University of Reading

Dopamine blockade impairs the exploration-exploitation trade-off in rats

Article

Published Version

Creative Commons: Attribution 4.0 (CC-BY)

Open Access

Cinotti, F. ORCID: https://orcid.org/0000-0003-2921-0901, Fresno, V., Aklil, N., Coutureau, E., Girard, B., Marchand, A. R. and Khamassi, M. (2019) Dopamine blockade impairs the exploration-exploitation trade-off in rats. Scientific Reports, 9. 6770. ISSN 2045-2322 doi: 10.1038/s41598-019-43245-z Available at https://centaur.reading.ac.uk/120880/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1038/s41598-019-43245-z

Publisher: Nature Publishing Group

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

University of Reading

Central Archive at the University of Reading

Reading's research outputs online

SCIENTIFIC REPORTS

OPEN

Dopamine blockade impairs the exploration-exploitation trade-off in rats

Received: 16 October 2018
Accepted: 18 April 2019
Published online: 01 May 2019

François Cinotti[1], Virginie Fresno[2,3], Nassim Aklil[1], Etienne Coutureau[2,3], Benoît Girard[1], Alain R. Marchand[2,3] & Mehdi Khamassi[1]

In a volatile environment where rewards are uncertain, successful performance requires a delicate balance between exploitation of the best option and exploration of alternative choices. It has theoretically been proposed that dopamine contributes to the control of this exploration-exploitation trade-off, specifically that the higher the level of tonic dopamine, the more exploitation is favored. We demonstrate here that there is a formal relationship between the rescaling of dopamine positive reward prediction errors and the exploration-exploitation trade-off in simple non-stationary multi-armed bandit tasks. We further show in rats performing such a task that systemically antagonizing dopamine receptors greatly increases the number of random choices without affecting learning capacities. Simulations and comparison of a set of different computational models (an extended Q-learning model, a directed exploration model, and a meta-learning model) fitted on each individual confirm that, independently of the model, decreasing dopaminergic activity does not affect learning rate but is equivalent to an increase in random exploration rate. This study shows that dopamine could adapt the exploration-exploitation trade-off in decision-making when facing changing environmental contingencies.

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

Dopamine holds a fundamental place in contemporary theories of learning and decision-making. The temporal evolution of phasic dopamine signals across learning has been extensively replicated, and is most of the time considered as evidence of a role in learning[6–8], but see alternative views in Coddington *et al.*[9]. Dopamine reward prediction error (RPE) signals have been identified in a variety of instrumental and Pavlovian conditioning tasks[10–13]. They affect plasticity and action value learning in cortico-basal networks[14–16] and have been directly related to behavioral adaptation in a number of decision-making tasks in humans, non-human primates[17] and rodents[18–21]. Accordingly, it is commonly assumed that manipulations of dopamine activity affect the rate of learning, but this could represent a misconception.

Besides learning, the role of dopamine in the control of behavioral performance is still unclear. Dopamine is known to modulate incentive choice (the tendency to differentially weigh costs and benefits)[22,23], and risk-taking behavior[24], as well as other motivational aspects such as effort and response vigour[25]. Because dopamine is one of the key factors that may encode success or uncertainty, it might modulate decisions by biasing them toward options that present the largest uncertainty[26,27]. This would correspond to a "directed" exploration strategy[5,28,29]. Alternatively, success and failure could affect tonic dopamine levels and control random exploration of all options, as recently proposed by Humphries *et al.*[30]. This form of undirected exploration, which is often difficult

[1]Institut des Systèmes Intelligents et de Robotique, Sorbonne Université, CNRS, F-75005, Paris, France. [2]CNRS, Institut de Neurosciences Cognitives et Intégratives d'Aquitaine (INCIA, UMR 5287), Bordeaux, France. [3]Université de Bordeaux, INCIA, Bordeaux, France. François Cinotti and Virginie Fresno contributed equally.Alain R. Marchand and Mehdi Khamassi jointly supervised this work. Correspondence and requests for materials should be addressed to M.K. (email: mehdi.khamassi@upmc.fr)

to distinguish from performance, may be viewed as "noise" in the choice process[31–33]. It is nevertheless known as a classical and efficient exploration strategy in machine learning[31]. Previous computational analyses of behavioral data in stochastic tasks have yielded mixed results, some suggesting a promotion of random[34] or directed[27] exploration by dopamine with possibly an effect on learning[27] others a reduction of random exploration[30,35,36], and yet others an effect on learning only[37].

In the present study, we first show formally that under fairly general assumptions, any manipulation that reduces the magnitude of dopamine positive reward prediction errors does not change the learning rate but instead changes the level of random exploration. We then proceed to test this idea experimentally in rats while applying a variety of computational models to the behavioral data. To dissociate learning and performance components, probabilistic tasks where the best option changes with time (known in Machine Learning as non-stationary bandit tasks) are particularly appropriate, because they require periodical exploration and relearning phases and are amenable to computational modeling using well-characterized reinforcement learning methods[38]. In this work, we develop such a 3-armed bandit task in rats with varying levels of uncertainty to investigate how dopamine controls the exploration level within an individual. We do this by examining the effects of dopamine blockade on learning and performance variables following injection of various doses of flupenthixol, a D1/D2 receptor antagonist which should cause a reduction in the effect of both tonic and phasic dopamine activity, in different sessions. This pharmacological intervention hopefully mimics, at least under low doses, natural variations in dopaminergic levels and can give us insight into their functional significance. We follow by replicating these data with a reinforcement learning model (Q-learning) extended with forgetting and verify our conclusions on a variety of alternative models such as a directed exploration model, an ε-greedy random exploration model, and a meta-learning model. This allows us to explicitly distinguish learning from exploration variables, and to show that dopamine activity is specifically involved in controlling the level of random exploration rather than the learning rate.

Results

Mathematical relationship between reward prediction errors and random exploration-exploitation. We first give here a formal demonstration (Supporting Material) that in a Q-model such as the one we applied to our task, a reduction of the amplitude of phasic dopaminergic responses to rewards directly translates into an increase in random exploration levels. In other words, it is mathematically equivalent to changing the value of the inverse temperature, and has no effect on the learning rate parameter. Briefly, in this demonstration, we assume that the value of the reward is pharmacologically reduced, resulting in decreased positive, but not negative, prediction errors[39]. On each trial, the Q-value of the performed action is revised in proportion to the RPE, so it is exactly a fraction of what it would be in the absence of pharmacological manipulation. For non-performed actions, if there is forgetting, the Q-value decreases in proportion to the Q-value itself, which preserves proportionality. As a result, throughout the learning process, all Q-values are downscaled in the same proportion as the reward. When these values are plugged into the softmax process, the result is exactly equivalent to a decrease of the inverse temperature, again in the same proportion. The learning and forgetting rates are not affected in any way. This result shows that under fairly general conditions the effects of a pharmacological manipulation of dopamine-dependent learning should be described as changes in exploration rate rather than as changes in learning rate. Indeed, manipulation of these two factors predicts distinct behavioral profiles: under different learning rates, both performance and win-shift curves (see Methods) should differ at early stages of blocks, but then converge to similar levels (Supplementary Fig. 1a,b). Conversely, when only the inverse temperature differs, performance curves should tend toward different asymptotes while win-shift curves should shift downwards as a whole when this parameter increases (Supplementary Fig. 1c,d).

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

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

We then looked at exploration, indexed as a first approximation by a win-shift index which only includes shifts from the current best lever (Fig. 2a). This index decreased within blocks ($F(5,85) = 27.7$, $p < 0.0001$) but increased with risk ($F(1,17) = 25.7$, $p < 0.0001$). Dopaminergic blockade dose-dependently elevated win-shift at all stages ($F(3,51) = 14.5$, $p < 0.0001$), without interacting with trial or risk ($p > 0.11$). Because win-shift in the early stages of blocks may not reflect exploration, but rather a return to a previously reinforced lever, we also limited our analysis of win-shift to the last 8 trials when performance has stabilized, meaning that the correct lever has been identified. In this case also there was a significant dose effect (Fig. 3a). Another index of shifting behavior, lose-shift (Fig. 4a), which may denote a correction strategy, was not significantly affected by the pharmacological condition ($F(3,66) = 1.77$, $p = 0.16$), possibly because of a ceiling effect.

We then examined whether the effects of flupenthixol on both win-shift and performance resulted from a negative effect on learning rate, or on inverse temperature[40]. In the present task, these two factors predict distinct behavioral profiles[41], in particular when considering asymptotic behavior (Supplementary Fig. 1). A cursory look at behavioral data in Figs 1 and 2 pleads in favour of a change in exploration rate, as individual performance levels during the last six trials of blocks were significantly affected by flupenthixol dose ($F(3,66) = 8.85$, $p < 0.0001$) (Supplementary Fig. 2a,b) instead of converging. Similarly, the win-shift rate in the last 6 trials of blocks was

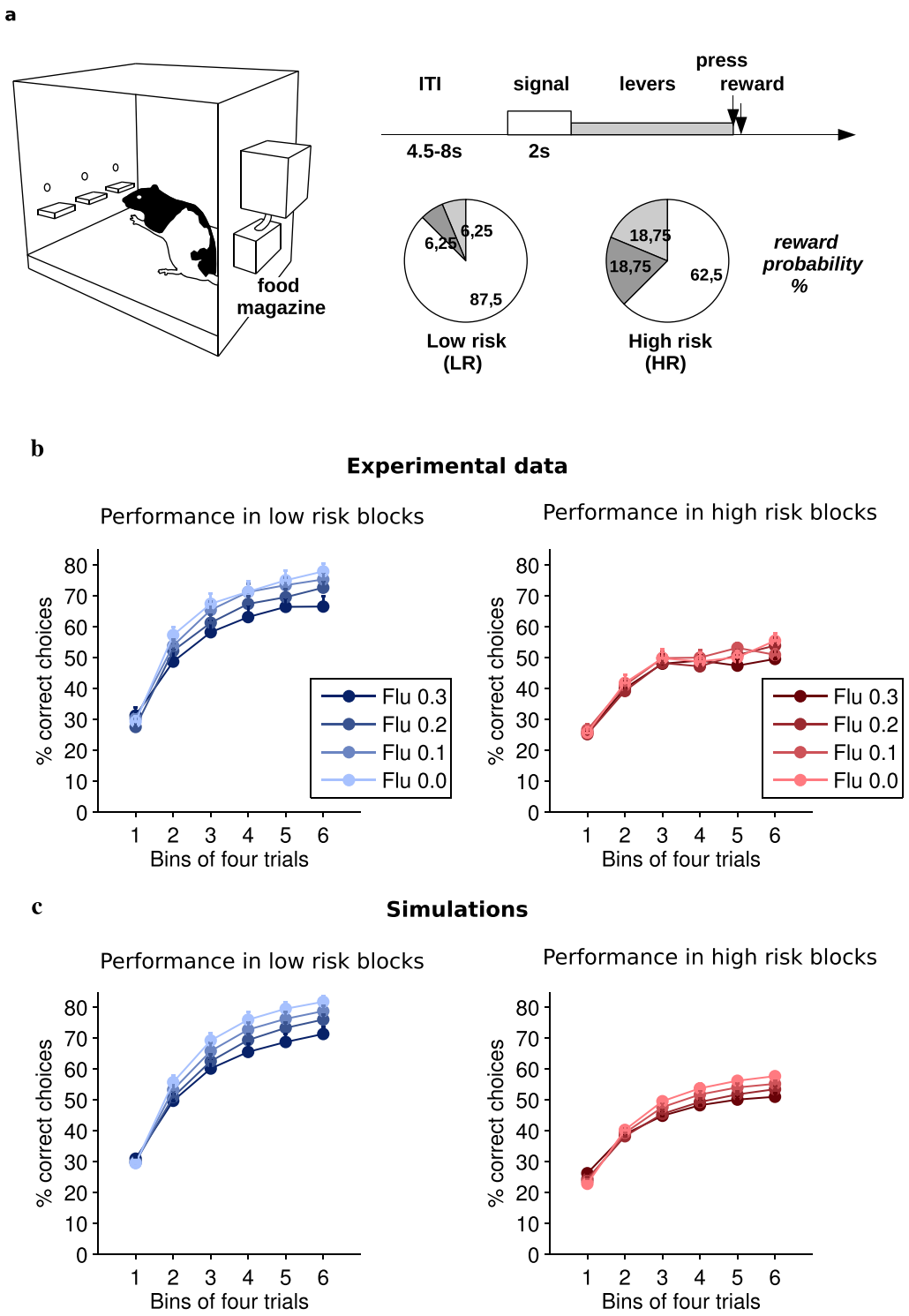

Figure 1. (**a**) Outline of the experimental task. (**b**) Average performance (mean + sem) of rats (n = 23) across a block as a function of risk and flupenthixol dose. (**c**) Average performance (mean + sem) of model simulations.

significantly affected by flupenthixol dose ($F(3,66) = 8.76$, $p < 0.0001$) (Supplementary Fig. 2c,d). This persistent difference in the curves at the end of each learning phase is strongly suggestive of a change in the proportion of random choices but could perhaps be explained by the limited number of trials in a block which might prevent full convergence. We therefore turned towards a computational approach of this issue.

Flupenthixol alters random exploration in simulated rats. We then examined how well a Q-learning model extended with a forgetting mechanism was able to account for behavioral data (see Methods), which is a prerequisite to drawing any conclusions from a parameter analysis. Average action values derived from the

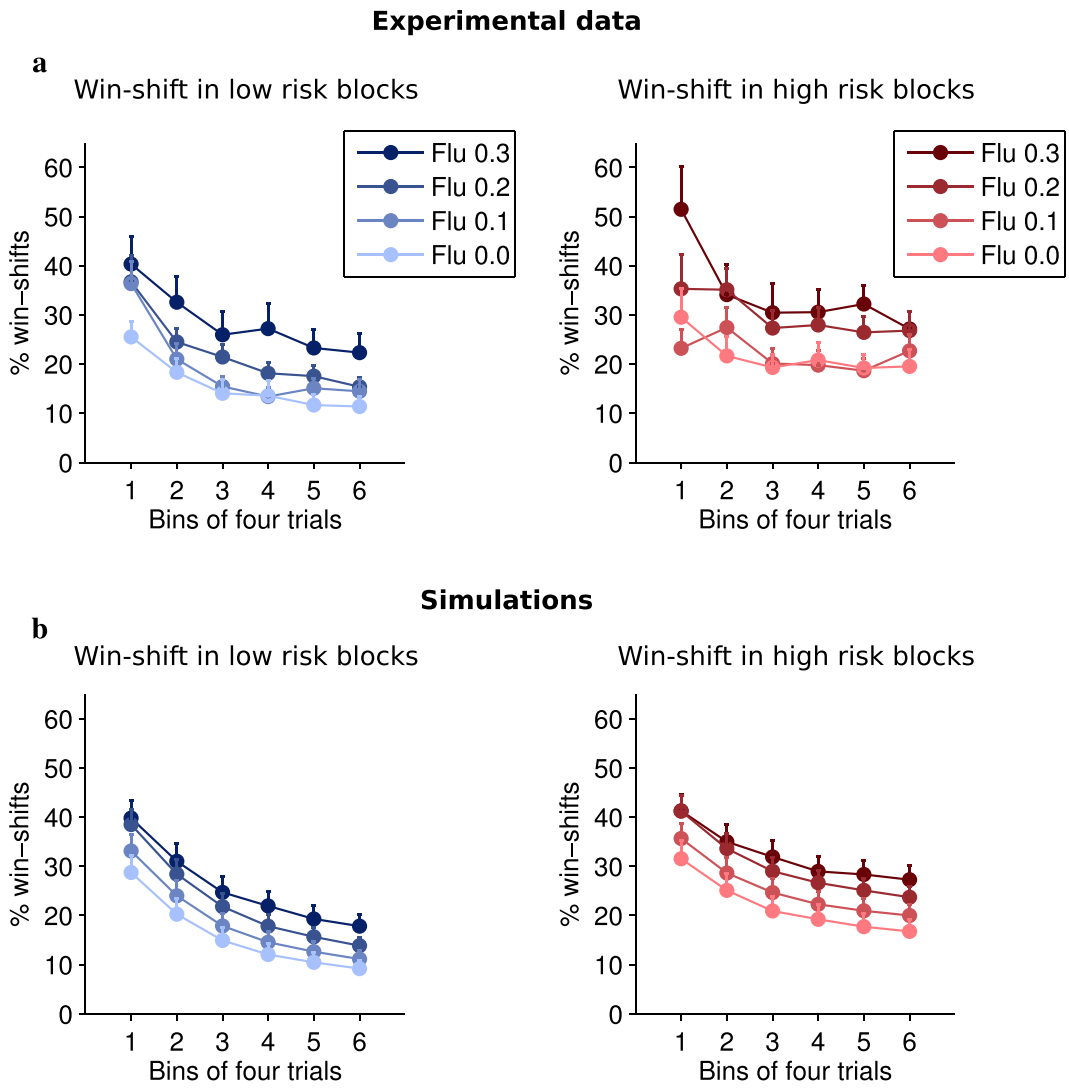

Figure 2. (**a**) Average proportion of win-shifts (mean + sem) of rats across a block as a function of risk and flupenthixol dose. (**b**) Average win-shift of model simulations.

model, individually fitted to each rat for each dose of flupenthixol and constrained to the rat's actual choices, predicted a very high proportion of the variance of individual rats' choices, (Supplementary Table 1 and Fig. 3). Moreover, unconstrained simulated data generated using these optimized parameters were highly similar to the actual behavior of the rats (Figs 1c, 2b, 3b, 4b, and Supplementary Fig. 5). When experimental and simulated data were pooled together for repeated-measures ANOVA, there was no significant main effect of simulations on within-block evolution of either performance (Fig. 1c, $F(1,22) = 4.27$, $p = 0.051$), win-shift (Fig. 2b, $F(1,17) = 0.006$, $p = 0.94$) or lose-shift (Fig. 4b, $F(1,22) = 1.63$, $p = 0.21$). However, a significant interaction between simulations and risk did emerge in the case of performance ($F(1,22) = 4.42$, $p = 0.047$) and there was also a significant interaction between trials and simulations for all three indicators ($F(5,110) = 6.98$, $p < 0.0001$ for performance; $F(5,85) = 2.90$, $p = 0.018$ for win-shift; $F(5,110) = 3.61$, $p = 0.005$ for lose-shift). Crucially however, no interaction involving simulations and flupenthixol dose could be detected (smallest $p = 0.07$). Moreover, when analyzed separately from experimental data, simulated data did in fact replicate the effects of flupenthixol on performance and win-shift, which both present significant main effects of flupenthixol ($F(3,66) = 5.23$, $p = 0.0027$ and $F(3,66) = 5.69$, $p = 0.0016$ for performance and win-shift respectively). Simulated lose-shift, however, was also sensitive to flupenthixol ($F(3,66) = 14.4$, $p < 0.0001$), an effect which is undetected in the experimental data and can be attributed to the much greater number of simulations compared to the number of rats. Having proved that this model replicates experimental results sufficiently well, we can then use it confidently to decipher the effects of flupenthixol on behaviour.

Model optimization dissociates random exploration from learning. To disentangle the effects of flupenthixol on learning versus performance, we then examined the values of the different parameters - α, β and α_2 - across pharmacological conditions (Fig. 5). Flupenthixol had no discernible effect on the learning rate α (Friedman ANOVA $\chi^2(3) = 4.04$, $p = 0.26$) or on the forgetting rate α_2 ($\chi^2(3) = 1.38$, $p = 0.71$), but clearly

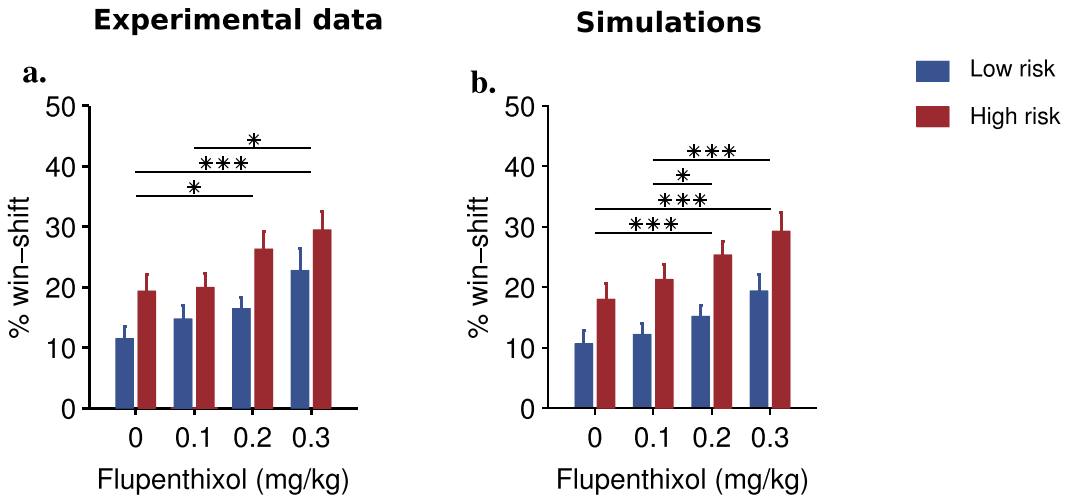

Figure 3. Average win-shift (mean + sem) in the last eight trials of blocks. *Post hoc* Bonferroni tests were performed when appropriate (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). **(a)** The average win-shift of rats is significantly greater in high risk than low risk blocks ($F(1,22) = 41.8$, $p < 0.0001$), and also increases with flupenthixol dose ($F(3,66) = 13.5$, $p < 0.0001$) suggesting that rats explore more when dopamine is inhibited. (**b**) Simulations of the extended Q-learning mode replicate experimental results with similar risk ($F(1,22) = 165.2$, $p < 0.0001$) and flupenthixol main effects ($F(3,66) = 18.6$, $p < 0.0001$). Furthermore, when confronting the two datasets directly through a repeated-measures ANOVA, there is no significant effect involving the simulations factor ($p > 0.14$ for all main and possible interaction with risk and dose effects).

decreased the exploration parameter β ($\chi^2(3) = 15.1$, $p = 0.0018$). *Post-hoc* tests revealed 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). Thus, the only parameter of the model significantly affected by dopaminergic blockade was the inverse temperature which decreased (i.e. exploration increased) as dopaminergic inhibition increased.

We next generalized this result by testing the optimized parameters of a range of different models which include a standard Q-learning model (Supplementary Fig. 6), an ε-greedy model of action selection (Supplementary Fig. 7), a model of directed exploration (Supplementary Fig. 8) taken from the literature[26] which includes an uncertainty bonus to bias decision-making towards options associated with a large value uncertainty and an original meta-learning model which sets the value of the inverse temperature based on accumulated reward prediction errors (Supplementary Fig. 9). In all of these cases, the only parameter that significantly varied with dose condition was the one responsible for controlling random exploration which flupenthixol invariably increased.

Finally, because of the strong interaction between learning rate and inverse temperature[40], we verified that our methodology was able to distinguish variations in inverse temperature β from variations in the learning rate α, by applying full parameter optimization (leaving all three parameters free) to an artificial data set generated with either α or β varying between doses while the other two parameters remain constant. The values of α and β which were used to generate the simulations had themselves been optimized on the experimental data so as to capture the maximum variability allowed by the data for this specific parameter. These optimized parameter values did indeed present significant variations between doses ($p < 0.0001$ in both cases, data not shown here), a finding which suggests that significant variations of one parameter could be at least partially replaced by variations of the other. Nevertheless, when only the learning rate α varied (Fig. 6a), the full optimization on the simulated data set was capable of identifying this effect on α ($\chi^2(3) = 16.1$, $p = 0.0011$), without confusion with the two other parameters ($\chi^2(3) \leq 1.54$, $p \geq 0.67$). Conversely, on an artificial data set where only the inverse temperature β varied (Fig. 6b), the subsequent optimization correctly identified β as the only varying parameter ($p = 0.0007$ but $p \geq 0.12$ for the other two parameters). These results clearly show that the computational analysis used above can disentangle the effects of dopamine manipulations on the inverse temperature from possible effects on learning rate despite their interdependence.

Discussion

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

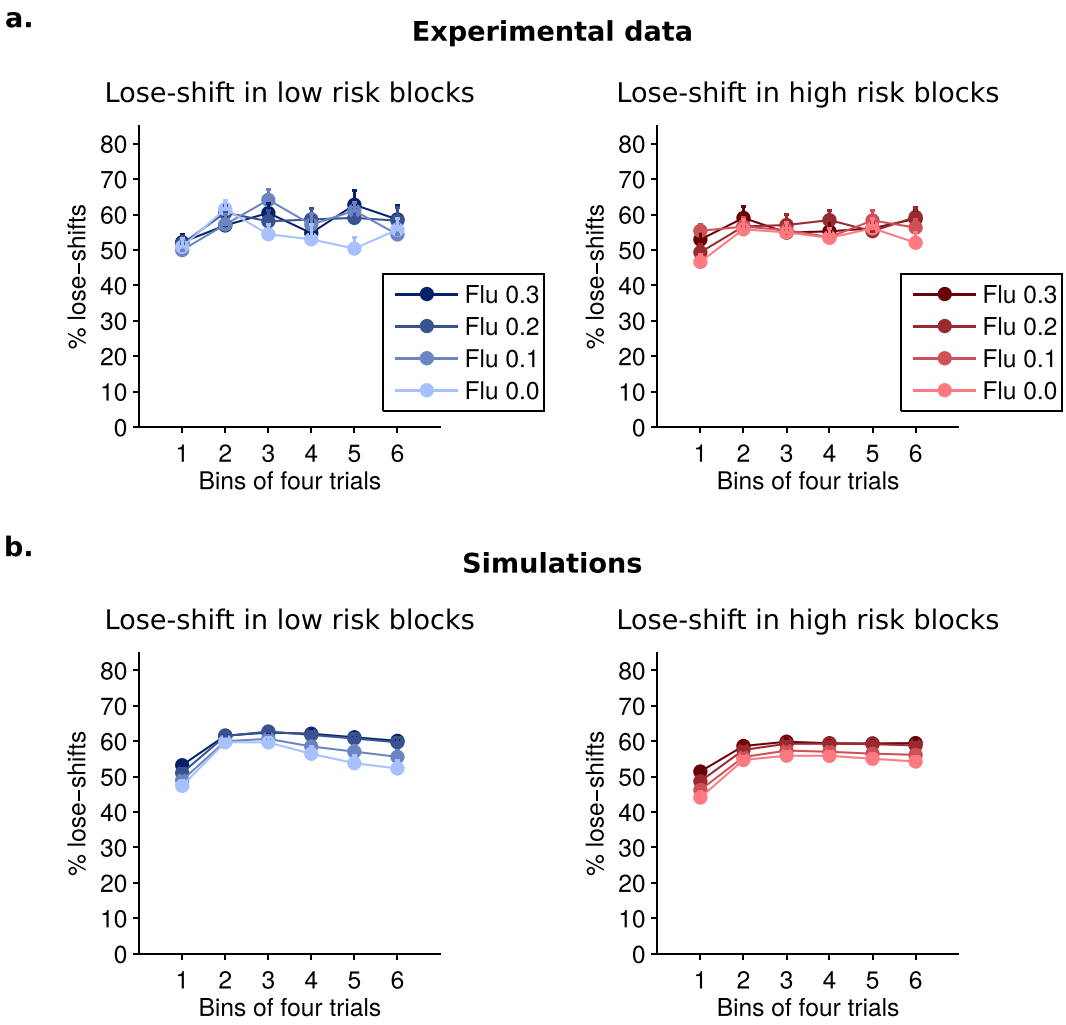

Figure 4. Lose-shift index, i.e. the proportion of trial a rat changes action after an unrewarded trial. (**a**) Average lose-shift (mean + sem) of rats across a block for different risk and dose condition. This index is not significantly affected by risk or pharmacological condition but a significant effect of trials is detected (repeated-measures ANOVA: $F(5,110) = 6.83$, $p < 0.0001$). A *post hoc* Bonferroni test indicates a significantly lower level of lose-shift in the first trial bin than for all other bins ($p < 0.0158$) except the fourth ($p = 0.17$). This effect is attributable to persistence as rats persist in selecting the previous target despite lack of reward before adjusting their behaviour. (**b**) Average lose-shift of model simulations. Simulations have a more complex lose-shift dynamic, with a significant decrease in lose-shift also occuring at the end of blocks, as well as significant main effects of dose ($p = 0.0007$) and risk ($p < 0.0001$). Similarly to win-shift (see Supp Fig. 5), simulations of the standard Q-learning (not shown here) systematically overestimated average lose-shift compared to the experimental data.

Previous experimental and computational work by Costa *et al.*[17] has demonstrated through blockade of dopamine transporter that increasing dopamine signalling increases novelty seeking which is a form of directed exploration. However, contrary to our results, they found no drug-related differences on the inverse temperature which controls undirected or random exploration. Our own task is not specially well-designed to tackle the question of novelty-driven exploration, as the only occurring change is the reward rates of the three levers, but we nonetheless optimized a model of directed-exploration[26] and still found an effect on random instead of directed exploration. Future work should focus on clearly disentangling these two processes of exploration and the role of dopamine in each of them[42]. In the present study, the effects of various doses of dopamine antagonist in the same individual indicate that normal dopamine levels limit undirected exploration in this task and therefore bias performance toward exploitation without affecting the learning rate, a result which is consistent with that of Lee *et al.*[35] using a different animal model, the macaque, and a different decision-making task. Undirected exploration may reflect several factors, and flupenthixol could for instance have reduced motivation[43,44], response vigor[45] or attention. However, the observation that learning rates were unchanged argues in favor of a selective effect on random exploratory choices.

In the task, performance improves within a block as one of the levers is gradually identified as target. Concurrently, win-shift decreases as learning progresses. At the beginning of each block, performance drops to chance levels or below, while win-shift increases. These high levels of win-shift in the absence of drug correspond

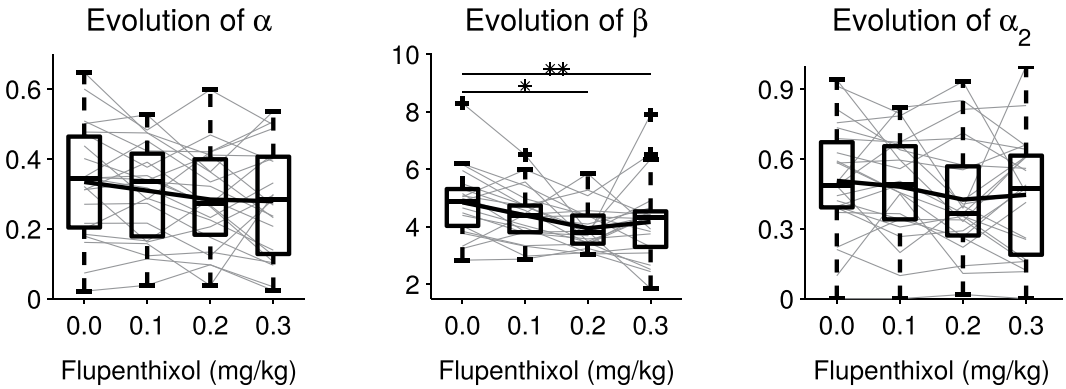

Figure 5. Distributions of the different model parameters for different flupenthixol doses. Gray lines connect the parameter values of the same individual rat and the bold black line plot the average. Box plots represent the median, interquartile and furthest observation not considered as outliers. Between dose comparisons were carried out using Friedman ANOVAs (*$p < 0.05$, **$p < 0.01$).

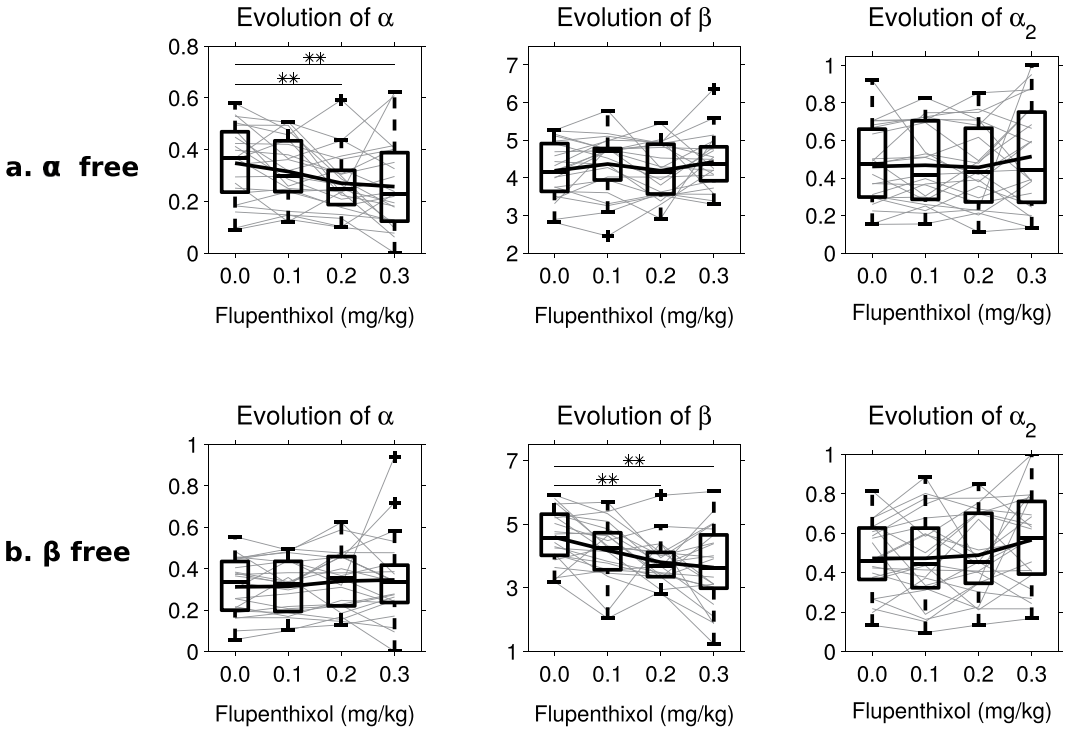

Figure 6. Validation of the optimization. A first version of the extended Q-learning model was optimized using the parameters from the 0 mg/kg condition with only α allowed to change between doses and a second version with only β allowed to vary. Once optimized, these two versions of the model were simulated to generate artificial datasets on which to optimize a model with all three parameters free to ensure that the only parameter that significantly varies between doses is indeed the one which was allowed to vary. (**a**) Distributions of the different model parameters optimized on simulations of the first version where α was free. Gray lines connect the parameter values of the same individual rat and the bold black line plot the average. Box plots represent the median, interquartile and furthest observation not considered as outliers. An individual with an extremely high β was excluded from the graph for readibility reasons but kept in the statistical tests. Between dose comparisons were carried out using Friedman ANOVAs (*$p < 0.05$, **$p < 0.01$) and only α presents a significant dose effect as originally designed. (**b**) Distributions of the different model parameters optimized on simulations of the second version where β was free. Gray lines connect the parameter values of the same individual rat and the bold black line plot the average. Box plots represent the median, interquartile and furthest observation not considered as outliers. An individual with an extremely high β also had to be excluded from this graph but was kept in the statistical tests. Between dose comparisons were carried out using Friedman ANOVAs (*$p < 0.05$, **$p < 0.01$) and only β presents a significant dose effect as originally designed.

to moments when the values of the various options have not been well identified and the rat's uncertainty is high. Indeed, in the high risk condition, identifying the correct lever appears more difficult and is associated with both lower performance and higher win-shift levels. Under dopamine blockade, the dose-dependent increase in win-shift appears independent of uncertainty, since it does not interact with either trial rank within a block or risk level. This is consistent with the notion that dopamine unconditionally scales action values and controls noise in the last stage of decision-making, where action values are converted to actual choices[30]. As to the effect of flupenthixol on lose-shift, the simulations (based on 100 runs) do show a slight increase corresponding to increased random exploration. This effect does not reach significance in the actual behavioral data, where it is probably more difficult to detect close to the ceiling value (the probability of lose-shift when choices are random is roughly 0.67).

Several past studies[27,34,46], have focused on the effects on choice of inter-individual differences in dopaminergic function. Our data thus stand in contrast with those of Beeler *et al.*[34,46], who observed that hyperdopaminergic mice allocate more time and energy to expensive options. On the basis of computational modeling, the authors interpreted their data as an increase in undirected exploration. However, data from primates and human subjects instead indicate increases in exploration with reduced dopaminergic activity, in agreement with our results. In particular, in an investigation of the behavioral disparities between human subjects due to genes controlling prefrontal and striatal dopamine function, Frank *et al.*[27] concluded that COMT alleles associated with lower dopamine levels increased the exploration component of their reaction time model. Eisenegger *et al.*[36] found in a between-subject design that a strong dose of sulpiride, a dopamine D2 receptor antagonist, in healthy human subjects favored exploration without affecting learning in a probabilistic task. These inter-individual studies supposed a fixed exploration level, and were inadequate to describe changes in exploration in the same subject. However, a similar result was achieved within subjects in an oculomotor decision task by Lee *et al.*[35] who demonstrated that injections of dopamine type 2 receptor antagonist in the dorsal striatum of two macaques deteriorated the animals' performance in a manner best explained by an increase of noise in the decision-making process rather than an effect on learning.

These experimental results, together with theoretical approaches[30], suggest that the behavioral effects of dopamine antagonists result from their action on the striatum, and Lee *et al.*'s results[35] point to a role of D2 receptors. In addition, Dickinson *et al.*[43] reported that a high dose of pimozide, a D2 antagonist appeared to have a direct effect on instrumental performance, while a weaker dose appeared to reduce the impact of the reinforcer on learning. However, Humphries *et al.*'s model[30] predicts that changes in D1 receptor activation are the main determinants of the exploration-exploitation trade-off. Our use of flupenthixol, which blocks both D1 and D2 receptors, as well as the systemic mode of administration do not allow us to clarify this issue. We could only speculate that dopamine receptor blockade by flupenthixol might exert its effect on random exploration through direct effects on postsynaptic activity, mimicking a reduction in dopaminergic activity, rather than through compensatory mechanisms involving D2 autoreceptors. By contrast, exploration directed toward novel objects[17] might be facilitated by increased dopamine levels if they enhance a dopamine-mediated novelty bonus[47].

In the three-armed bandit task used in the present study, we modeled learning of the correct action using an extended Q-learning model with forgetting, and we modeled choice behavior using a softmax mechanism. Our model was sufficient to account for behavioral performances as shown by (i) the high similarity of the simulated (unconstrained) and experimental behavioral data (Figs 1–4) and (ii) the high correlation between the modeled value (constrained by the rats' choices) of the different levers and actual choice probability, even during periods of low value when the target lever was not yet identified (Supplementary Fig. 3). We show that a forgetting mechanism is required to adequately account for the rats' behavior as a simple Q-learning mechanism appears unable to cope with the multiple target shifts (reversals) involved throughout the task (Supplementary Fig. 4). In our model, forgetting is important to reduce the value of competing actions even when these actions are not chosen any more, unlike simple Q-learning which only adjusts the value of actions actually performed. The observation that the forgetting rate is generally larger than the learning rate (Fig. 5) implies that the rats tend to persist on a choice even in the absence of reward[48]. This process stands in contrast with some theories of directed exploration[5] which predict that unchosen options become attractive as uncertainty about their outcome increases.

Our model does not include any mechanism to track uncertainty about action values, unlike several models of choice behavior in humans[4,5,27,29,49,50]. Our simulations furthermore show that the gradual reduction in win-shift within a block does not reflect a dynamic adaptation of the model parameters since it is reproduced in the simulations where these parameters are kept constant. Instead, this decrease is a consequence of the interaction between value learning and the softmax mechanism. Choice is more variable when actions values are relatively similar, and becomes less variable as the values of the various actions are better differentiated. We did observe a significant effect of risk on performance and win-shift in the behavioral data, but because the same effect was present in the simulations, it is attributable to a slower acquisition of value in the high risk situation due to increased stochasticity.

As expected from the formal analysis, dopamine was found to specifically control the exploration parameter β (inverse temperature) which represents undirected exploration or random noise in the choice process converting values into actions[28,32,33], rather than directed exploration driven by uncertainty[5,26,27,29]. Furthermore, this result is still valid with other models such as the standard Q-learning, the ε-greedy version of Q-learning and even a directed exploration model[26]. Our data agree with the theoretical proposal by Humphries *et al.*[30] that tonic dopamine in the basal ganglia could modulate the exploration-exploitation trade-off during decision-making. On the basis of a prior, biologically inspired model of the basal ganglia[51], they showed that changing simulated tonic dopamine levels had similar effects as changes in the β parameter. The mechanism they proposed rested on a different mechanism though, namely the direct modulation of the excitability of striatal neurons instead of a change in the learning mechanism as we propose here. This simply highlights the fact that our study cannot disentangle these two possibilities: either dopamine inhibition diminishes neuron excitability which in turn increases noisy exploration, or dopamine inhibition causes a down-revision of the reward value which has the same effect

downstream on decision-making. Computationally at least, these two possibilities are indistinguishable as proved by our formal analysis.

While Humphries *et al*.'s model focuses on tonic dopamine activity[30], our interpretation in terms of RPE depends on phasic dopamine signals. These two types of signals may be subject to independent control[52]. However, our formal demonstration requires that tonic dopamine levels do not directly affect learning. Instead, we assume that RPEs used in learning only consist of phasic changes with respect to current tonic activity. In other words, we assume that the threshold for RPEs tracks tonic activity, even under flupenthixol, which is expected to reduce the impact of both tonic and phasic dopaminergic activity. There is some evidence that the threshold for RPEs is adaptive[53,54]. It has been proposed that tonic dopamine changes with the overall rate of rewards in a task[25,55], and Tobler *et al*.[56] reported that phasic dopamine signals adapt to the range of expected rewards and that they could even reverse sign for an identical amount of reward. It is not clear whether the same adaptation occurs under dopaminergic blockade[57]. Nevertheless, failing to adapt the threshold would bias all RPEs towards negative values and mimic low levels of tonic dopamine activity. This would lead to more decay of learned values within and between trials, in contrast to what we observed under flupenthixol. Indeed, a model taking into account this possible effect of tonic dopamine changes on RPEs did not produce satisfactory results (negligible values of the parameter meant to simulate the effect of dopamine and worse likelihood scores than the standard model of this paper).

On the other hand, a hypothesis that our study can successfully rule out is that dopamine controls the learning rate. Indeed, our study highlights a common misconception that equates the well-established role of dopamine in learning[58] with an effect on learning rate. To the extent that learning is based on reward prediction errors and action selection on a softmax mechanism, as is typically assumed in model-free reinforcement learning, our formal analysis indicates that the inverse temperature, the parameter controlling random exploration, is the only parameter affected by simple manipulations of the reward prediction error signal. Notably, in our task, fitted learning rate was unaffected by flupenthixol, and behavioral performance was largely preserved, while the win-shift index markedly increased. In a probabilistic learning task, Pessiglione *et al*.[59] showed that administration of L-DOPA, a chemical precursor of dopamine known for enhancing dopaminergic functions, improved performance in accumulating gains compared to subjects under a dopamine antagonist. This improvement was attributed to an increase in learning from positive reward errors, but increased dopamine could also have reduced exploration. Similarly, Krugel *et al*.[36] reported that COMT alleles increasing dopamine levels were associated with better performance in a reward-based learning task with reversals, and they explained their results by a modulation of learning rate. In contrast, there are reports in humans and monkeys[60] that probabilistic learning is not impaired by dopamine antagonists[61,62]. Our results call for careful modeling of the impact of dopaminergic manipulations in behavioral tasks as changes in random exploration rates could easily be mistaken for changes in learning rate.

As fluctuations in tonic dopamine levels appear to track the average reward rate[25], it seems natural to use such a signal to regulate the exploration-exploitation trade-off[33]: high reward rates suggest that the current policy is appropriate and the subject could crystallize its behavior by exploiting more. Conversely, sudden drops in reward rate leading to tonic dopamine decreases may lead to increased exploration of the environment in search for better options. Dopamine levels could thus contribute to dynamically regulate exploratory choices in volatile environments where option values change with time. Here, we show that dopamine blockade affects undirected exploration independently from the changes in uncertainty levels within blocks. Dopaminergic regulation of exploration appears to occur at a longer time scale than that of a few trials, which would constitute a form of meta-learning[3,33], adapting behavior to the general characteristics of the task rather than to immediate events.

Methods

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

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

During the behavioral experiments, rats were maintained at 90% of their original weight by restricting their food intake to ~15 g/day. For pre-training, all rats were trained for 3 days to collect rewards during 30 min magazine training sessions. Rewards were delivered in the magazine on a random time 60 sec schedule. The conditioning cage was lit for the duration of each session. The rats then received training for 3 days under a continuous reinforcement, fixed ratio schedule FR1 (i.e. each lever press was rewarded with one pellet) until they had earned

30 pellets or 30 min had elapsed. At this stage, each lever was presented continuously for one session and the magazine was placed adjacent to the lever (side counterbalanced across rats). Thereafter, all three levers were on the left wall and the magazine on the right wall. The levers were kept retracted throughout the session except during the choice phases. On the next two sessions, levers were successively presented 30 times in a pseudo-random order (FR1-trials). One press on the presented lever produced a reward and retraction of the lever. On the next eight sessions, levers were presented 30 times but each time five presses were required to obtain the reward (FR5-trials). As a result, all rats readily pressed the levers as soon as they were presented. The rats then underwent 24 training sessions of training in the probabilistic choice task, 20 sessions of six trial blocks each and four double sessions of 12 blocks each.

Following 24 sessions of training in the task, rats received i.p. injections (1 ml/kg) of D1-D2 receptor antagonist Flupenthixol (FLU) or saline, 20 min prior to each double session of test, for a total of 16 injections. Four doses of Flupenthixol (Cis-(Z)-Flupenthixol dihydrochloride, Sigma, dissolved in saline at 0, 0.1, 0.2 or 0.3 mg/ml) were selected according to a pilot experiment. All rats received each dose (0, 0.1, 0.2 or 0.3 mg/kg) in separate sessions according to a latin square design with at least two days of recovery between injections. After two days of recovery, a second series of injections was performed under similar conditions. In addition, on the day preceding each of the eight tests, a retraining session under saline was performed.

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

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

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

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

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

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

In order to improve the intra-block dynamics of the model, a forgetting mechanism[63] was added:

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

Thus, the Q-values of non-selected actions gradually regress to 0, at which they are initialized, at a rate fixed by constant α_2. These values would otherwise only be updated when the corresponding action is selected. This mechanism was found to be necessary to achieve a good fit of the dynamics of win-shift (see Supplementary Figs 4 and 5 comparing the win-shift curves of forgetting and non-forgetting models to the experimental data) and corresponds to a perseverance mechanism independent of reward history[48]. Additionally, once combined with the action selection model described further down, we compared the log-likelihoods of the two optimized models adjusted for the number of extra parameters using either the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) calculated for the whole population (n = 23) and the four conditions (the better model being the one with lower scores). The Q-learning model with forgetting parameters had lower scores than the model without forgetting (AIC: 74534 *vs*. 82772; BIC: 76149 *vs*. 84386). We also compared individual log-likelihood scores for each rat and dose using the likelihood ratio test for nested models[40]. Given a model M1 nested into a more complex model M2 and their log-likelihood after optimization ll1 and ll2, d = 2*(ll2 - ll1) follows a chi-squared distribution with 1 degree of freedom (1 added parameter in M2) under the null hypothesis that the log-likelihood of M2 is not better than M1. In all rats except one, the forgetting model brought a highly significant improvement in likelihood when compared to the simpler model.

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

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

The key parameter of this function, the inverse temperature β, determines the relationship between action values and action selection, or in other terms between learning and performance. Low values of β result in almost equiprobable action selection (hence exploration), independently of the learned Q-values, while high values of β greatly favour the best action over all the others (i.e. exploitation). This Eq. (5) is especially crucial for optimization because it defines the probability of a rat's action at each trial given its parameter-set Θ (including learned values), which is used to calculate the likelihood of each rat's entire history of choices, H, under the parameters of each model:

$$P\langle H|\theta\rangle = \prod_{t=1}^{n_{trials}} P\langle a_t|\theta\rangle = \prod_{t=1}^{n_{trials}} \frac{e^{\beta Q_t(a_t)}}{\sum_j e^{\beta Q_t(a_j)}} \quad (6)$$

Parameter optimization was then carried out separately for each one of the four pharmacological conditions (0, 0.1, 0.2 and 0.3 mg/kg administrations of flupenthixol) by maximizing the log-likelihood of the model for each rat so as to get individual parameter-sets for each subject. Each parameter was initialized in three different points of its parameter space to avoid being trapped in local maxima, for a total of $3^3 = 27$ initializations for the different combinations of parameter initializations. A first validation of this optimization is to check the adequation between observed action probabilities and those predicted by the model when constrained to the original experimental data, meaning that the model makes the same choices as the animal and we extract from it the softmax probabilities for comparison. We then fit a simple linear model without an intercept (observed = b_1 *simulation) in the hope that b_1 would be close to 1 and the proportion of explained variance satisfactory. The results of this comparison were successful and are reported in Supplementary Table 1 and Fig. 3. In addition, because it is possible that rats quickly adapt their learning rate when reward contingencies abruptly change, we carried out this same comparison on trials 2–6 of each block, immediately after such a switch in contingencies, and found that there was still a very strong linear relationship between observed and predicted actions (b1 = 0.99, 5% confidence interval = [0.97 1.01] and adjusted $R^2 = 0.77$).

To verify the conclusions drawn from the characteristics of the optimized model, we also had to check that the model was able to properly reproduce the experimental data by running unconstrained simulations using the individual parameters fitted to each rat 100 times on the full experiment. These simulations were then averaged and compared to the original data as reported in the main text. This verification procedure has recently been advocated[64] as a standard and crucial requirement when modeling experimental data.

We used the same optimization method on four additional models. The first of these additional models is the standard Q-learning model which is identical to the model just presented except it has no forgetting mechanism. The second additional model is the forgetting Q-learning model presented earlier with an ε-greedy action selection mechanism instead of the softmax rule. This mechanism simply selects the action with highest Q-value with probability 1-ε and the remaining two actions with probability ε/2. Thus the larger the ε parameter, the more exploratory the behavior of the animal. The third model we tested is borrowed from the literature[26] and consists in combining Q-values, which estimate the expected payoff of a given action, with an uncertainty bonus υ, which estimates the variance of these payoffs. Q-values are updated in the same way as the extended Q-learning model (Eqs (2) and (3)). Each time an action a is selected and rewarded, we calculate the uncertainty prediction error $\xi_t(a)$ based on squared reward prediction errors (taken from Eq. 1) and the previous estimate of the variance:

$$\xi_t(a) = \delta^2{}_t - \upsilon_{t-1}(a) \quad (7)$$

Using squared prediction errors means that it is the magnitude of reward prediction errors which matters and provides us with an estimate of variance. The expected uncertainty of the same action can then be updated using its own learning rate parameter α^{φ}:

$$\upsilon_t(a) = \upsilon_{t-1}(a) + \alpha^{\phi}\xi_t(a) \quad (8)$$

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

$$P(a_{t+1} = a_i) = \frac{e^{\beta(Q_t(a_i)+\phi\upsilon_t(a_i))}}{\sum_j e^{\beta(Q_t(a_j)+\phi\upsilon_t(a_j))}} \quad (9)$$

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

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

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

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

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

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

Data analysis. Trials from each 24-trial block were grouped into six bins of four trials and averaged for each of the four flupenthixol doses and each of the two risk levels in each rat. Asymptotic performance and win-shift levels were estimated on the last six trials of blocks, respectively. Performance was represented by the proportion of trials where the target lever was selected. Exploration was indexed by win-shift, the proportion of trials where rats changed lever choice after being rewarded on the target lever. One rat did not complete the task under 0.3 mg/kg of flupenthixol and was therefore removed from analysis. Five other rats missed occasional parts of the blocks (no win), so statistical analysis of win-shift concerned 18 rats. Experimental or simulated data were submitted to repeated-measures ANOVAs with three factors (flupenthixol dose, risk and trial) with an additional factor (experiment/model) when appropriate. *Post-hoc* comparisons were performed using simple t-tests with Bonferroni correction. Model parameters were analyzed as a function of dose using non-parametric Friedman's ANOVA as the distribution of β values violated the assumptions of normality (Shapiro-Wilk test of normality $p < 0.0001$ for 0.2 mg/kg of flupenthixol), and sphericity (Mauchly test: $\chi^2(5) = 24.68$, $p = 0.0002$) required for a repeated-measures ANOVA. First-order error risk was set at 0.05 (two-sided tests).

Data Availability

The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

References

1. Rushworth, M. F. S. & Behrens, T. E. J. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience* **11**(4), 389–397 (2008).
2. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**(7095), 876–879 (2006).
3. Schweighofer, N. & Doya, K. Meta-learning in reinforcement learning. *Neural Networks* **16**(1), 5–9 (2003).
4. Nassar, M. R., Wilson, R. C., Heasly, B. & Gold, J. I. An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. *J. Neurosci* **30**(37), 12366–12378 (2010).
5. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore-exploit dilemma. *J exp Psychol Gen* **143**(6), 2074–2081 (2014).
6. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science (80-.).* **275**, 1593–1599 (1997).
7. Schultz, W. Updating dopamine reward signals. *Current Opinion in Neurobiology* **23**(2), 229–238 (2013).
8. Watabe-Uchida, M., Eshel, N. & Uchida, N. Neural Circuitry of Reward Prediction Error. *Annu. Rev. Neurosci.* **40**(1), 373–394 (2017).
9. Coddington, L. & Dudman, J. T. The timing of action determines reward prediction signals in identified midbrain dopamine neurons. *Nature Neuroscience* **21**(11), 1563–1573 (2018).
10. Bayer, H. M. & Glimcher, P. W. Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron* **47**(1), 129–141 (2005).
11. Morris, G., Nevet, A., Arkadir, D., Vaadia, E. & Bergman, H. Midbrain dopamine neurons encode decisions for future action. *Nat. Neurosci.* **9**(8), 1057–1063 (2006).
12. Roesch, M. R., Calu, D. J. & Schoenbaum, G. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat. Neurosci.* **10**(12), 1615–24 (2007).
13. Matsumoto, M. & Hikosaka, O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* **459**(7248), 837–841 (2009).
14. Centonze, D., Picconi, B., Gubellini, P., Bernardi, G. & Calabresi, P. Dopaminergic control of synaptic plasticity in the dorsal striatum. *Eur. J. Neurosci.* **13**(6), 1071–1077 (2001).
15. Reynolds, J. N., Hyland, B. I. & Wickens, J. R. A cellular mechanism of reward-related learning. *Nature* **413**(6851), 67–70 (2001).
16. Izhikevich, E. M. Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cereb. Cortex* **17**(10), 2443–2452 (2007).
17. Costa, V. D., Tran, V. L., Turchi, J. & Averbeck, B. B. Dopamine modulates novelty seeking behavior during decision making. *Behav. Neurosci.* **128**(5), 556–566 (2014).
18. Haluk, D. M. & Floresco, S. B. Ventral striatal dopamine modulation of different forms of behavioral flexibility. *Neuropsychopharmacology* **34**(8), 2041–52 (2009).
19. Flagel, S. B. *et al*. A selective role for dopamine in stimulus–reward learning. *Nature* **469**(7328), 53–57 (2011).
20. Papageorgiou, G. K., Baudonnat, M., Cucca, F. & Walton, M. E. Mesolimbic Dopamine Encodes Prediction Errors in a State-Dependent Manner. *Cell Rep.* **15**(2), 221–8 (2016).
21. Jenni, N. L., Larkin, J. D. & Floresco, S. B. Prefrontal Dopamine D1 and D2 Receptors Regulate Dissociable Aspects of Decision Making via Distinct Ventral Striatal and Amygdalar Circuits. *J. Neurosci.* **37**(26), 6200–6213 (2017).
22. Salamone, J., Correa, M., Mingote, S. & Weber, S. Beyond the reward hypothesis: alternative functions of nucleus accumbens dopamine. *Curr. Opin. Pharmacol.* **5**(1), 34–41 (2005).
23. Berridge, C. W. & Arnsten, A. F. T. Psychostimulants and motivated behavior: Arousal and cognition. *Neurosci. Biobehav. Rev.* **37**(9), 1976–1984 (2013).
24. Stopper, C. M., Tse, M. T. L., Montes, D. R., Wiedman, C. R. & Floresco, S. B. Overriding Phasic Dopamine Signals Redirects Action Selection during Risk/Reward Decision Making. *Neuron* **84**(1), 177–189 (2014).
25. Niv, Y., Daw, N. D., Joel, D. & Dayan, P. Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology (Berl).* **191**(3), 507–520 (2007).
26. Naudé, J. *et al*. Nicotinic receptors in the ventral tegmental area promote uncertainty-seeking, *Nat. Neurosci*. October 2015 (2016).

27. Frank, M. J., Doll, B. B., Oas-Terpstra, J. & Moreno, F. The neurogenetics of exploration and exploitation: Prefrontal and striatal dopaminergic components, In. *Nature Neuroscience* **12**(8), 1062–1068 (2009).
28. Zajkowski, W. K., Kossut, M. & Wilson, R. C. A causal role for right frontopolar cortex in directed, but not random, exploration. *Elife* **6**, 1–18 (2017).
29. Cogliati Dezza, I., Yu, A. J., Cleeremans, A. & Alexander, W. Learning the value of information and reward over time when solving exploration-exploitation problems. *Sci. Rep.* **7**(1), 16919 (2017).
30. Humphries, M. D., Khamassi, M. & Gurney, K. Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Front. Neurosci* **6**(FEB), 1–14 (2012).
31. Sutton, R. & Barto, A. Reinforcement Learning: An Introduction, *MIT Press, Cambridge, Massachusetts* (1998).
32. Doya, K. Modulators of decision making. *Nat. Neurosci.* **11**(4), 410–416 (2008).
33. Khamassi, M., Enel, P., Dominey, P. F. & Procyk, E. Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. *Prog Brain Res* **202**, 441–464 (2013).
34. Beeler, J. A., Daw, N., Frazier, C. R. M. & Zhuang, X. Tonic dopamine modulates exploitation of reward learning. *Front. Behav. Neurosci.* **4**(November), 170 (2010).
35. Lee, E., Seo, M., Dal Monte, O. & Averbeck, B. B. Injection of a Dopamine Type 2 Receptor Antagonist into the Dorsal Striatum Disrupts Choices Driven by Previous Outcomes, But Not Perceptual Inference. *J. Neurosci.* **35**(16), 6298–6306 (2015).
36. Eisenegger, C. *et al*. Role of dopamine D2 receptors in human reinforcement learning. *Neuropsychopharmacology* **39**(10), 2366–75 (2014).
37. Krugel, L. K., Biele, G., Mohr, P. N. C., Li, S.-C. & Heekeren, H. R. Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proc. Natl. Acad. Sci. USA* **106**(42), 17951–6 (2009).
38. Averbeck, B. B. Theory of Choice in Bandit, Information Sampling and Foraging Tasks. *PLoS Comput. Biol.* **11**(3), 1–28 (2015).
39. Lesaint, F., Sigaud, O., Flagel, S. B., Robinson, T. E. & Khamassi, M. Modelling Individual Differences in the Form of Pavlovian Conditioned Approach Responses: A Dual Learning Systems Approach with Factored Representations. *PLoS Comput. Biol.*, **10**, 2 (2014).
40. Daw, N. D. Trial-by-trial data analysis using computational models. *Decis. Making, Affect. Learn. Atten. Perform.* **XXIII**, 1–26 (2011).
41. Averbeck, B. B. & Costa, V. D. Motivational neural circuits underlying reinforcement learning. *Nat. Neurosci.* **20**(4), 505–512 (2017).
42. Gershman, S. J. & Tzovaras, B. G. Dopaminergic genes are associated with both directed and random exploration. *Neuropsychologia*, **120**, 97–104 (2018).
43. Dickinson, A., Smith, J. & Mirenowicz, J. Dissociation of Pavlovian and instrumental incentive learning under dopamine antagonists. *Behav. Neurosci* **114**(3), 468–83 (2000).
44. Barbano, M. F., Le Saux, M. & Cador, M. Involvement of dopamine and opioids in the motivation to eat: influence of palatability, homeostatic state, and behavioral paradigms. *Psychopharmacology (Berl).* **203**(3), 475–487 (2009).
45. Niv, Y. Cost, benefit, tonic, phasic: What do response rates tell us about dopamine and motivation? *Ann. N. Y. Acad. Sci.* **1104**, 357–376 (2007).
46. Beeler, J. A., Frazier, C. R. M. & Zhuang, X. Putting desire on a budget: dopamine and energy expenditure, reconciling reward and resources. *Front. Integr. Neuroscir.* **6**(July), 49 (2012).
47. Kakade, S. & Dayan, P. Dopamine: generalization and bonuses. *Neural Netw., Jun-Jul* **15**(4–6), 549–559 (2002).
48. Katahira, K. The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior. *J. Math. Psychol.* **66**, 59–69 (2015).
49. Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**(9), 1214–21 (2007).
50. Jepma, M. *et al*. Catecholaminergic Regulation of Learning Rate in a Dynamic Environment. *PLOS Comput. Biol.* **12**(10), e1005171 (2016).
51. Gurney, K. N., Humphries, M., Wood, R., Prescott, T. J. & Redgrave, P. Testing computational hypotheses of brain systems function: a case study with the basal ganglia. *Network* **15**(4), 263–90 (2004).
52. Grace, A. A., Floresco, S. B., Goto, Y. & Lodge, D. J. Regulation of firing of dopaminergic neurons and control of goal-directed behaviors. *Trends Neurosci.* **30**(5), 220–227 (2007).
53. Park, S. Q. *et al*. Adaptive coding of reward prediction errors is gated by striatal coupling. *Proc Natl Acad SCI USA* **109**, 4285–4289 (2012).
54. Lak, A., Stauffer, W. R. & Schultz, W. Dopamine neurons learn relative chosen value from probabilistic rewards. *Elife* **5** (2016).
55. Guitart-Masip, M., Beierholm, U. R., Dolan, R., Duzel, E. & Dayan, P. Vigor in the Face of Fluctuating Rates of Reward: An Experimental Examination. *J. Cogn. Neurosci.* **23**(12), 3933–3938 (2011).
56. Tobler, P. N., Fiorillo, C. D. & Schultz, W. Adaptive coding of reward value by dopamine neurons. *Science* **307**, 1642–1645 (2005).
57. Diederen, K. M. J. *et al*. Dopamine Modulated Adaptive Prediction Error Coding in the Human Midbrain and Striatum. *J. Neurosci.* **37**(7), 1708–1720 (2017).
58. Schultz, W. Neuronal Reward and Decision Signals: From Theories to Data. *Physiol. Rev.,* **95**(3), 853–951 (2015).
59. Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J. & Frith, C. D. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* **442**(7106), 1042–1045 (2006).
60. Costa, V. D., Tran, V. L., Turchi, J. & Averbeck, B. B. Reversal learning and dopamine: a bayesian perspective. *J. Neurosci* **35**(6), 2407–2416 (2015).
61. Shiner, T. *et al*. Dopamine, salience, and response set shifting in prefrontal cortex. *Cereb. Cortex* **25**(10), 3629–3639 (2015).
62. Smittenaar, P. *et al*. Decomposing effects of dopaminergic medication in Parkinson's disease on probabilistic action selection - learning or performance? *Eur. J. Neurosci.* **35**(7), 1144–1151 (2012).
63. Ito, M. & Doya, K. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *J. Neurosci.* **29**(31), 9861–9874 (2009).
64. Palminteri, S., Wyart, V. & Koechlin, E. The Importance of Falsification in Computational Cognitive Modeling. *Trends Cogn. Sci.* **21**(6), 425–433 (2017).

Acknowledgements

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

Author Contributions

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

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-43245-z.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019