyart, and E. Koechlin, “Computational cognitive neuroscience: what is a
579 good model of brain processes?,” pp. 1–20, 2016.
- 580 [54] F. Lesaint, O. Sigaud, S. B. Flagel, T. E. Robinson, and M. Khamassi, “Modelling Individual
581 Differences in the Form of Pavlovian Conditioned Approach Responses: A Dual Learning Systems
582 Approach with Factored Representations,” *PLoS Comput. Biol.*, vol. 10, no. 2, 2014.
- 583

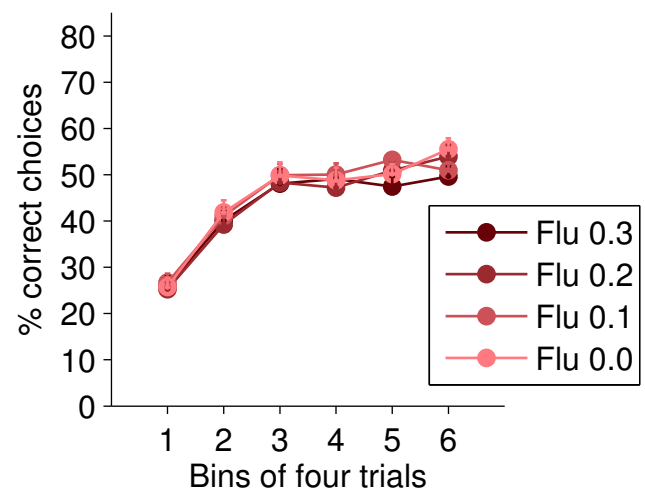
a



b Performance in low risk blocks



Performance in high risk blocks



c

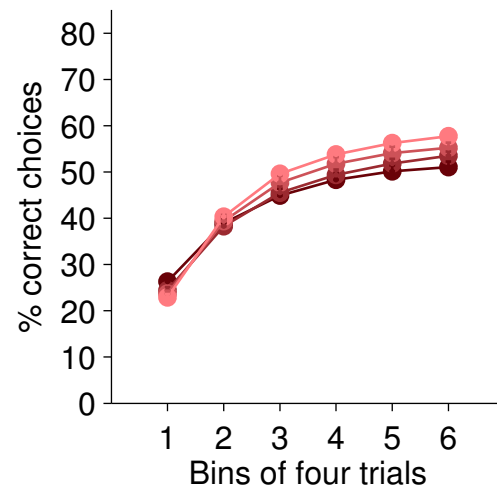
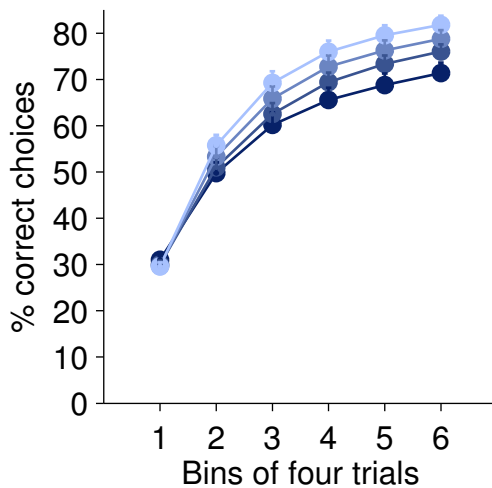


Figure 1. Actual and simulated performances. **(a)** Outline of the experimental task. **(b)** Average performance of rats ($n=23$) across a block as a function of risk and flupentixol dose (mean + SEM). **(c)** Average performance of the model (unconstrained simulations).

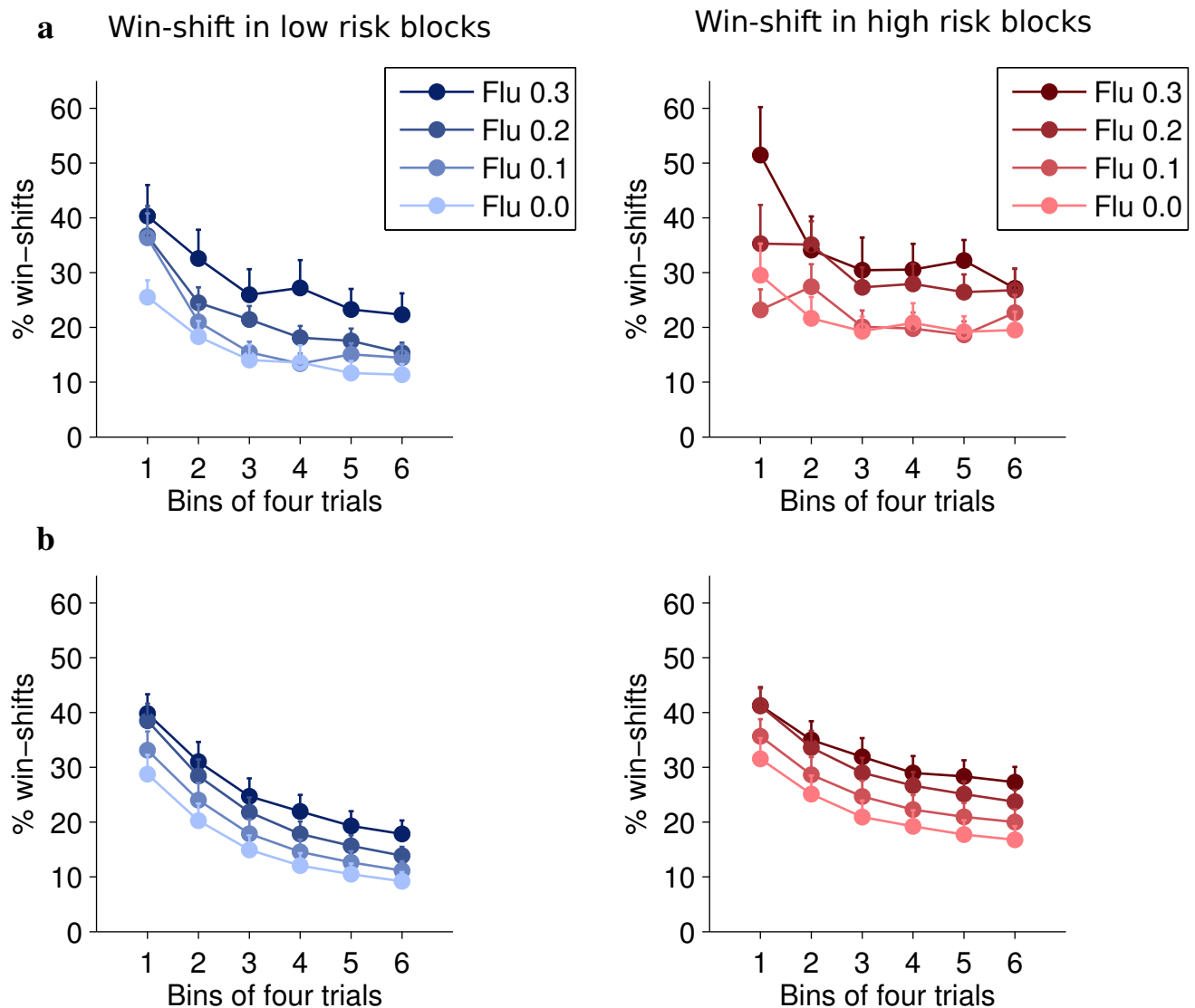


Figure 2. Actual and simulated exploratory choices. **(a)** Average win-shift of rats ($n=23$) across a block as a function of risk and flupenthixol dose (mean + SEM). **(b)** Average win-shift of the model (unconstrained simulations).

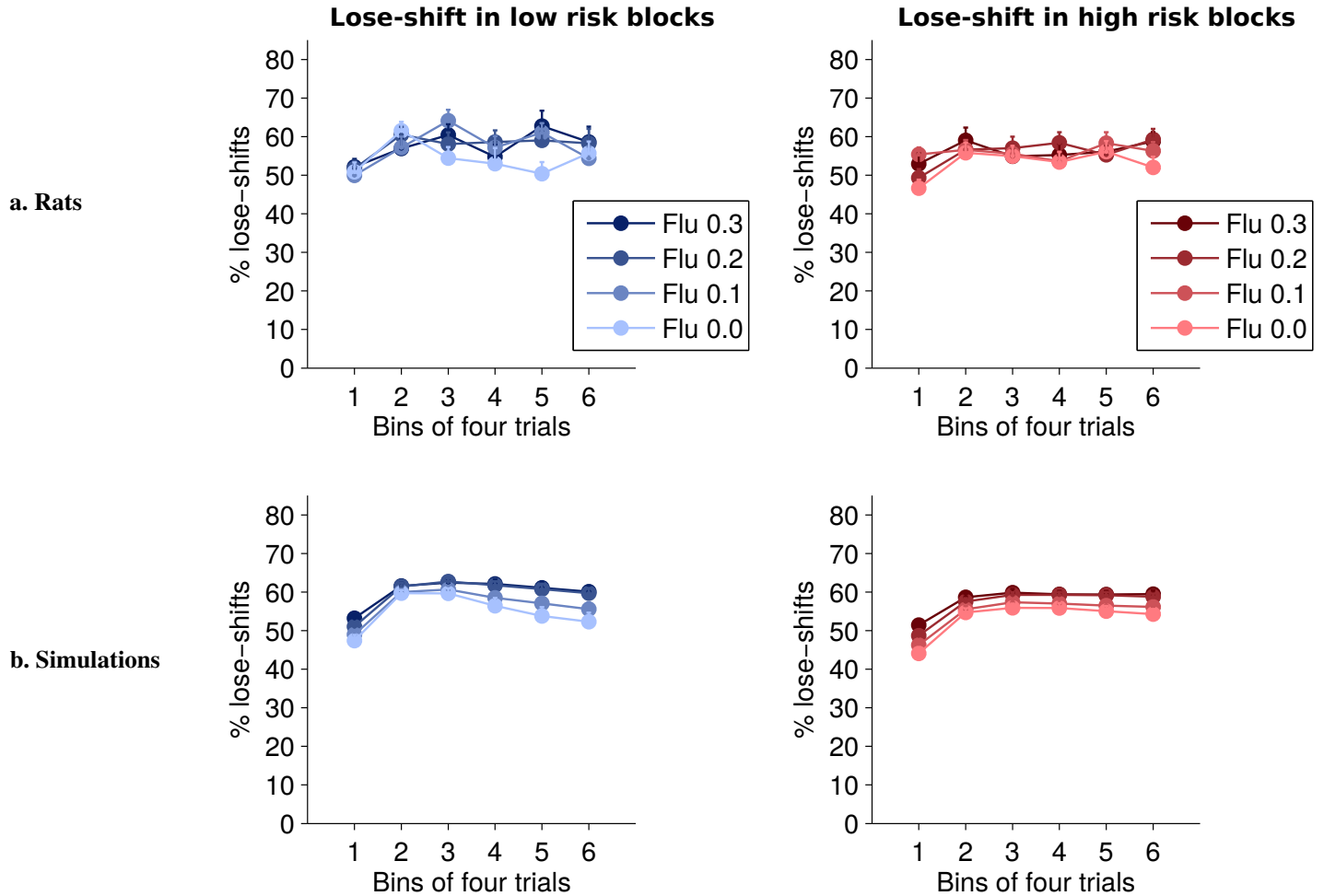


Figure 3. Lose-shift index, i.e. the proportion of trials a rat makes a different choice after an unrewarded trial. **(a)** Average lose-shift of rats ($n=23$) across a block for different risk and dose conditions (mean \pm SEM). This index is not significantly affected by risk, and pharmacological conditions ($p > 0.11$) but a significant trial effect is observed ($F(5,110) = 6.83$, $p < 0.0001$). A post hoc test corrected for multiple comparisons using the Bonferroni procedure indicates that lose-shift in the first bin is significantly smaller than at all other stages of blocks ($p < 0.0158$) except bin number 4 ($p = 0.17$). This effect might be attributable to persistence, as rats probably continue to press the previous target lever unsuccessfully before adjusting their behaviour. **(b)** Simulations of the extended Q-learning model reveal a more complex lose-shift dynamic with a significant decrease in lose-shift also occurring at the end of blocks. In the simulations, significant main effects of dose ($p = 0.0007$) and risk ($p < 0.0001$) are also detected. Similarly to win-shift (see **supp. fig. 1** and **7**), simulations of the standard Q-learning model, which are not shown here, systematically overestimate lose-shift levels.

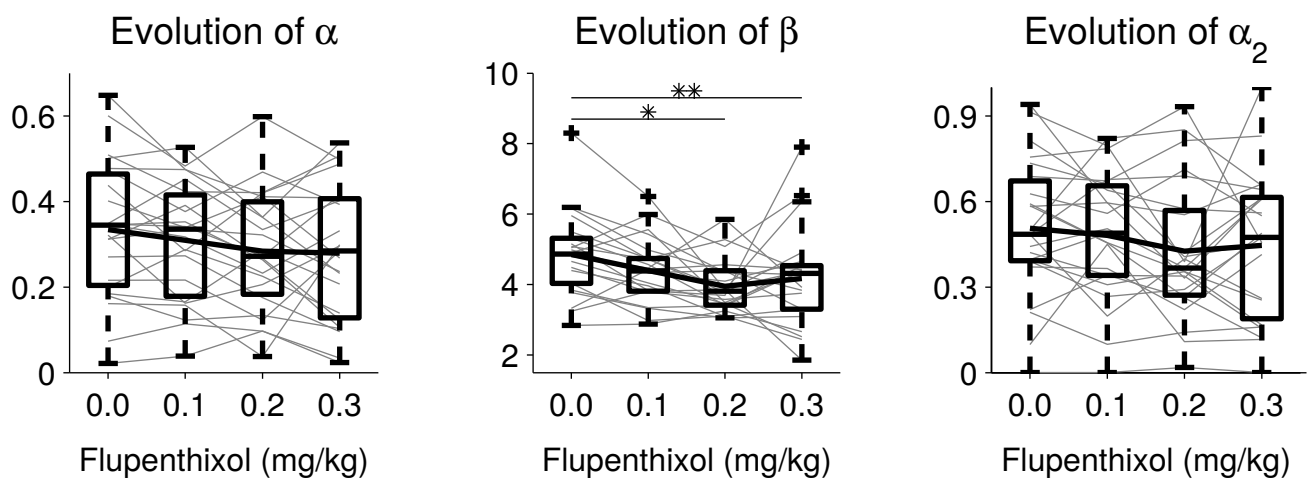


Figure 4. Changes of the different model parameters across flupenthixol doses in all subjects. Gray lines connect parameter values of a single individual while bold lines represent the average. Box plots represent the median, interquartile and furthest observations not considered as outliers. *: $P < 0.05$; **: $P < 0.01$.

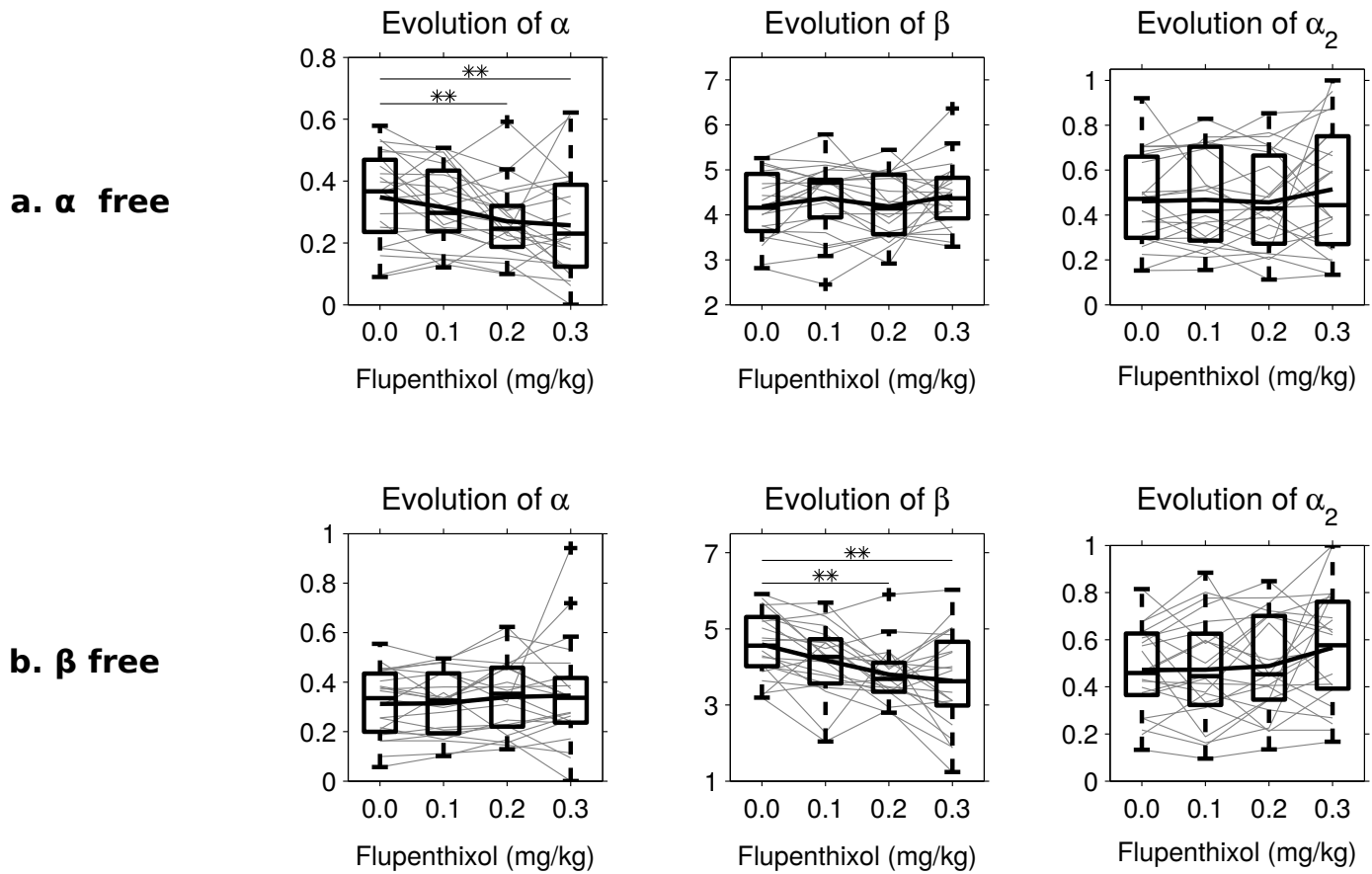


Figure 5. Validation of the optimization procedure. A first version of the extended Q-learning model was optimized with only α allowed to vary between doses and a second version with only β allowed to vary. This way, only one parameter was allowed to capture behavioural changes without compensatory adjustments from other parameters. These optimized parameters were then used to generate artificial sets of data corresponding to a case of learning rate variation or of inverse temperature variation. We then reoptimized a model with all three parameters let free between doses as was done in the paper to ensure that only the correct parameter changed significantly. **(a)** Variation of the different model parameters across flupenthixol doses optimized on the data set obtained from simulations where only α was free. Box plots of median, interquartile and furthest observations not considered as outliers which are represented as crosses. Bold line represents average parameter values. An individual with an extremely high value of β was removed from the graph to make it more readable but was nonetheless included in the statistical analysis. Only α presents a significant dose effect (Friedman Anova test: $\chi^2(3) = 16.1$, $p=0.0011$) as was originally designed. **(b)** Variation of the different model parameters across flupenthixol doses optimized on the data set obtained from simulations where only β was free. An individual with an extreme value of β was also removed from this graph. Only β presents a significant dose effect (Friedman Anova test: $\chi^2(3) = 17.1$, $p = 0.0006$).

584 **Supporting material**

585 **Effects of positive reward prediction errors on exploration**

586 In this section, we show mathematically that a reduction of the amplitude of reward prediction errors[54] directly
 587 translates into changes in random exploration levels in a standard Q-learning model, with or without a forgetting
 588 mechanism. In other words, we prove that inhibition of positive reward prediction errors is mathematically
 589 equivalent to changing the value of the inverse temperature and not to a change in the learning rate.

590 The general formula for RPE is:

$$591 \quad \delta_t = \begin{cases} r - Q_t(a_t) & \text{if a reward } r \text{ is given} \\ 0 - Q_t(a_t) & \text{otherwise} \end{cases} \quad (12)$$

592 We assume that the effect of dopaminergic inhibition is to attenuate positive RPEs:

$$593 \quad \delta_t = r - Q_t(a_t) - f \quad \text{with } 0 \leq f \leq r \quad (13)$$

594 which practically results in a new reward function $r' = g \cdot r$ where $g = \frac{r-f}{r}$ ($0 \leq g \leq 1$). We also assume
 595 the factor g to be constant during the learning process, which is reasonable under pharmacological or genetic
 596 manipulations if relearning is periodically required. This will change the consequences of positive RPEs (when
 597 the reward is present), but not negative RPEs (in the absence of reward). Importantly, it is not equivalent to
 598 reducing learning rate, which would affect the consequences of both positive and negative RPEs. In addition, this
 599 dopaminergic manipulation is assumed not to affect the revision of value for non-selected actions if a forgetting
 600 mechanism is at play (eq. 4).

601 We will now show by induction that under dopaminergic inhibition, the Q-values obtained at any time during
 602 learning Q' , are just scaled-down versions of the original Q-values, Q .

603 Starting with $Q_0=0$ and $Q_0'=0$, $Q_0'=g \cdot Q_0$ is true.

604 After each rewarded step of learning which only affects performed actions

$$605 \quad Q'_{t+1} = Q'_t + \alpha(r' - Q'_t) = g \cdot Q_t + \alpha(g \cdot r - g \cdot Q_t) = g(Q_t + \alpha(r - Q_t)) = g \cdot Q_{t+1} \quad (14)$$

606 After each non-rewarded step of learning, which only affects performed actions,

$$607 \quad Q'_{t+1} = Q'_t + \alpha(0 - Q'_t) = g \cdot Q_t + \alpha(0 - g \cdot Q_t) = g(Q_t + \alpha(0 - Q_t)) = g \cdot Q_{t+1} \quad (15)$$

608 After each forgetting step, which only affects non-performed actions (if applicable)

$$609 \quad Q'_{t+1} = (1 - \alpha_2) Q'_t = (1 - \alpha_2) g \cdot Q_t = g \cdot Q_{t+1} \quad (16)$$

610 Then plugging the scaled Q-values into the softmax function we get:

$$611 \quad P(a_{t+1}=a_i) = \frac{e^{\beta Q'_{t+1}(a_i)}}{\sum_j e^{\beta Q'_{t+1}(a_j)}} = \frac{e^{\beta' Q_{t+1}(a_i)}}{\sum_j e^{\beta' Q_{t+1}(a_j)}} \quad \text{with} \quad \beta' = g \cdot \beta \quad (17)$$

612 Therefore, dopaminergic inhibition, by scaling down the reward function by a factor g , is formally equivalent to
 613 multiplying the inverse temperature by the same factor g . Noticeably, although Q-values (including asymptotic
 614 values) are reduced by dopaminergic inhibition, this effect cannot be mimicked by a simple change in learning
 615 rate. Indeed, changes in learning rate do not affect asymptotic values for a constant reward function as $\delta_t=0$ if
 616 and only if $Q_t=r$.

617 The same demonstration entails that increasing reward size directly reduces random exploration, thereby
 618 favoring exploitation.

619

620 **Supplementary Table 1.** Coefficients of determination of the models under various conditions. For each block,
 621 we labeled the three possible actions as either the target (action 1 in **Supplementary Fig. 4**), the previous target
 622 (action 2; for the very first block of the session, the target of the last block was used so as to ensure all lever-
 623 label combinations were equally represented) or the remaining lever (action 3). For each rat ($n=23$), we then
 624 averaged the experimental probabilities of each action for bins of four trials per block, and compared them to the
 625 corresponding average theoretical probabilities as determined by current Q-values plugged into the softmax
 626 function. This procedure was applied to separate doses, separate risk levels and to the entire experiment as
 627 reported in the table. In every case, the model explains a high proportion of the individual variance.

628

629

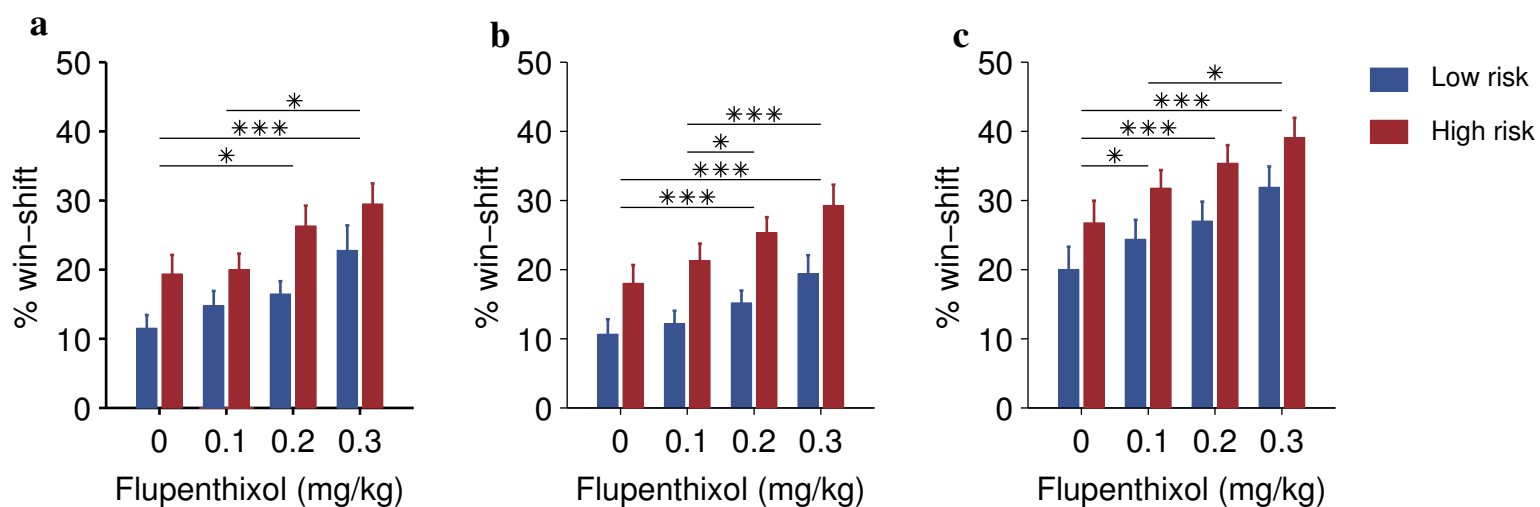
630

631

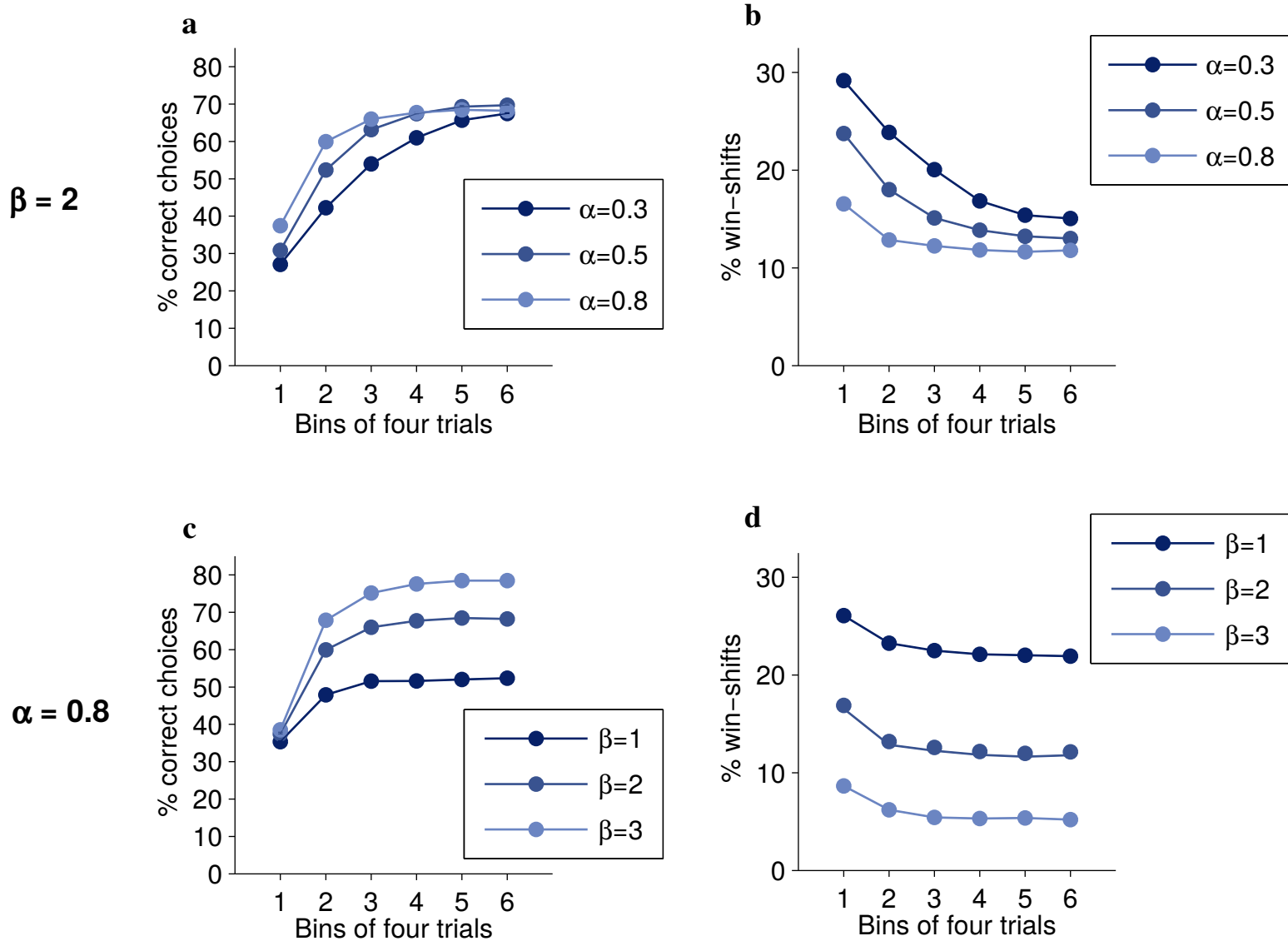
632

633

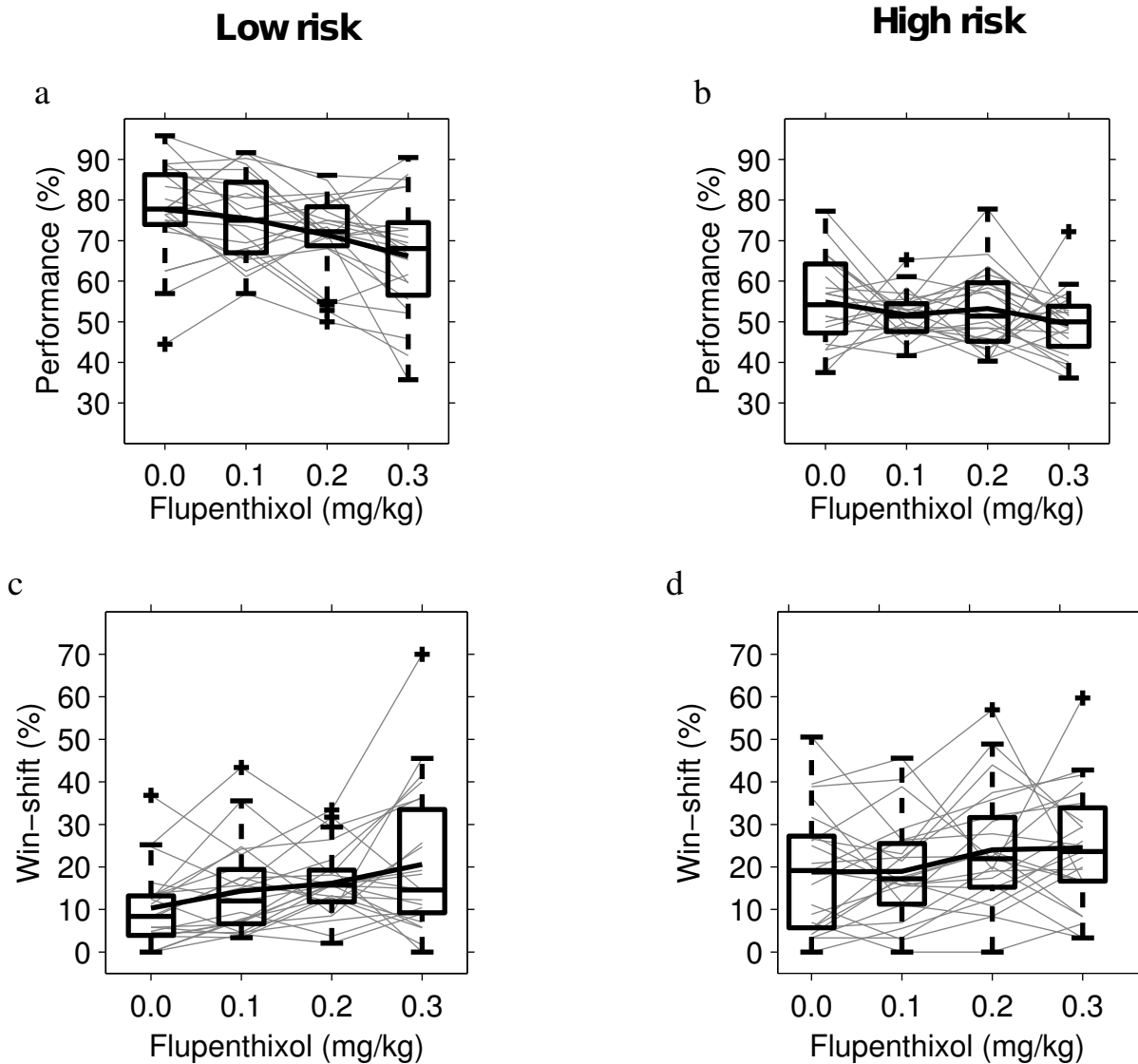
		LR	Risk HR	All
	0	0.9635	0.8412	0.9572
Flupenthixol	0.1	0.9549	0.8161	0.944
Dose	0.2	0.9245	0.8104	0.935
(mg/kg)	0.3	0.8549	0.677	0.8751
	All	0.929	0.7882	0.9314



Supplementary Figure 1. Average win-shift (mean + SEM) in the last eight trials of blocks. Post hoc tests were corrected using Bonferroni procedure (* : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$). **(a)** The average win-shift of rats ($n=23$) is greater in high risk blocks compared to low risk blocks ($F(1,22) = 41.819$, $p < 0.0001$), and also increases with the dose of flupenthixol ($F(3,66) = 13.47$, $p < 0.0001$) and by pharmacological condition ($p < 0.0001$) suggesting that rats explore more when dopamine is inhibited. **(b)** Simulations of the extended Q-learning model replicate experimental results with similar risk ($F(1,22)=165.16$, $p < 0.0001$) and flupenthixol effects ($F(3,66) = 18.605$, $p < 0.0001$) on average win-shift. Furthermore, when directly confronted to experimental data using a 3-factor repeated measures Anova including a model factor, there is no significant effect of simulations ($F(1,22)=1.2714$, $p=0.27$; no significant interaction factor involving model $p > 0.14$). **(c)** Simulations of the simple Q-learning model show significantly increased levels of win-shift compared to the original data ($F(1,22) = 45.562$, $p < 0.0001$). Nonetheless, the effects of risk and dose are also captured accurately (no significant interaction between model and risk and/or dose detected, $p > 0.61$).

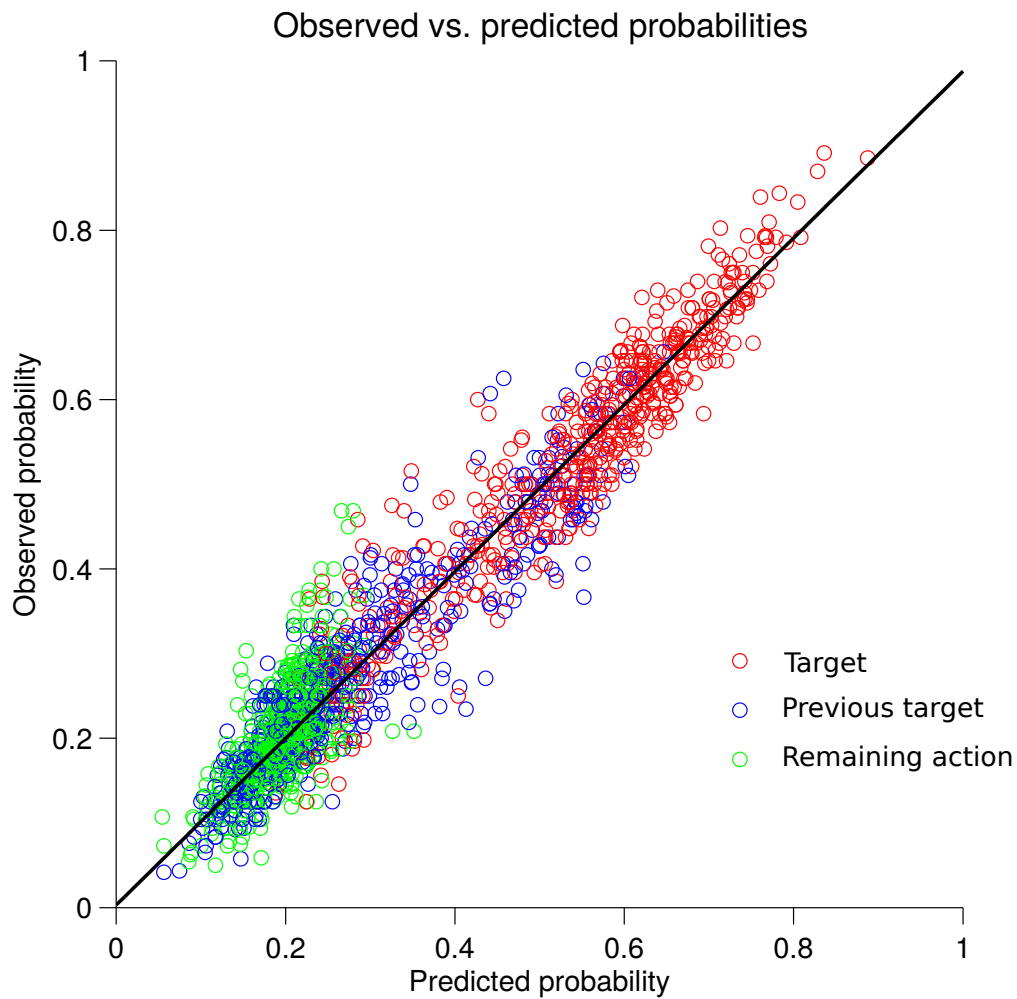


Supplementary Figure 2. Illustration of the effects of varying either the learning rate or the exploration parameter on performance and win-shift curves based on simulations of the standard Q-learning model in low risk blocks. The range of parameters was chosen for illustrative purposes and does not reflect values actually found when the model was optimized on the experimental data. **(a)** Under different learning rates, performance curves initially increase at different rates but end up converging to similar asymptotic levels. **(b)** Win-shift curves start at different levels and decrease at different rates and eventually converge. **(c)** Under different exploration rates, performance curves converge toward different asymptotes. **(d)** Win-shift curves lie parallel to one another without ever converging.



Supplementary Figure 3. Experimental effects of flupenthixol on final performance and win-shift levels. Gray lines represent individual animals, boxes show the median, interquartile range (box edges) and most extreme data points not considered as outliers (whiskers), and crosses are suspected outliers. Bold lines represent the mean.

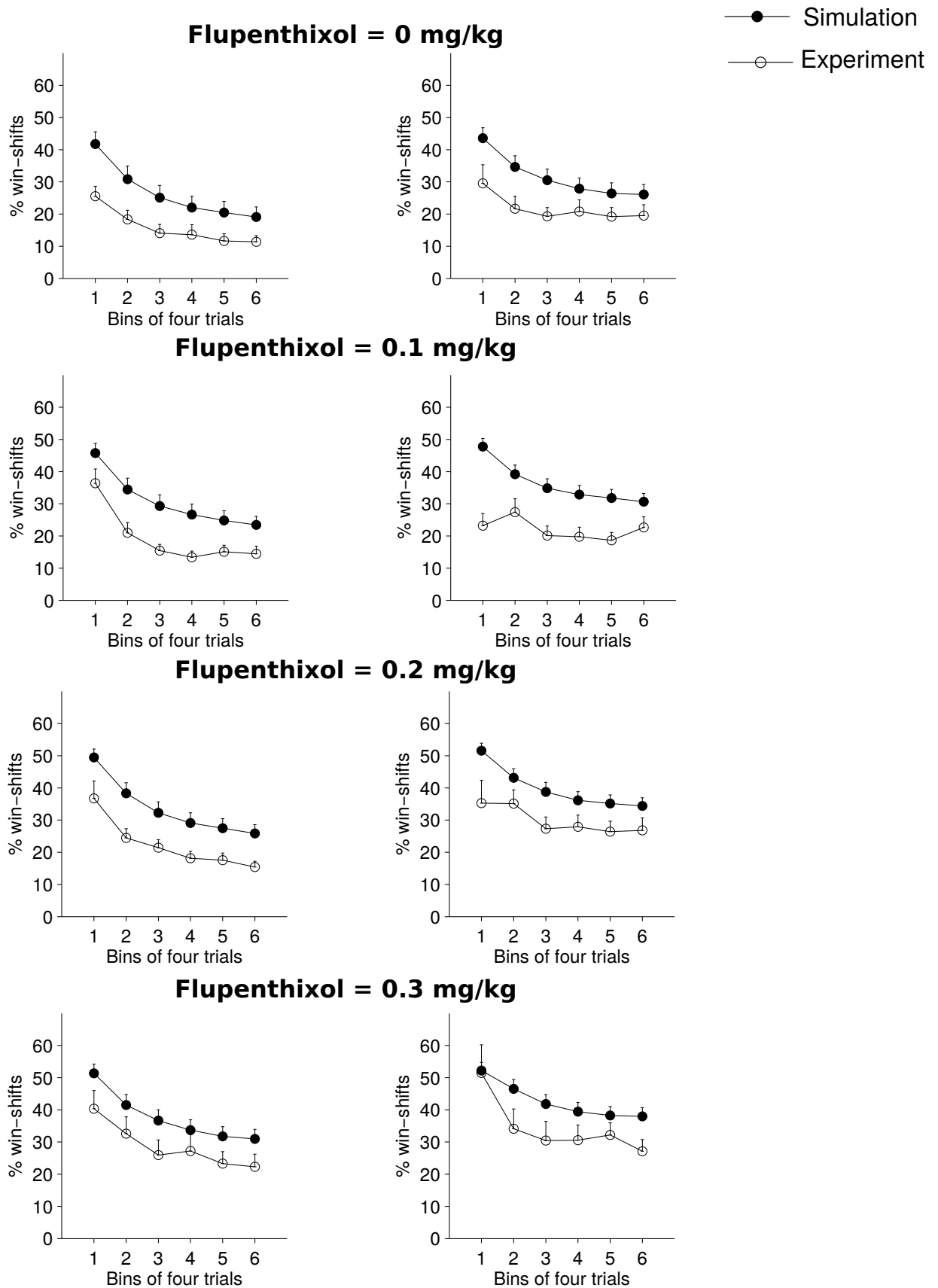
(a) & (b) Effect of flupenthixol on average performance levels in the last six trials of low and high risk blocks. A significant negative effect of flupenthixol appears ($F(3,66)=8.85$, $p<0.0001$) and there is no interaction with risk ($F(3,66)=1.96$, $p=0.13$). Post hoc t-tests revealed that average performance under 0.3 mg/kg was significantly smaller than all other doses (highest $p=0.020$) and that average performance under 0.2 mg/kg was significantly worse than under 0 mg/kg ($p=0.027$). **(c) & (d)** Effect of flupenthixol on average win-shift in the last six trials of low and high risk blocks. A significant dose effect was detected ($F(3,66)=8.76$, $p<0.0001$) without any interaction with risk ($F(3,66)=0.77$, $p=0.51$). Post hoc t-tests showed that average win-shift increased for 0.3 mg/kg of flupenthixol when compared to 0 and 0.1 mg/kg (highest $p=0.0017$ and for 0.2 mg/kg when compared to 0 mg/kg ($p=0.0021$)).



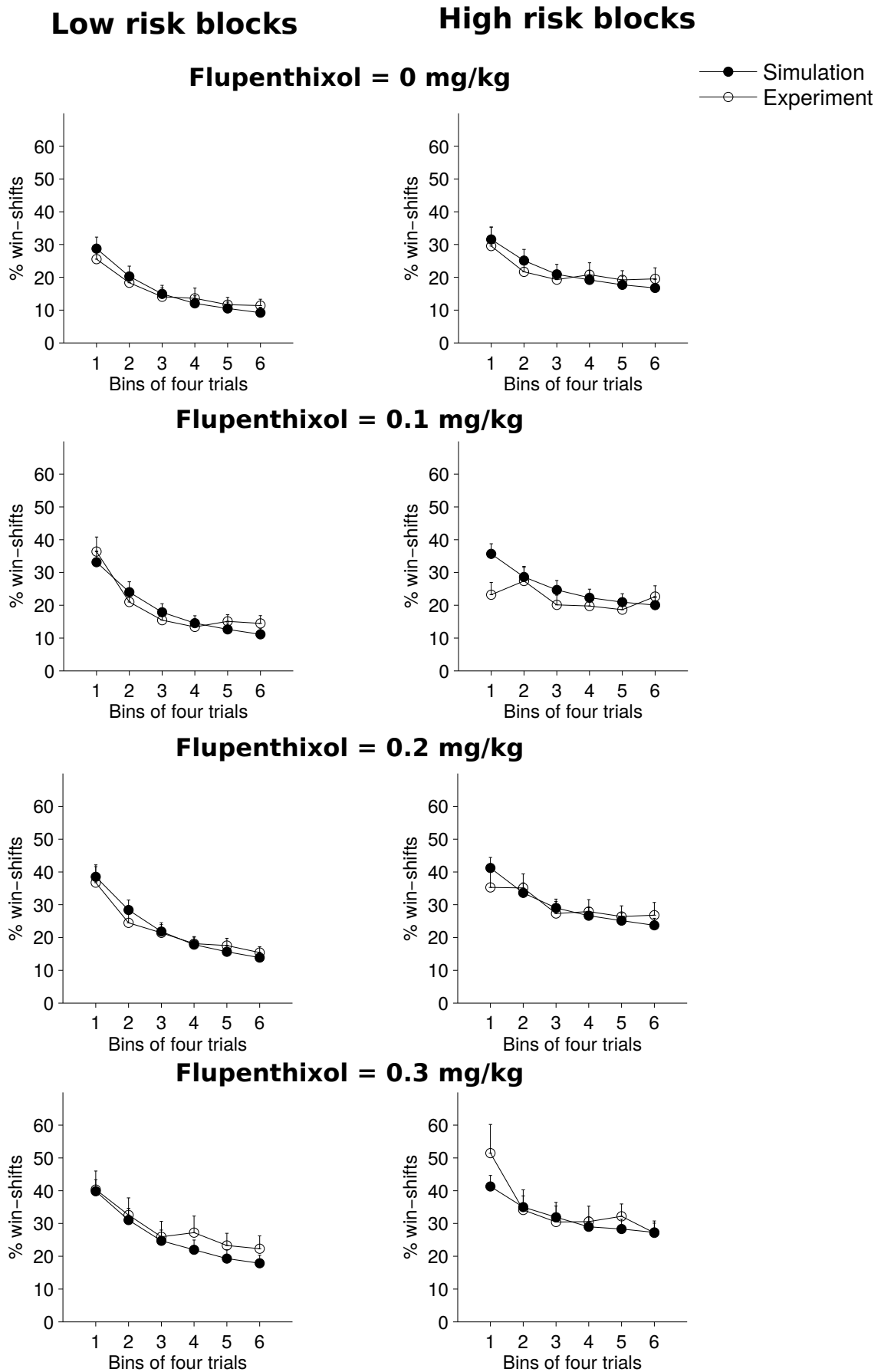
Supplementary Figure 4. Relationship between predicted and observed probabilities of the three possible actions within a block. Predicted probability is given by the softmax function averaged in a given bin of four trials within a block. Observed probability is based on the number of times this specific action was chosen in this bin. Each data point represents the average of blocks for a given combination of rat, bin, risk level, dose and response type. Response types are identified by colors representing correct (target, red), previously correct but now incorrect (previous target, blue) or other actions (green). The same linear regression curve ($y = 0.9536x + 0.0174$) is shown in black) fits all three response types.

Low risk blocks

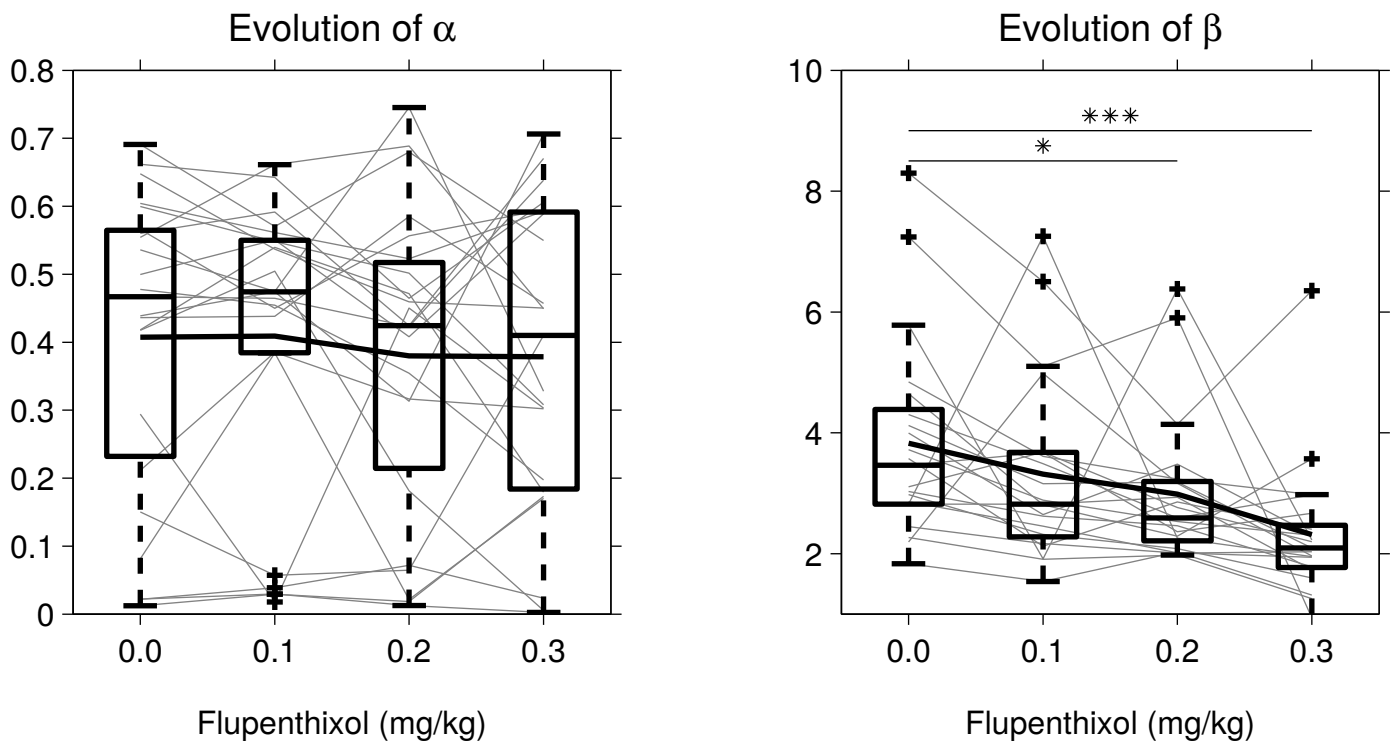
High risk blocks



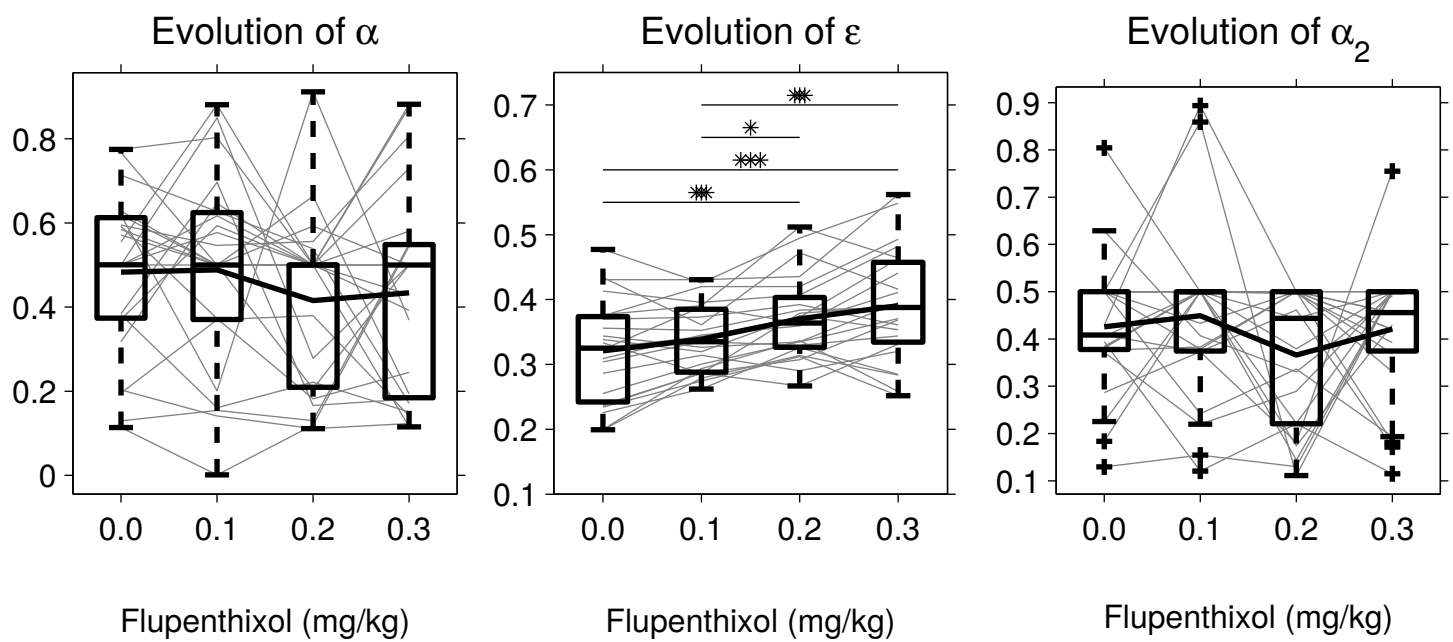
Supplementary Figure 5: Win-shift (mean + s.e.m.) of simulations of the standard Q-learning model plotted against the experimental data for the different risk and pharmacological conditions. Contrary to the forgetting model (see **Supplementary Fig. 6**), this model was unable to reproduce this key aspect of behaviour and shifted from correct rewarded actions far more frequently than the subjects.



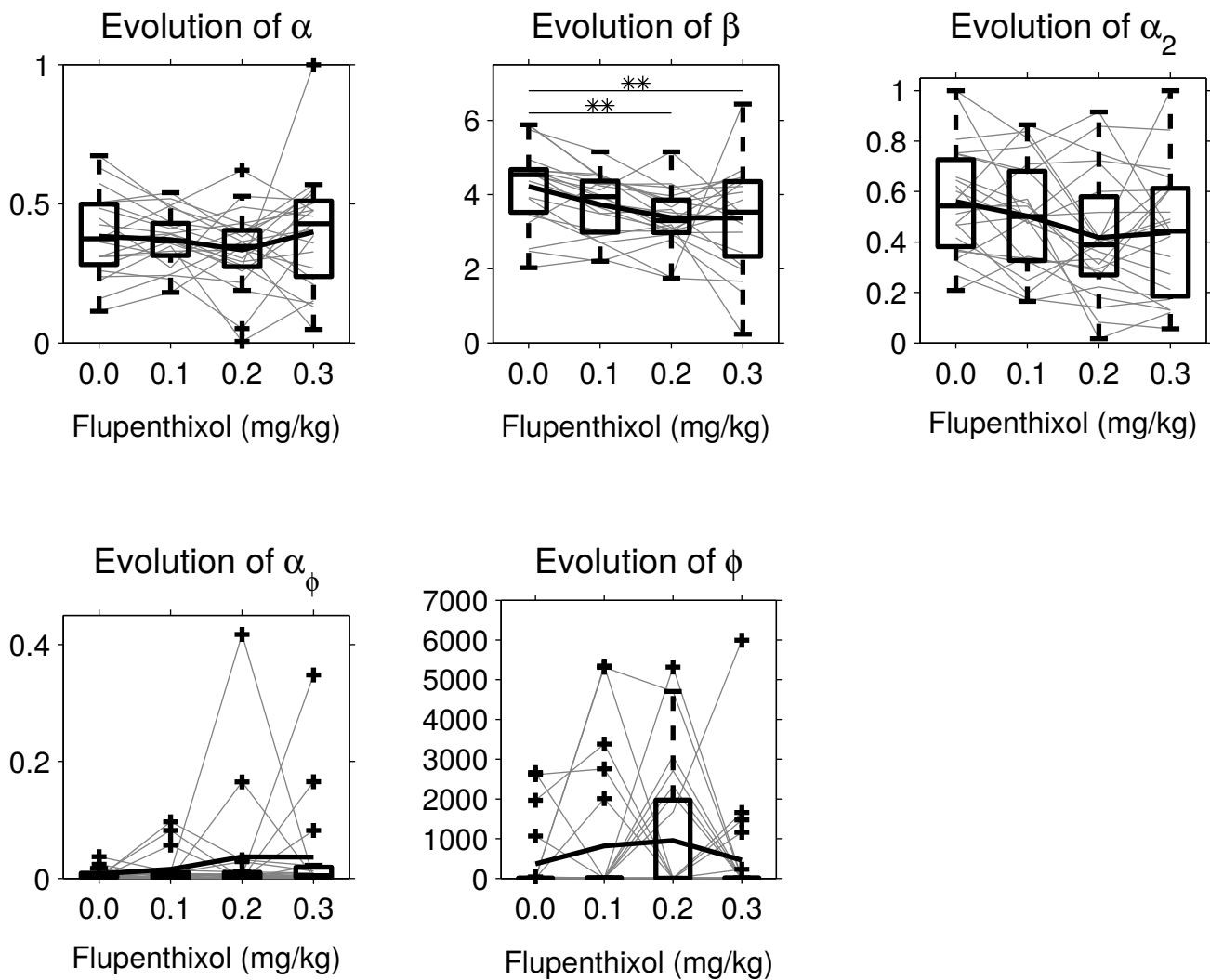
Supplementary Figure 6: Win-shift (mean + s.e.m.) of simulations of the Q-learning model extended with a forgetting mechanism plotted against the experimental data for different doses of flupenthixol and both risk levels. This figure is in fact a different representation of **Fig. 2** for comparison with **Supplementary Fig. 5**.



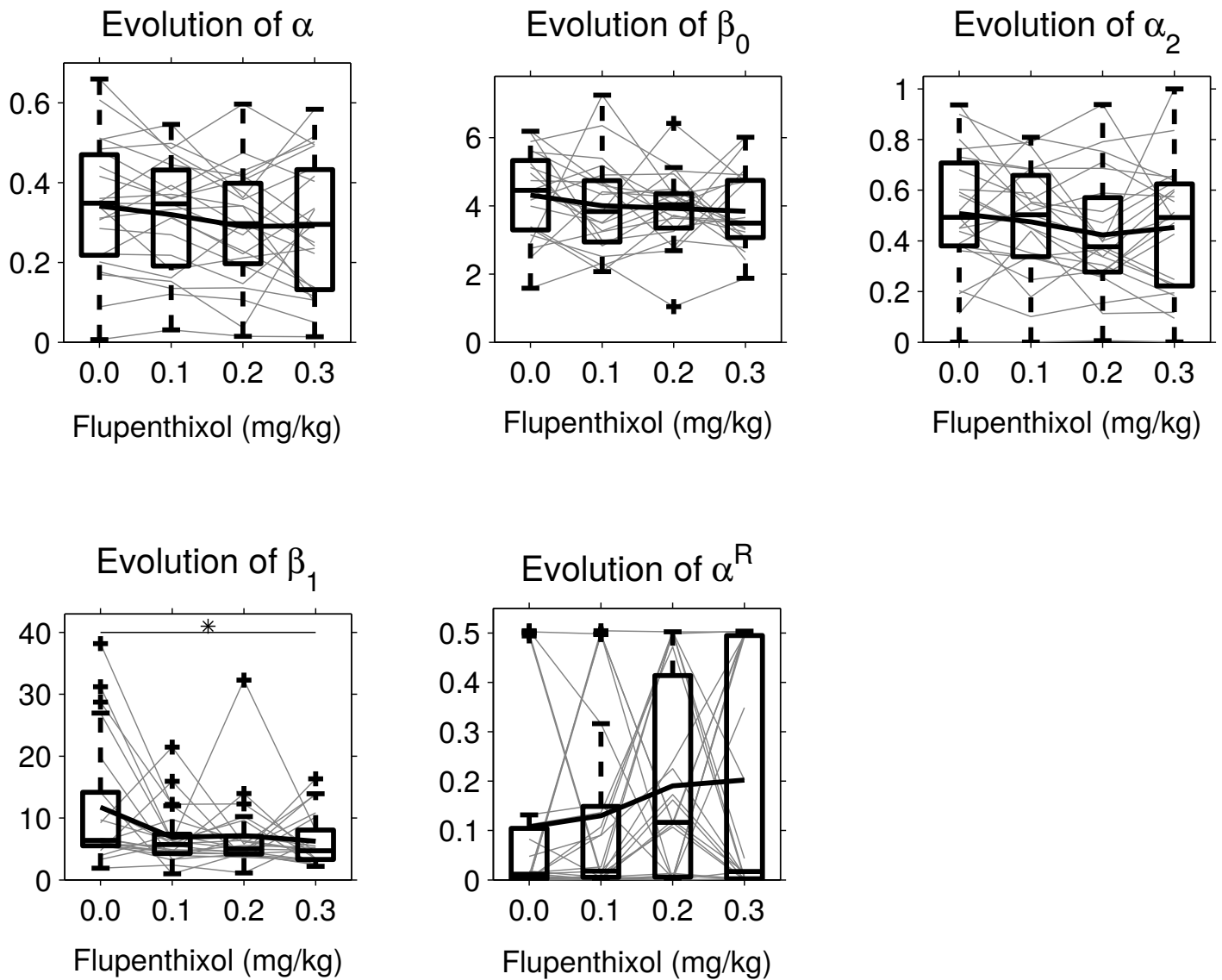
Supplementary Figure 7. Variations of the parameters of the standard Q-learning model. Gray lines connect parameter values of a same individual. Box plots of median, interquartile and furthest values not considered as outliers represented as crosses. Bold lines plot average parameter values. As with the extended model, there is no significant dose effect on α (Friedman Anova test: $\chi^2(3) = 1.90$, $p = 0.59$) whereas β is significantly affected ($\chi^2(3) = 23.5$, $p < 0.0001$).



Supplementary Figure 8. Variations of the parameters of the ϵ -greedy Q-learning model. Gray lines plot the parameter variations of a single individual while bold line plots average parameter values. Box plot of median, interquartile and extreme values not considered as outliers represented as crosses. As with the softmax version of this model (**Fig. 4**), the only parameter affected by flupenthixol is the one responsible for controlling exploration, ϵ (Friedman Anova test: $\chi^2(3) = 34.7$, $p < 0.0001$).



Supplementary Figure 9. Variations of the parameters of the uncertainty bonus model. Grey lines represent single individuals and bold lines the average parameter variation. Box plots of median, interquartile and most extreme values not considered as outliers. As with previous models, the only parameter affected by flupenthixol is β (Friedman Anova test: $\chi^2(3) = 16.3$, $p = 0.0010$) which controls random exploration. Exploration targeted at uncertain options is not affected.



Supplementary Figure 10. Variations of the parameters of the meta-learning model. Grey lines correspond to individual rats while the bold lines represent the average of parameter values. Box plots of median, interquartile and most extreme values which are not outliers. The only parameter affected by flupenthixol is β_1 although the statistical effect is in fact quite modest (Friedman Anova test: $\chi^2(3) = 9.3$, $p = 0.026$).