Regulation of reinforcement learning parameters captures long-term changes in rat behaviour

François Cinotti, Etienne Coutureau, Mehdi Khamassi, Alain R Marchand, Benoît Girard

Received: 30 September 2023 | Revised: 14 May 2024 | Accepted: 5 June 2024

DOI: 10.1111/ejn.16449

RESEARCH REPORT

EJN European Journal of Neuroscience FENS WILEY

Regulation of reinforcement learning parameters captures long-term changes in rat behaviour

François Cinotti[1,2] | Etienne Coutureau[3] | Mehdi Khamassi[1] | Alain R. Marchand[3] | Benoît Girard[1]

[1]Institut des Systèmes Intelligents et de Robotique, Sorbonne Université, CNRS, Paris, France

[2]University of Reading, School of Psychology and Clinical Language Sciences, Whiteknights, Reading, UK

[3]INCIA, CNRS UMR 5287, Université de Bordeaux, Bordeaux, France

Correspondence
François Cinotti, Institut des Systèmes Intelligents et de Robotique, Sorbonne Université, CNRS, Paris, France.
Email: francois.cinotti@gmail.com

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-11-BSV4-006; CRCNS 2015 Project, Grant/Award Number: ANR-15-NEUC-0001; Agriculture, Food and Health research theme

Edited by: Maxime Assous

Abstract

In uncertain environments in which resources fluctuate continuously, animals must permanently decide whether to stabilise learning and exploit what they currently believe to be their best option, or instead explore potential alternatives and learn fast from new observations. While such a trade-off has been extensively studied in pretrained animals facing non-stationary decision-making tasks, it is yet unknown how they progressively tune it while learning the task structure during pretraining. Here, we compared the ability of different computational models to account for long-term changes in the behaviour of 24 rats while they learned to choose a rewarded lever in a three-armed bandit task across 24 days of pretraining. We found that the day-by-day evolution of rat performance and win-shift tendency revealed a progressive stabilisation of the way they regulated reinforcement learning parameters. We successfully captured these behavioural adaptations using a meta-learning model in which either the learning rate or the inverse temperature was controlled by the average reward rate.

KEYWORDS
decision-making, dopamine, exploration-exploitation trade-off, meta-learning

1 | INTRODUCTION

Faced with an uncertain environment in which resources fluctuate continuously, animals must repeatedly decide whether to stabilise learning and exploit what they currently believe to be their best option, or instead explore potential alternatives and adapt fast in case better opportunities are in fact available. Such a trade-off could either reflect an equilibrium between exploration and exploitation (Cohen et al., 2007), between learning fast or slow (Behrens et al., 2007), or both, given these two processes are known to be largely interdependent (Daw, 2011). Such a trade-off could nevertheless itself be tuned to the circumstances: if the animal is currently

Abbreviations: FR, Fixed Ratio; LR, Low Risk; HR, High Risk; AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion; MSE, Mean Squared Error; QL, Q-Learning; OFC, Orbito-Frontal Cortex; RPE, Reward Prediction Error; TS, Thompson Sampling; DTS, Dynamic Thompson Sampling; SWTS, Sliding Window Thompson Sampling..

Mehdi Khamassi, Alain R. Marchand and Benoît Girard contributed equally to this work.

experiencing a high reward rate, then it would seem in its best interest to keep exploiting its current strategy and reduce the learning rate so as to be immune to noisy feedback; in contrast, if the reward rate drops, this could be a signal that it is time to start exploring new strategies and adopt a high learning rate to adapt faster to environmental changes.

While the exploration-exploitation trade-off has attracted a lot of interest in recent years, the precise mechanisms by which this trade-off is tuned to the animal's current experience are still unknown. This is partly due to its tight intertwining with learning and inference processes (Findling & Wyart, 2021), which makes it difficult to disentangle them. In humans facing stochastic decision-making tasks with non-stationary reward probabilities, choice variability has been investigated in terms of regulation of the learning rate in response to volatility (Behrens et al., 2007; Cazé & Van Der Meer, 2013). Importantly, if in a learning task, an animal's performance is seen to deteriorate, this can arguably be explained either by a decrease in its ability to learn and identify the best action, or by a reduced tendency to actually use what it has learnt to guide its action. Furthermore, sub-optimal choices, whether due to learning or decision-making defects, necessarily impact the converse process: the animal can only learn about actions that the decision-making process has sampled, and conversely, if a learning deficiency makes actions less discriminable, then even a greedy decision-making strategy will produce apparently random behaviour. Therefore, the current predominance of theories on the regulation of learning should not close the door to the potential role of the regulation of exploration as a mechanism of adaptation to the environment.

Reinforcement learning (Sutton & Barto, 1998), a class of algorithms for learning what actions to take based on discrete outcomes in the form of rewards and punishment, offers a very useful framework for tackling this question, because it explicitly separates the learning mechanism, controlled by a learning rate parameter, from the decision-making process, typically modelled as a softmax rule (Daw et al., 2006) controlled by a parameter called the inverse temperature. Although the two parameters are still correlated, so that increasing one can be partly compensated for by decreasing the other, this compensation is not a strict equivalence. Thus, it becomes possible to distinguish an effect on learning from an effect on the exploration-exploitation balance. For instance, when simulated, manipulation of the learning rate affects the slope of learning curves, while that of the inverse temperature affects the value towards which such learning curves converge.

In a previous paper (Cinotti et al., 2019), we indeed showed through careful modelling of rat behavioural adaptation to changing reward probabilities in a three-armed bandit task that pharmacological inhibition of dopamine via flupenthixol, a non-discriminative D1 and D2 receptor antagonist, caused an increase in exploration without affecting learning itself. It remained to be seen whether this relationship between dopamine and the exploration-exploitation trade-off was a functional one or merely an experimental artefact. It was at least conceivable that even the lowest levels of inhibition did not really mimic the natural fluctuations of dopamine within the brain. Dopamine plays a well-established role in its phasic form in signalling reward prediction errors, which are crucial to reinforcement learning (Hart et al., 2014; Schultz et al., 1997). In addition, it has been postulated to carry information about uncertainty (Gilbertson & Steele, 2021) or average reward rate (Niv, 2007; Niv et al., 2007) in its background or tonic activity. This led us to the hypothesis that animals might regulate the exploration-exploitation trade-off via an effect of the average reward rate on dopamine levels (Humphries et al., 2012; Khamassi et al., 2011).

In this paper, we aim to explore this hypothesis by looking at long-term changes in behaviour as rats learned to choose a rewarded lever in a three-armed bandit task across 24 days of the experiment. These days constitute the pretraining phase of the experiment presented in Cinotti et al. (2019). Here, we investigate how animals adapted their behaviour while progressively learning the task, and whether such an adaptation resulted from a form of meta-learning (Schweighofer & Doya, 2003) in which either the exploration-exploitation trade-off or the learning rate was being regulated by the current performance level.

2 | METHODS

2.1 | Experimental methods

Experimental methods are as reported in a previously published study (Cinotti et al., 2019). Male Long Evans rats (n = 24) were obtained from Janvier Labs (France) at the age of two months. 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 LEDs (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 behavioural 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 was on the right wall. The levers were kept retracted throughout the session except during the choice phases. In 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. In 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 sessions of the probabilistic choice task, which make up our experiment, 20 sessions of six trial blocks each and four double sessions of 12 blocks each.

The experimental task (Figure 1) consisted of 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 intervals varied randomly within a 4.5 to 8 second range. Trials were grouped into unsignalled blocks of fixed length (24 trials each) characterised by a constant combination of target lever and risk. The target lever always changed between blocks. Therefore, rats had to re-learn the target lever after each block change. Blocks were ordered pseudo-randomly within a session with all combinations of target and risk counterbalanced and tested once (or twice in the last four double sessions).

2.2 | Data analysis

In order to smooth the appearance of block average performance and win-shift, trials were binned into groups of

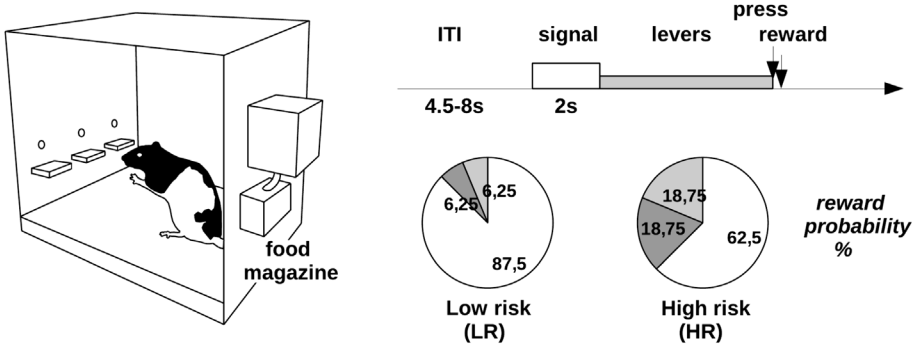

FIGURE 1 Outline of the experimental task, reproduced from Cinotti et al. (2019).

4 trials. In the case of win-shift, the average was obtained by pooling all potential win-shift events belonging to a given bin between blocks (*e.g.*, the ratio of the number of win-shifts, which occurred in the first four trials of low-risk blocks to the number of win trials in the same period). Individual performance and win-shift curves were calculated first, then the population average, so that error bars correspond to inter-individual variability.

Smoothed performance and win-shift curves were analysed using repeated-measures ANOVAs, the between factor consisting of individual subjects and the within factors being risk, session (grouped into bins of 6) and trial bins. *Post hoc* t-tests with a Bonferroni correction comparing sessions for trial x risk combinations were performed whenever the interaction between trial, risk and session was significant, the only exception being experimental win-shift (Figure 2c,d) for which, because the interaction between risk and session was still significant, we compared average win-shift over all bins instead, as shown in Figure 2c,d. These same methods were used when analysing the different model simulations.

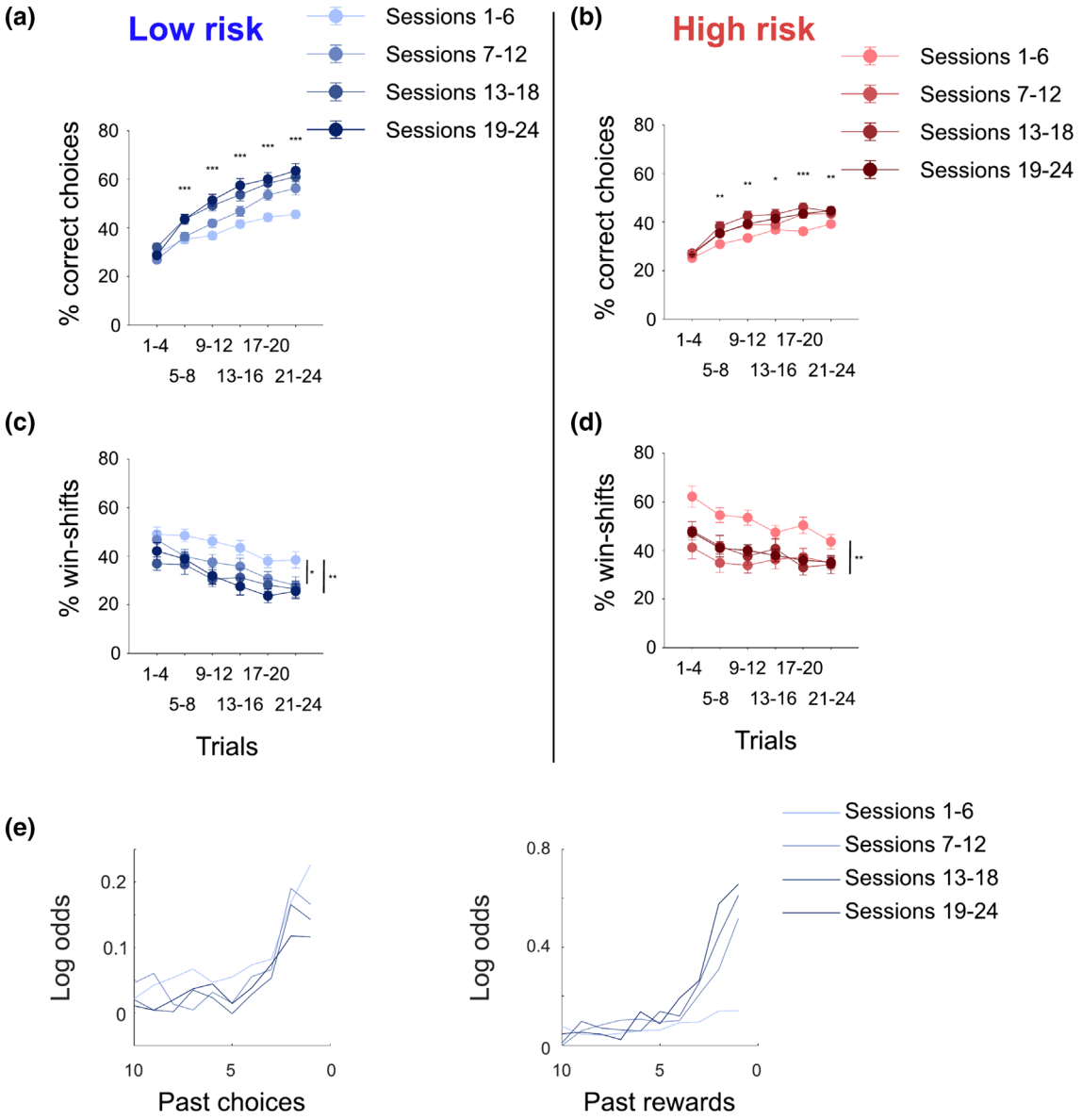

FIGURE 2 Changes in behaviour across task sessions. (a,b) Mean performance ± s.e.m. (n = 24 subjects) increases between sessions in low- and high-risk blocks, respectively. Trials are binned together into groups of four for smoothing purposes. Stars indicate that there is a significant difference between at least two groups of sessions for a given bin of four trials. (c,d) Mean win-shift ± s.e.m. decreases between sessions in low- and high-risk blocks, respectively. Because no significant trial x sessions x risk interaction was detected, win-shift is not compared trial by trial as with performance, but we instead report significant session x risk differences. (e) Average logistic regression weights for effects of past choices and rewards on current trial. Significance levels as follows: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

To separate the effects of past choices and rewards on choosing one of the three levers, we fitted individual logistic regression models in which the dependent variable was whether or not the lever in question had been chosen on a given trial, and the independent variables were whether or not that same lever had been chosen in each of the past ten trials, and whether or not it had received a reward in each of the past ten trials. We then had 3 sets (1 for each lever) of 20 regression coefficients, which we averaged between levers. We also attempted but failed to fit a multinomial regression model, which is why we adopted this approach instead. These models were fit using the *glmfit* function in MATLAB and any individual for which this function failed to converge within the default maximum number of iterations was removed from the analysis. The same method was applied to simulations.

2.3 | Model fitting

With the exception of the two Thompson sampling models, all other models relied on a softmax action selection process, which defines the trial-by-trial likelihood of the model (Daw, 2011):

$$P(a_t) = \frac{e^{\beta Q_t(a_t)}}{\sum_i e^{\beta Q_t(a_i)}}$$

For the Thompson sampling models, we computed the likelihood of selecting an action given the posterior density functions f_1, f_2 and f_3, and their corresponding cumulative probability functions F_1, F_2 and F_3, e.g. P (a_t = action 1) is the probability that the sample from f_1 is greater than the two other samples, which is the integral over all possible sample values of the product between f_1 and the two cumulative distribution functions F_2 and F_3:

$$P(a_t = a_1) = \int_0^1 f_1(x).F_2(x).F_3(x)dx$$

Likelihood over the entire 24 sessions of the experiment is then defined as the product of these trial likelihoods, and the log-likelihood as the total sum of the trial log-likelihoods. Q-learning derived models were optimised through minimisation of the negative log-likelihood using the built-in *fmincon* function in MATLAB, which implements a gradient descent method. To avoid falling into a local minimum and missing the global minimum, three different fixed initial points per parameter were combined for different initialisations of the gradient descent (*i.e.*, 27 different initialisations for the three-parameters forgetting model, 243 for the various 5-parameters models and 729 for the six-parameters sigmoid meta-learning model). The fixed initialisation points for the different parameters and their bounds are given in Table 1.

The two Thompson sampling models were not optimised in this manner. Instead, because they only have one free parameter, it was easy to explicitly compute the log-likelihood for integer values of T and C between 2 and 100 to find an optimal parameter value.

2.4 | Model comparisons

It is possible to directly compare models with the same number of parameters by looking at their log-likelihood, the better model simply being the one with the highest log-likelihood. When the number of parameters is different, it is necessary to take this into account to avoid overfitting. Two well-known criteria were used in this study: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). On an individual level, the BIC, which also depends on the number of trials, proved more conservative than the AIC, but when summed over all individuals to select the best model at the population level, both criteria were always in agreement, thus sparing us a discussion over the different merits and drawbacks of these criteria (Lebarbier & Mary-huard, 2006).

Ultimately, models were judged by their ability to produce simulations similar to the original experimental data (Humphries & Gurney, 2007; Palminteri et al., 2017). For each individual, we ran 100 independent simulations using the optimised set of parameters and the same block schedule as the individual subjects; in these simulations, the agent's choices were made stochastically based on the distribution resulting from the softmax function (or random samples taken from the posterior density functions for the two Thompson sampling methods) and rewards were randomly given according to the current block reward distribution. We then averaged block performance and win-shift of the 100 simulations to get 24 individual average simulations. These were then averaged again to produce the different simulated performance and win-shift curves shown in this study. The standard error of the mean thus corresponds to the variability between average individual simulations. Simulations were judged based on how well they reproduced the experimental data, which we quantified using mean squared errors (MSE) relative to the original average curves:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$$

TABLE 1 Initialisation points and bounds of the parameters of the different models.

	Common parameters			Meta-learning models			Time-based models		
	α	α_2	β	α_R	β_{max}	β_{min}	a	k	β_0/β_{max}
Initial values	0.1, 0.5, 0.9	0.1, 0.5, 0.9	1, 5, 20	0.1, 0.5, 0.9	1, 5, 20	1, 5, 20	0.01, 0.05. 0.1	0.1, 0.5, 2	1, 5, 20
Bounds	[0, 1]	[0, 1]	[0, +∞[	[0, 1]	[0, +∞[	[0, +∞[	[0, 1]	[0, +∞[	[0, +∞[

With n the number of data points (*i.e.*, 6 trial bins × 4 groups of sessions × 2 risk levels), Y_i the experimental values and $\widehat{Y}_i$ the simulation values.

For the staggered models, we tested the effect of sessions on α, α_2 and β with a Friedman ANOVA. The Friedman ANOVA was used because out of the four distributions of β, three were significantly different from a normal distribution according to the Shapiro–Wilk test (*swtest* MATLAB function written by Ahmed BenSaïda [2014]), and the assumption of sphericity was also violated according to a Mauchly test ($p = 1.10^{-4}$).

3 | RESULTS

3.1 | Experimental results

The rats were presented with a three-armed bandit task, which consisted of discrete trials in which they had to choose one of three levers in order to get a reward (Figure 1). Each session (a total of 24) was comprised of six blocks of 24 trials, two blocks per lever: one high-risk block in which the most rewarded lever had a probability of reward of 5/8 while the other levers were rewarded 3/16th of the time; and one low-risk block in which the best lever was rewarded 7 times out of 8 versus 1 out of 16 for the two other levers. Therefore, discrimination of the correct lever was much easier in the low-risk than in the high-risk condition. Blocks were ordered pseudo-randomly within each session so that the same lever was never the best twice in a row. The last four sessions contained 12 blocks instead of 6, so that each lever x risk combination was tested twice rather than once.

In Figure 2a,b, we tracked the rats' average performance within blocks, which is defined as the number of times they selected the lever with the highest reward probability in the current block. Average performance at the beginning of blocks started at around 26% and 29% in high-risk and low-risk blocks, respectively, which is significantly below chance levels of 33% (t-test that the average performance in either low- or high-risk blocks equals 1/3: $p < 10^{-6}$) indicating that rats were unaware that a block change had occurred and were persisting with the previously best-rewarded option. As rats learned to find the best lever, performance then increased more or less rapidly depending on risk condition, as expected, but also depending on the stage within the experiment. In the first six sessions, performance levels reached 45% and 40% at the end of low- and high-risk blocks, respectively, compared to 63% and 48% in the last six sessions. These observations were supported by repeated-measures ANOVA with significant trial ($F[5115] = 112.1$, $p < 10^{-4}$), session ($F[3,69] = 29.4$, $p < 10^{-4}$) and risk

($F[1,23] = 98.4$, $p < 10^{-4}$) effects. All possible combinations of these three main factors were also significant ($p < 0.0463$). *Post hoc* Bonferroni t-tests on low-risk blocks showed that, with the exception of the first four trials, performance in the first six sessions was always significantly worse than at least one other group of six sessions. Similarly, performance in high-risk blocks did not differ between any sessions in the first four trials but was significantly worse for all subsequent trials in sessions 1–6 compared to at least one other group of sessions. To summarise, in addition to the better performance in low-risk blocks compared to the high-risk blocks as expected, the analysis of performance reveals a long-term improvement in performance. In contrast to the post-training results in Cinotti et al. (2019), because here the curves do not have time to visibly converge within the blocks of 24 trials, we cannot say at this point whether this is attributable to an increased learning rate – which tends to steepen the slopes of learning curves – or to an increase in the inverse temperature – which causes higher final performance levels.

Win-shift, an exploratory strategy, also changed significantly throughout the experiment. It consists of the probability of changing the lever, after being rewarded for a correct choice of the current best lever. As depicted in Figure 2c,d, win-shift decreased within blocks as uncertainty surrounding the identity of the correct lever also decreased (significant trial effect found with a repeated-measures ANOVA: $p < 10^{-4}$). Win-shift in high-risk blocks was greater than in low-risk blocks (significant effect of risk: $F[1,23] = 63.9$, $p < 10^{-4}$). Contrary to performance, the interaction between sessions and risk was the only significant interaction ($F[3,69] = 6.1$, $p = 0.0021$) involving sessions. Win-shift in the first six sessions was significantly higher than for all subsequent sessions in both low- (*post hoc* Bonferroni test, $p < 0.025$) and high- ($p < 0.006$) risk blocks (Figure 2c,d). We also analysed lose-shift behaviour, a corrective strategy consisting of switching choices immediately after failing to receive a reward, but found that this indicator was affected neither by session nor by risk level, with an average value of 54% (not shown).

To complete our analysis of experimental behaviour, we performed logistic regressions on the choices made on each trial using both past choices and past rewards in the last ten trials as regressors (Hattori et al., 2023; Lau & Glimcher, 2005). The average coefficients (excluding two outlier subjects for which the fitting algorithm failed to converge) are plotted in Figure 2e. In the first six sessions, there was a strong tendency to repeat the previous choice with log odds above 0.2 to repeat the exact same choice as the previous trial. This tendency dropped as the log odds fell to about 0.1 in the last six sessions. Even more striking is the relationship between past rewards and choices. In the first six sessions, the coefficients for past rewards are small and their variation is quite flat within the 10-trial window so that a reward at t-1 has barely more impact than at t-10. In later sessions, however, there is a very strong tendency to repeat a choice which has been rewarded, and this tendency quickly drops the further back in time the reward occurred. Therefore, behaviour in the first sessions tended to be more persistent independently of rewards, while it became increasingly guided by rewards in later sessions.

These results indicate that long-term changes in behaviour occurred both in terms of performance and win-shift. Because an increase in exploration comes at the expense of picking the best action less often, there is a reciprocal relationship between these two measurements, which makes it impossible to say whether the changes resulted from an increase in learning rate or a decrease in exploration. Computational modelling can help us more precisely address this issue.

3.2 | A forgetting Q-learning model with fixed parameters is incapable of replicating the experiment

Reinforcement learning provides a framework to disentangle learning and exploration effects on behaviour. These models rely on continuously updating estimates for the value of the different possible actions, so-called Q-values. One of the most popular of these algorithms, Q-learning, states that given a trial t during which the agent performs action a_t and receives a reward r_t, the learning rule should be written as:

$$Q(a_t) \leftarrow Q(a_t) + \alpha.(r_t - Q(a_t))$$

The learning rate, α, determines the impact the immediate outcome has on the previous estimate. The higher it is, the more heavily the new information weighs in the current value resulting in faster learning with the risk of undesirable volatility in a stochastic environment. We added a forgetting mechanism through which the values of the two unchosen levers decrease towards 0, the initial value all levers were set at:

$$Q(a_{\sim t}) \leftarrow (1 - \alpha_2).Q(a_{\sim t})$$

with α_2 the forgetting rate. Thanks to this additional mechanism, the fit between model simulations and experimental data in Cinotti et al. (2019) – which used the same task structure – was significantly improved, and this was again found to be the case here (data not

shown). This mechanism has been linked to persistence independently of reinforcement by Katahira (2015), with values of the forgetting rate α_2 smaller than α causing increased persistence. Finally, at each trial, the decision process is modelled using a softmax function of the Q-values:

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

in which the inverse temperature β determines the level of randomness in the exploration-exploitation trade-off by increasing the contrast between the action with the highest Q-value and the other competing options.

We first examined how well this model fitted experimental data when we allowed the acquired Q-values to carry over from one session to the next. We therefore optimised the three parameters – α, α_2 and β – with initial Q-values of 0, which were not reset in between sessions thus allowing Q-values to gradually build up. We then ran unconstrained simulations of this model using the same sequences of blocks but allowing the model to choose its actions at each trial randomly based on the softmax equation, rather than constraining it to the actions made by the corresponding animal. As shown in Figure 3a,b, the simulated average performance and win-shift are very different from the corresponding experimental curves of Figure 2. In the case of performance, although there is a significant session effect (repeated-measures ANOVA: $F[3,69] = 29.2$, $p < 10^{-4}$), the differences between sessions are not only far smaller but also inconsistent with the experimental data. In particular, performance in the last six sessions is worse than in earlier sessions, in complete contradiction with the experimental data. Concerning win-shift, the simulated data completely fail to reproduce the effect of sessions present in the experimental data. This analysis rules out the possibility that inter-session effects are a simple effect of accumulated learning without any changes to the model parameters.

3.3 | Animals do not use Dynamic Thompson Sampling to regulate exploration

Thompson sampling provides a normative strategy for balancing exploration and exploitation and has recently gained popularity in neuroscience for stationary tasks with normal reward probability distributions (Gershman, 2018). After observing the inter-session decline in win-shift, we tested whether this strategy might explain our results. Briefly, when applied to a bandit task with binomial rewards, Thompson sampling consists of replacing single-point Q-values estimating the expected reward of an action with a Beta distribution defining the Bayesian posterior probability distribution over the expected reward of each action. This richer representation provides the agent with an estimate of his subjective uncertainty of the value of each arm, which is leveraged to guide his decision-making. Given three posterior density functions for the three levers in the task with distinct sets of shape parameters $(\alpha,\beta)_{1,2,3}$, three random samples are drawn and the lever corresponding to the highest sample is chosen. Actions that have been sampled relatively little will have wide posterior distributions, increasing the likelihood of sampling a large value that will favour their selection. The outcome that is observed is then used to update the shape parameters of the corresponding lever:

$$\begin{cases} \alpha_i \leftarrow \alpha_i + r_t \\ \beta_i \leftarrow \beta_i + 1 - r_t \end{cases}$$

This results in a narrowing of the beta distribution, which quickly converges around the true probability of reward of that lever. Of course, in this experiment where reward probabilities change at the end of blocks of 24 trials, an extra mechanism must be introduced to avoid merely estimating the average reward over the whole course of the experiment. Dynamic Thompson sampling or DTS (Gupta et al., 2011) solves this problem by introducing a threshold C, which determines the minimum variance of the beta distributions. A similar solution is given by the Sliding Window Thompson Sampling (SWTS) algorithm which, as the name indicates, only counts rewarded and non-rewarded trials in a sliding window of size T (Trovo et al., 2020). We tested these two models by calculating their log-likelihood for different values of C and T for each individual. For comparison, we also calculated the log-likelihood of a naive model selecting each lever with a fixed probability of 1/3, which has a simple log-likelihood of $\sum_{i=1}^{N_{trials}} log\left(\frac{1}{3}\right)$. Two representative plots of these individual log-likelihoods are shown in Figure 4a,d. for the DTS and SWTS methods, respectively. In this example, the negative log-likelihood of DTS decreases rapidly for small values of C and then decreases far more slowly as C further increases. SWTS on the other hand has a concave shape, with the best value of T being the lowest bound of our tested values. These observations were not made for all subjects, with some for instance having an identifiable best value for C. Nevertheless, most importantly, and this applied to all individuals, for all tested values of C and T, the negative log-likelihood is above that of the naive model,

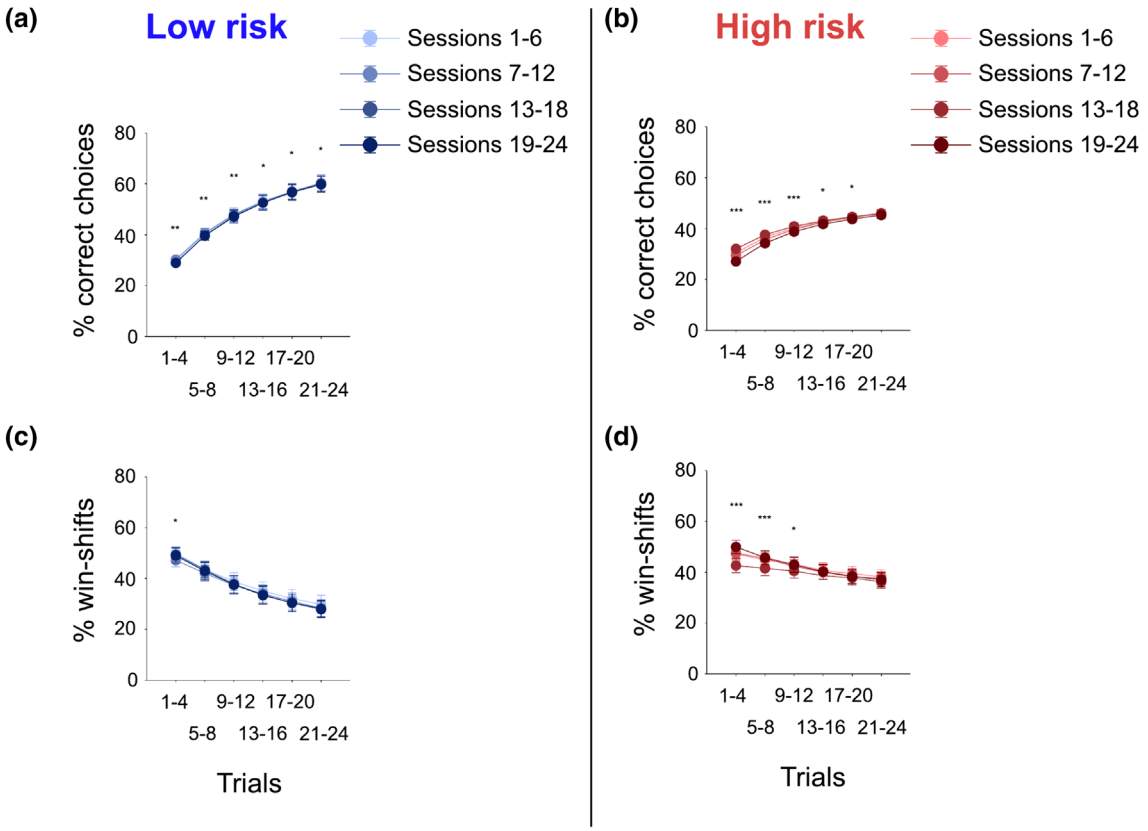

FIGURE 3 Simulations of the forgetting Q-learning model. (a,b) Mean performance ± s.e.m. (n = 24 average simulations) in low- and high-risk blocks, respectively. (c,d) Mean win-shift ± s.e.m. in low- and high-risk blocks, respectively. Stars indicate that there is a significant difference between at least two groups of sessions for a given bin of four trials. Significance levels as follows: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

which means that these models are worse at predicting the choice really made by the rats than randomly selecting a lever with probability 1/3. This disastrous performance can be explained by looking at simulations of these models solving this task for different parameter values in Figure 4b,c. for DTS and Figure 4e,f for SWTS. For the DTS, the performance of these methods at the beginning of blocks is far worse than animals at around 20%, but performance for small values of C quickly increases to reach levels above 80% and 60% in low-risk and high-risk blocks, respectively, which are far better than the final performance levels of animals. Increasing the value of C results in a lower final performance curve, but the increase in performance remains linear in time, contrary to the experimental data. Because this increase effectively means allowing the Beta distributions to become narrower, this effect probably betrays the tendency of the posterior density functions to track the average reward probabilities over longer time scales than the 24 trial blocks. Concerning the SWTS, a size window of just two trials produce a surprisingly good average performance of about 50 and 40% in low-risk blocks and high-risk blocks, respectively, but the simulations reach this biologically plausible value of performance far faster than experimentally observed. Increasing T allows the model to learn more slowly at the cost of again increasing final performance well above what the animals achieved. In conclusion, it seems that the reason why DTS and SWTS are so bad at fitting the data lies precisely in how much more efficient they are in finding and honing in on the correct lever compared to real animal subjects, justifying their use as normative rather than descriptive models of behaviour. We thus discarded these two models from the remaining analyses.

3.4 | Distinct learning rates and inverse temperatures but not forgetting rates can replicate animal behaviour

Having ruled out the possibility that these session effects are simply due to accumulated learning, we tested our starting hypothesis that the exploration-exploitation trade-off is regulated by optimising a "staggered β" model, which has six parameters: a learning

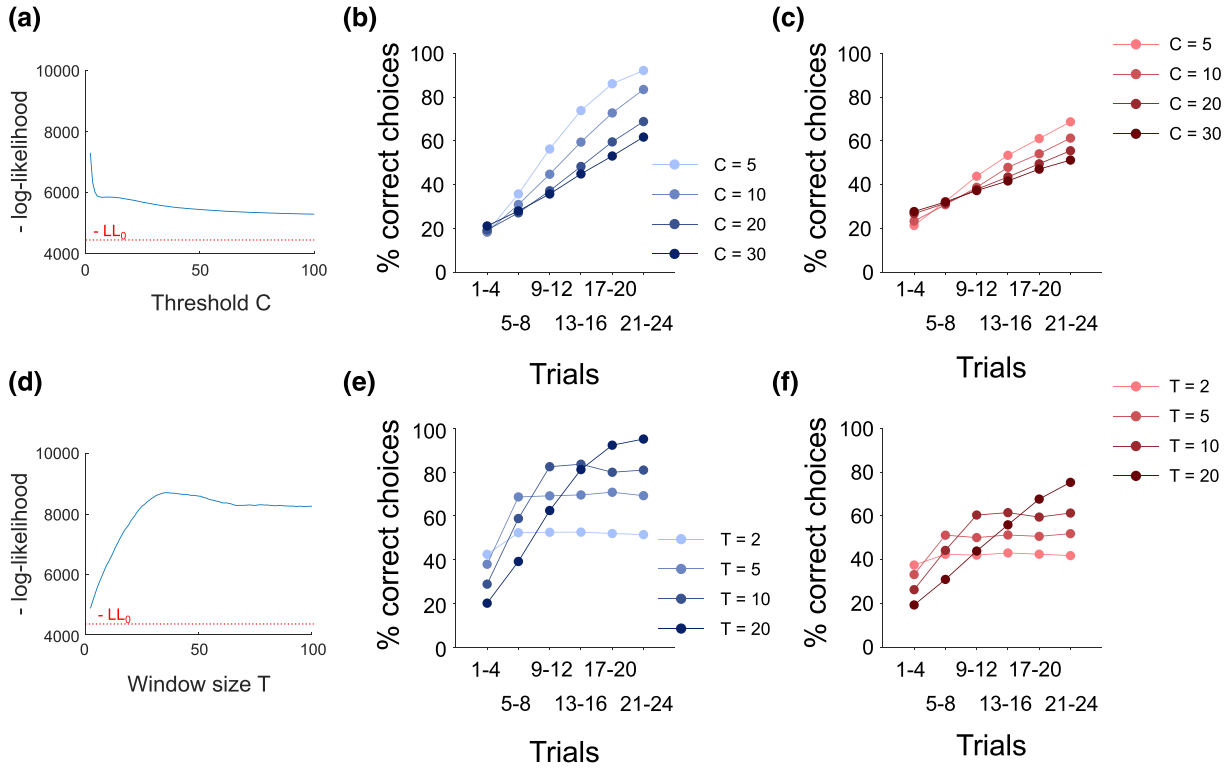

FIGURE 4 (a) Negative log-likelihood of Dynamic Thompson Sampling (DTS) for a representative individual as a function of the threshold parameter C. LL_0: log-likelihood of a naïve model selecting a random lever with probability one-third on each trial. (b) Simulated performance of DTS in low-risk blocks for different values of C. (c) Simulated performance of DTS in high-risk blocks for different values of C. (d) Negative log-likelihood of Sliding Window Thompson Sampling (SWTS) for a representative individual as a function of T, the number of trials in the sliding window. LL_0: log-likelihood of a naïve model selecting a random lever with probability one-third on each trial. (e) Simulated performance of SWTS in low-risk blocks for different values of T. (f) Simulated performance of SWTS in high-risk blocks for different values of T.

rate and forgetting rate, and four inverse temperatures β_1, β_2, β_3 and β_4 for sessions 1–6, 7–12, 13–18 and 19–24, respectively. Another possibility is that, rather than the exploration-exploitation trade-off, it is one of the other two parameters that is changing between sessions. For this reason, we also optimised a "staggered α" model and a "staggered α_2" model. We compared these models and the previously described forgetting QL model using the Akaike (AIC) and Bayesian Information Criterion (BIC) shown in Figure 5a,b, respectively. Despite being more heavily penalised for having three extra parameters, all three staggered models have lower AIC and BIC scores than the forgetting QL model with fixed parameters between sessions. Between the three staggered models, the staggered α model has the lowest AIC (17789) and BIC (17880) scores, while the staggered α_2 model is very slightly better than the staggered β model.

Models with better optimisation scores may actually prove less good when simulated (Palminteri et al., 2017; Wilson & Collins, 2019), which is why we also ran simulations of these models for comparison. Contrary to simulations of the simple forgetting Q-learning model, simulated performance of the staggered α and β models were significantly different between sessions for later trials of both low- and high-risk blocks (Figure 6a,b and Figure 7a,b) in accordance with the experimental results. The simulated win-shift curves of these two models also fitted experimental data well, with win-shift in the first six sessions being significantly higher than in subsequent sessions (Figure 6c,d and Figure 7c,d). By contrast, despite having a better AIC and BIC than the staggered β model, the staggered α_2 model produced simulations that clearly failed to capture the inter-session changes of interest (Figure 8). These simulations were more similar to the forgetting QL model, with no effect of sessions on win-shift difference, and effects on performance that were most pronounced in the first trials of high-risk blocks rather than the last. Given the fact that α_2 controls persistence (Katahira, 2015) and that we had previously found a decrease in reward-independent persistence (Figure 2e), this is a surprising result that suggests that a dynamic α_2 parameter is not sufficient to capture rats' behaviour in our task.

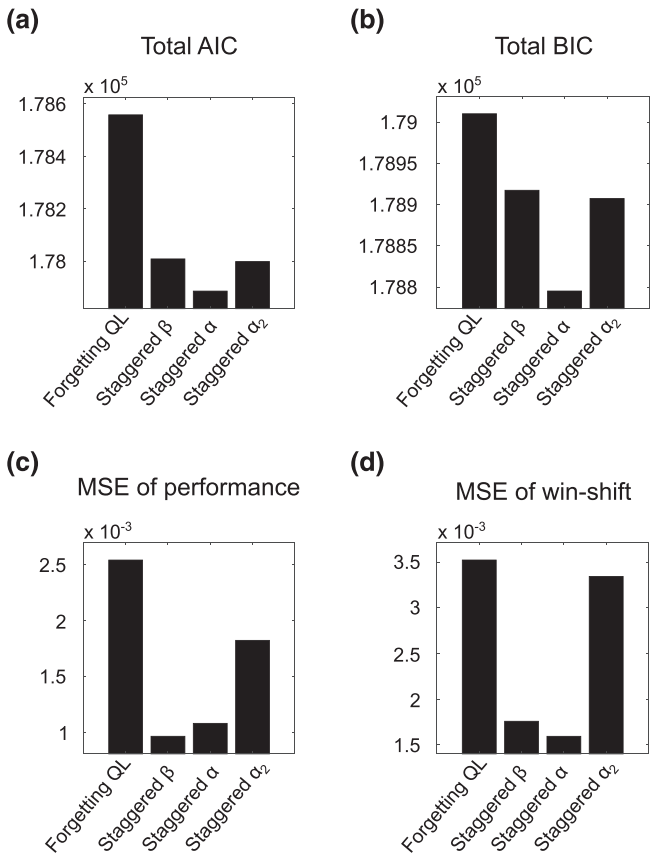

FIGURE 5 Comparison of the forgetting Q-learning and staggered models using (a) the Akaike Information Criterion (AIC) summed over all subjects, (b) the Bayesian Information Criterion (BIC) summed over all subjects, (c) the Mean Squared Error (MSE) of average simulated performance with respect to experimental performance, and (d) the MSE of average simulated win-shift with respect to experimental win-shift.

Thus, allowing either the learning rate or inverse temperature to adapt between sessions while keeping other parameters fixed is sufficient to replicate the animals' improvements in performance and decrease in win-shift. The between-session evolution of individual learning rates and inverse temperatures are plotted in Figures 6e and 7e, respectively. There is a significant effect of sessions on the learning rate of the staggered α model (Friedman ANOVA: $p = 6.10^{-6}$), and the optimised values of α in the first six sessions were indeed significantly smaller than for all other sessions (Figure 6e). This would mean that the changes in performance and win-shift can be explained by a gradual increase in the learning rate, session after session, so that Q-values change more rapidly with respect to feedback in later sessions. According to Figure 7e, the ability of the staggered β model to replicate the experiment is also attributable to an increase in the inverse temperature (Friedman ANOVA, sessions effect: $p = 0.004$) as β in the first six sessions is significantly smaller than in sessions 7–12 and 19–24. This result indicates that long-term changes in behaviour might also be explained by an increased tendency to exploit the action with the highest Q-value. In contrast to the staggered α model, where the increase in α seems linear (see black average curve in Figure 6e), β seemed to quickly converge to a plateau after an initial increase between the first six sessions and the next ones. Variations of the forgetting rate for the staggered α_2 model are shown in Figure 8e, and have no discernible pattern, in line with the fact that this model is unable to replicate inter-session changes.

To quantify how well these different models fitted the original data, we computed for each the mean-squared error, or MSE, as detailed in the methods, which measures the average distance between each point of a simulated curve and a reference experimental curve. Results are shown in Figure 5c,d for performance and win-shift, respectively, and reveal that the staggered β model provides the best fit to performance, while it is the staggered α model that is best at fitting win-shift. As expected, the staggered α_2 and forgetting QL give the worst fits to both measurements.

3.5 | Models of meta-learning based on an average reward rate

In Cinotti et al. (2019), we showed that dopamine inhibition causes an increase in random exploration without impacting learning. In addition, tonic dopamine has been hypothesised to integrate the reward prediction errors and thus represent an average reward rate. For these reasons, we were inspired to design a meta-learning model in which random exploration, which is set by the parameter β, is controlled by a running average reward rate R_t:

$$R_{t+1} = R_t + \alpha_R.(r_t - R_t)$$

with α_R the reward rate learning parameter. Because trial outcomes are either 1 or 0 and $\alpha_R < 1$, R_t is itself bounded between 0 and 1. We then model the dependence of β on this reward rate as a simple linear function as in Blackwell and Doya (2023):

$$\beta_t = \beta_0 + m.R_t$$

The slope m of this function was not constrained to be positive, so the tendency to exploit might either increase or decrease with the average reward rate.

Having shown the potential of the staggered α model, we also optimised a meta-learning on α model in which it is the learning rate rather than the inverse

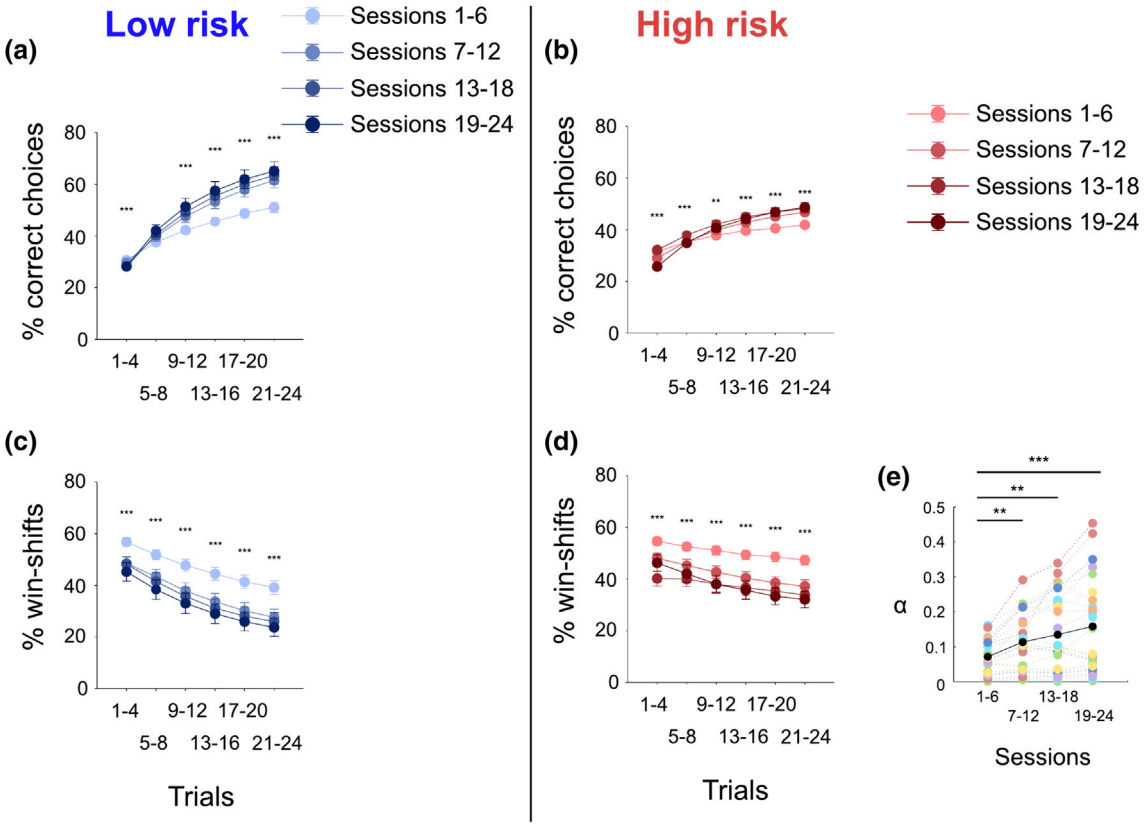

FIGURE 6 Simulations of the staggered α model. (a,b) Mean performance ± s.e.m. (n = 24 average simulations) in low- and high-risk blocks, respectively. (c,d) Mean win-shift ± s.e.m. in low- and high-risk blocks, respectively. Stars indicate a significant difference between at least two groups of sessions for a given bin of four trials. (e) Variations of the optimised values of the learning rate between sessions. Coloured dashed lines represent individuals, and the black line is the mean. Significance levels as follows: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

temperature, which evolves as a linear function of the average reward rate:

$$\alpha_t = \alpha_0 + m.R_t$$

Just as with the meta-learning on β model, the slope was not constrained to be positive or negative. This means that as the average reward rate increases, α could either increase or decrease depending on the values optimised for each subject. A similar α_2 meta-learning model was also optimised.

An alternative to these meta-learning models is that rats were simply increasing their tendency to exploit or their learning rate over time irrespective of their performance (Lloyd et al., 2023; Moin Afshar et al., 2020). To confront our meta-learning models to these alternatives, we also optimised monotonic trial-dependent regulation models of α and β based on the shape of the evolution of these parameters according to the staggered models. For β, because this parameter appears to grow more quickly at the beginning of sessions before converging (Figure 7e), we designed a trial-dependent geometric increase model in which:

$$\beta_t = \beta_{t-1} + m.(\beta_{max} - \beta_{t-1})$$

This model has five unknown parameters to optimise, the learning and forgetting rates α and α_2, the initial value of the inverse temperature β_0, the maximum value to which it converges β_{max}, and the rate m at which it approaches this maximum. Other trial-dependent regulation of β models such as a linear function of trial and a logarithmic function were tested but we retained this one as the best amongst these options.

In contrast, the learning rate apparently increases more linearly over time (Figure 6e) so that a linear trial-dependent function seemed more appropriate for this parameter:

$$\alpha_t = \alpha_0 + m.t$$

which has just four parameters, α_0, m, β and α_2. Other tested models not presented here include meta-learning models with a sigmoid function of the reward rate and meta-learning models based on the running average of squared RPEs as an estimate of uncertainty. All models were optimised using the same procedure given in the

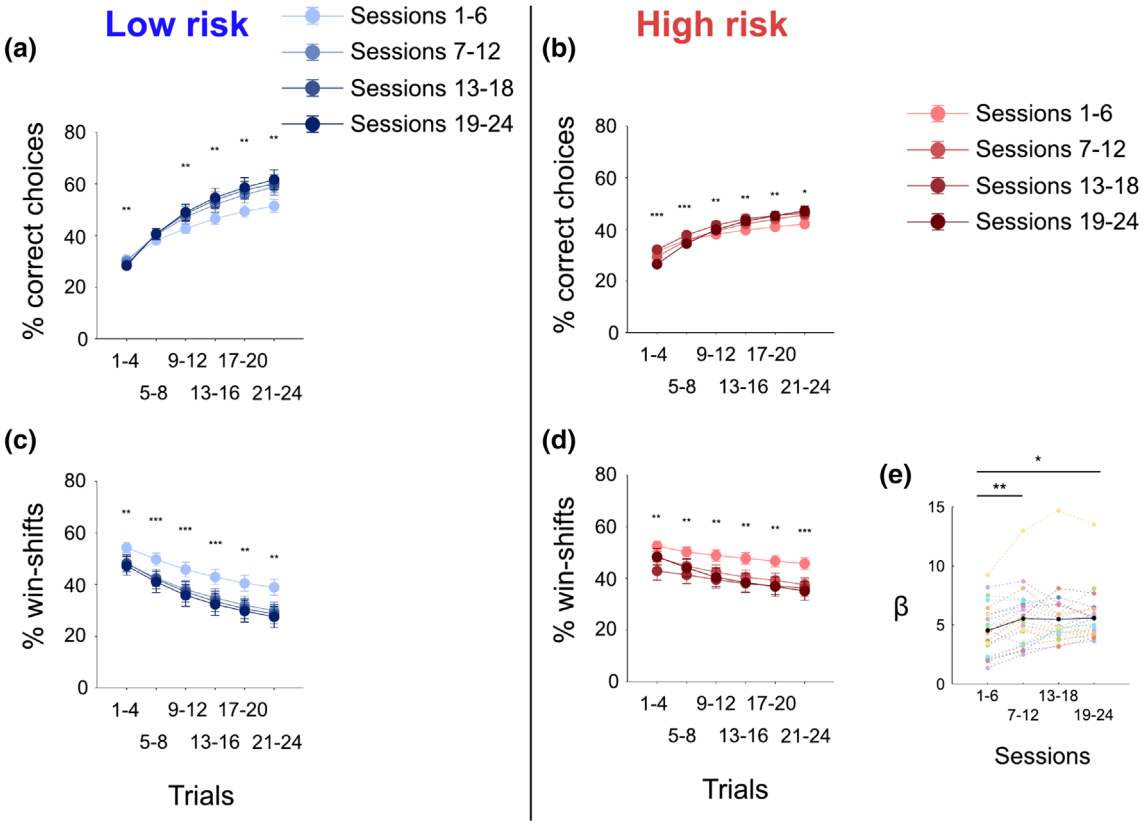

FIGURE 7 Simulations of the staggered β model. (a,b) Mean performance ± s.e.m. (n = 24 average simulations) in low- and high-risk blocks, respectively. (c,d) Mean win-shift ± s.e.m. in low- and high-risk blocks, respectively. Stars indicate a significant difference between at least two groups of sessions for a given bin of four trials. (e) Variations of the optimised values of the inverse temperature between sessions. Coloured dashed lines represent individuals, and the black line is the mean. Significance levels as follows: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

Methods and were then compared using the AIC (Figure 12a) and BIC (Figure 12b), with the previously discussed staggered α model also included as a reference. According to both criteria, the best model was the β meta-learning model, and the worst was the α_2 meta-learning model. However, after running and averaging 100 simulations with optimised parameters, the model with the best MSE for both performance and win-shift was in fact the α meta-learning model (Figure 12c,d). The distributions of optimised parameters of the α and β meta-learning models are shown in Figure 13. The slope m of the β meta-learning model was found to be positive for all but two individuals, meaning that as the average reward rate increases, so does the inverse temperature leading to increased exploitation. In the case of the α meta-learning model, the situation was more ambivalent, with a majority of 16 rats for which the slope was similarly positive. This means Q-values update more quickly as the reward rate increases. But the opposite was true for the remaining 8 rats. For both meta-learning models, the reward rate learning parameter α_R tended to be very small (median: 0.0032 and 0.0243 for α and β meta-learning, respectively) in comparison with the standard learning rate α (median α_0: 0.023 and 0.056 for α and β meta-learning, respectively), indicating that the reward rate changes far more slowly over time than the Q-values. This explains the ability of these models to capture between-session changes, as shown later. Finally, the forgetting rates α_2 had very similar distributions between the two models and were usually greater than the learning rates (median 0.22 and 0.19 for α and β meta-learning, respectively), which reflects a strong tendency to persevere independently of reinforcement.

Simulations of the α meta-learning model, plotted in Figure 9, successfully reproduced the between-session increases in performance and decline in win-shift. In addition, when we ran a logistic regression predicting the probability of choices on each simulated trial given the past ten choices and rewards, we found that choices in the first six sessions were far less sensitive to the last reward than in later sessions, as observed experimentally (Figure 9e right). Contrary to the experimental data, however, these simulations did not exhibit any obvious decrease in the effect of past choices (Figure 9e left). In fact, the regression coefficients for these effects were

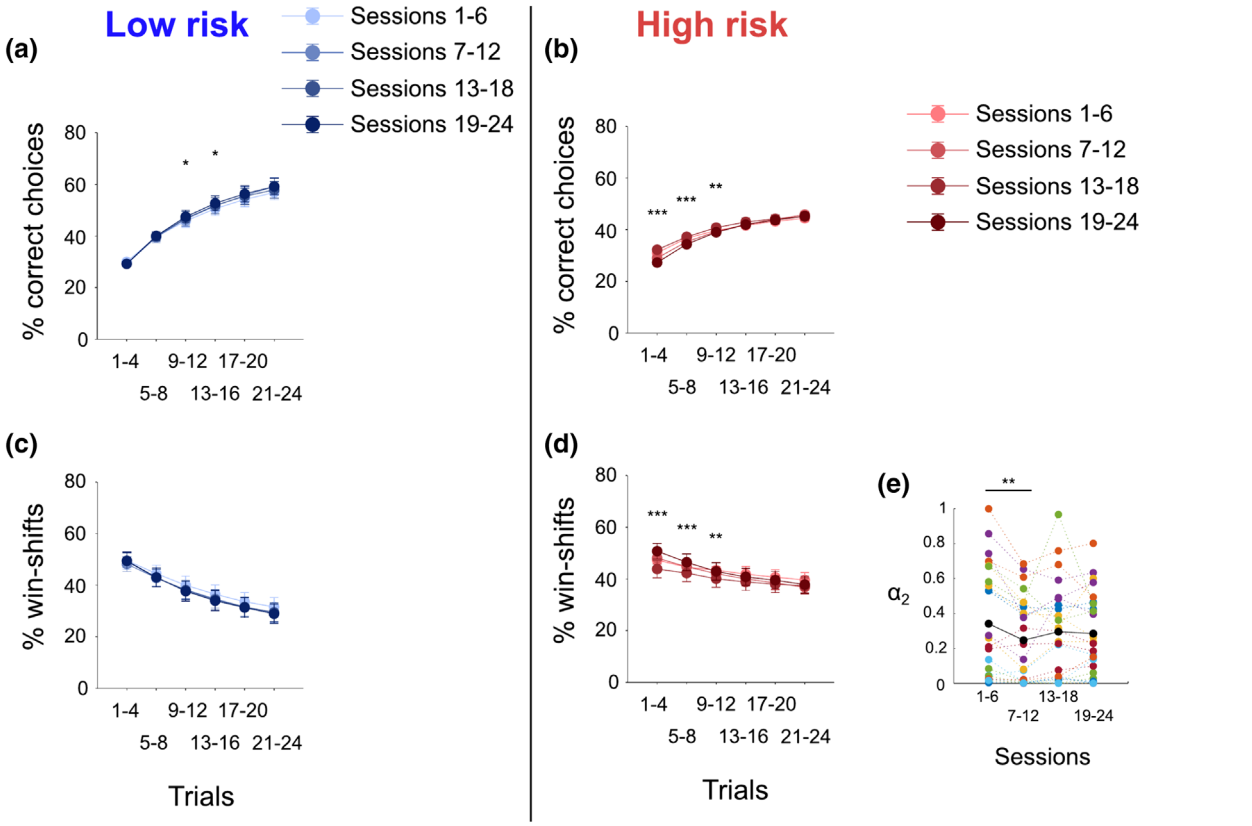

FIGURE 8 Simulations of the staggered α_2 model. (a,b) Mean performance ± s.e.m. (n = 24 average simulations) in low- and high-risk blocks, respectively. (c,d) Mean win-shift ± s.e.m. in low- and high-risk blocks, respectively. Stars indicate a significant difference between at least two groups of sessions for a given bin of four trials. (e) Variations of the optimised values of the forgetting rate between sessions. Coloured dashed lines represent individuals, and the black line is the mean. Significance levels as follows: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

smaller in size for all sessions than in the experimental data, suggesting that the absence of decrease was due to a floor effect and that an additional tendency to persistence was present in the experimental data, which was absent in the model.

Despite having a slightly worse total MSE, the β meta-learning model was also very good at generating simulations similar to the original data (Figure 10), with similar improvements in performance and a decrease in win-shift in both low- and high-risk blocks. In terms of the logistic regression coefficients (Figure 10e), although there was a tendency for the last reward to more significantly bias decision-making as the experiment progressed between sessions, this increase was less marked. Whereas the curve of the coefficients over the last 10 rewards tended to be quite flat for both the experimental and simulated α meta-learning model in sessions 1–6, there is already a stronger impact of the most recent rewards for the β meta-learning model in these early sessions. Concerning past choices, the evolution of these coefficients between sessions was in fact opposite to that observed experimentally as the tendency to persist actually increased. In addition, as with the α meta-learning model, the magnitude of the effects of past trials seemed far smaller with log-odds all below 0.1.

Finally, simulations of the α_2 meta-learning model (Figure 11) did not replicate the inter-session effects of interest. They showed very little inter-session changes, apart from an unexpected decline in performance in the first trials of high-risk blocks (Figure 11b), in parallel to an increase in win-shift on these same trials (Figure 11d). Logistic regression coefficients also did not show much variation in between sessions (Figure 11e).

In a last attempt to separate the two standout meta-learning models, we computed individual MSE by comparing individual simulations to individual performance and win-shift curves (not shown). These comparisons gave us distributions of MSE plotted in Figure 12e,f, which could then be compared statistically. For both performance and win-shift, we found no statistically significant difference in MSE of the α and β meta-learning model, using either a paired t-test ($p = 0.64$ and $p = 0.89$ for comparisons of performance and win-shift MSE, respectively) or a Wilcoxon signed-rank test ($p = 0.81$ and $p = 0.84$ for performance and win-shift MSE, respectively).

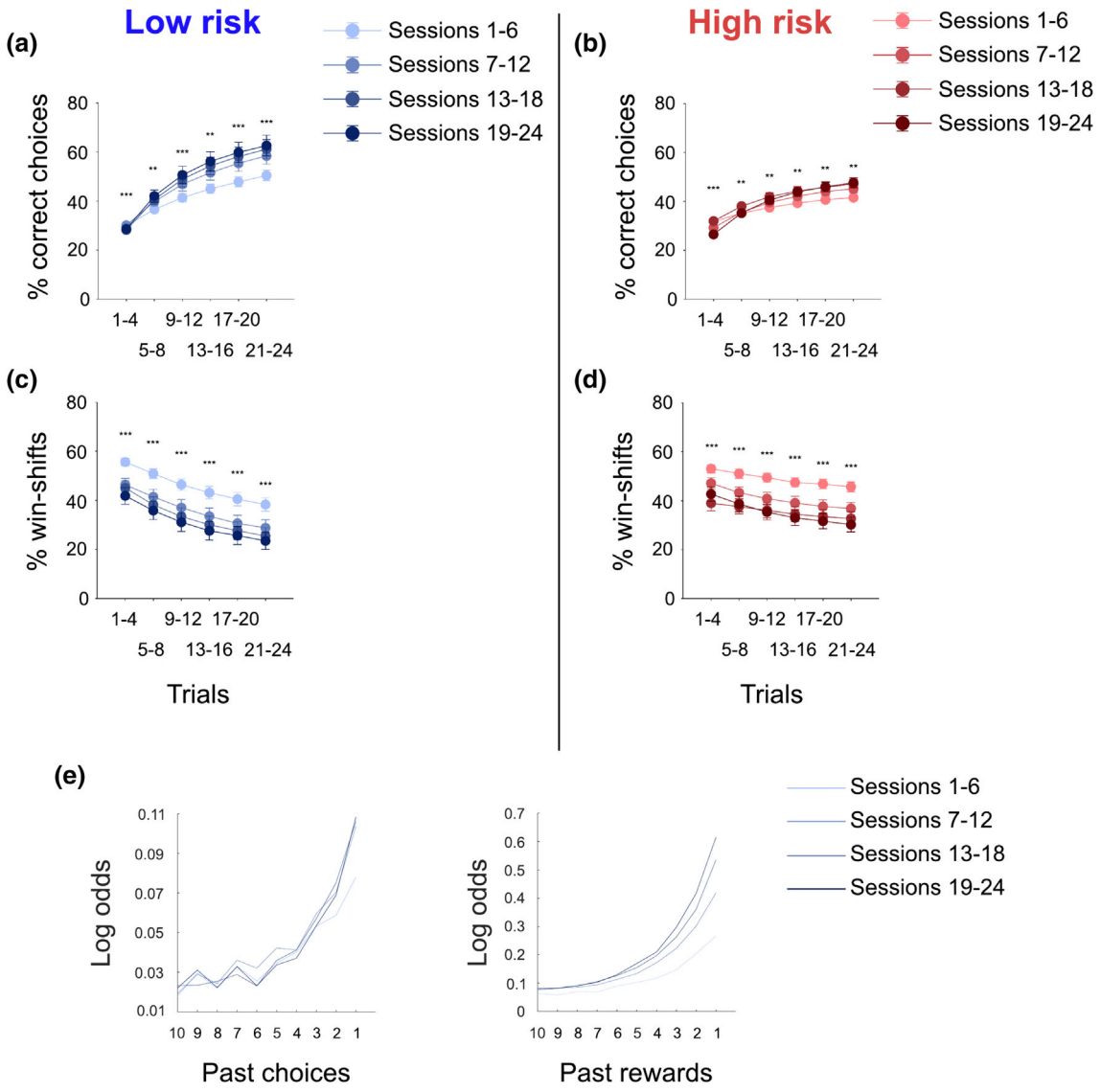

FIGURE 9 Simulations of the α meta-learning model. (a,b) Mean performance ± s.e.m. (n = 24 average simulations) in low- and high-risk blocks, respectively. (c,d) Mean win-shift ± s.e.m. in low- and high-risk blocks, respectively. Stars indicate a significant difference between at least two groups of sessions for a given bin of four trials. (e) Average logistic regression weights for effects of past choices and rewards on current trial. Significance levels as follows: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

4 | DISCUSSION

In this work, we compared the ability of different computational models to account for rats' progressive tuning of reinforcement learning parameters while they were learning the structure of a three-armed bandit task. Our task included three levers with different reward probabilities, and two risk conditions: a low-risk condition and a high-risk condition. The task was moreover non-stationary in that the reward probabilities of the levers changed without signal every 24 trials.

We found that rats' performance significantly improved within- and between-sessions and that performance improvement was sharper in low-risk conditions. We moreover found that the percentage of exploratory trials (i.e., win-shift trials after a rewarded choice of the correct lever) was higher during the first 6 sessions, without further significant changes during the remaining 18 sessions. This indicated that either the exploration-exploitation trade-off or the trade-off between learning fast or slow was progressively learned and stabilised in adaptation to the task. Such behavioural tendencies cannot be captured by a standard reinforcement learning model with fixed parameters. Instead, we found that a meta-learning model, which linearly tunes either the inverse temperature or the learning rate parameter, based on variations in the average reward rate, provided the best account of these long-term variations in rats'

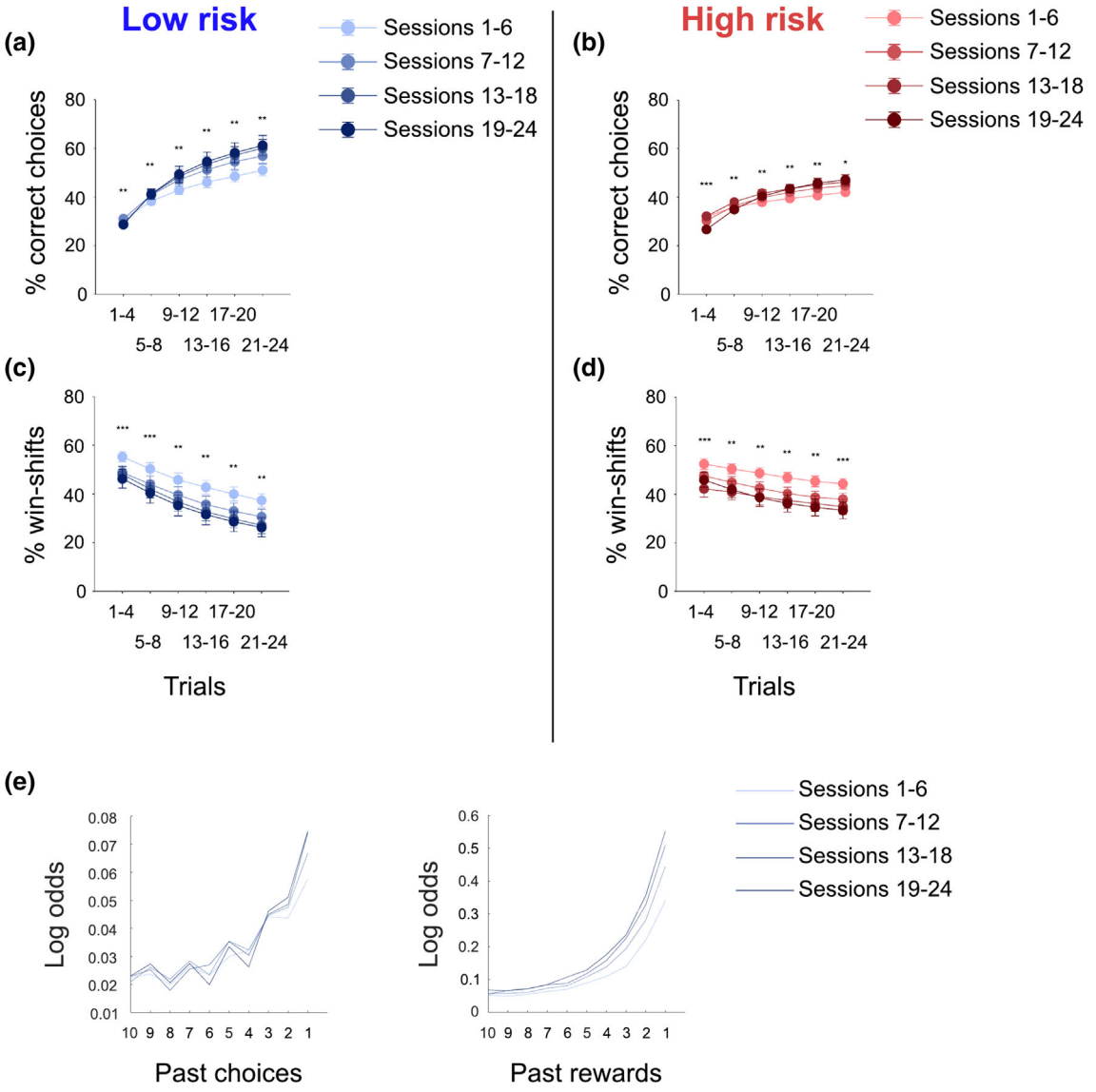

FIGURE 10 Simulations of the β meta-learning model. (a,b) Mean performance ± s.e.m. (n = 24 average simulations) in low- and high-risk blocks, respectively. (c,d) Mean win-shift ± s.e.m. in low- and high-risk blocks, respectively. Stars indicate a significant difference between at least two groups of sessions for a given bin of four trials. (e) Average logistic regression weights for effects of past choices and rewards on current trial. Significance levels as follows: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

behaviour. We further confirmed these modelling results by model simulations and analyses. These results suggest that rats progressively tune their reinforcement learning parameters while learning the structure of new decision-making tasks.

In this study, we tested the hypothesis that rat long-term behavioural adaptation could be captured by a meta-learning process dynamically tuning reinforcement learning parameters session after session. We initially focused on meta-learning concerning the inverse temperature, which regulates the exploration-exploitation trade-off. In making this choice, we pursued a line of inquiry begun by Humphries et al. (2012), a theoretical study which presented a model of the basal ganglia. In that model, the entropy of action selection, i.e. random exploration, decreased with average dopamine levels. This hypothesis was investigated experimentally in Cinotti et al. (2019), where we showed that systemic pharmacological inhibition of dopamine enhanced exploration without affecting the learning rate. Together with the assumption that tonic dopamine represents the average reward rate (Hamid et al., 2016; Niv et al., 2007), this leads to the idea that the reward rate might control exploration through tonic dopamine levels. Another possibility is that it is the learning rate or the forgetting rate or a combination of these parameters that is being regulated over time. While rat behaviour in our task was inconsistent with the model with meta-learning of the forgetting

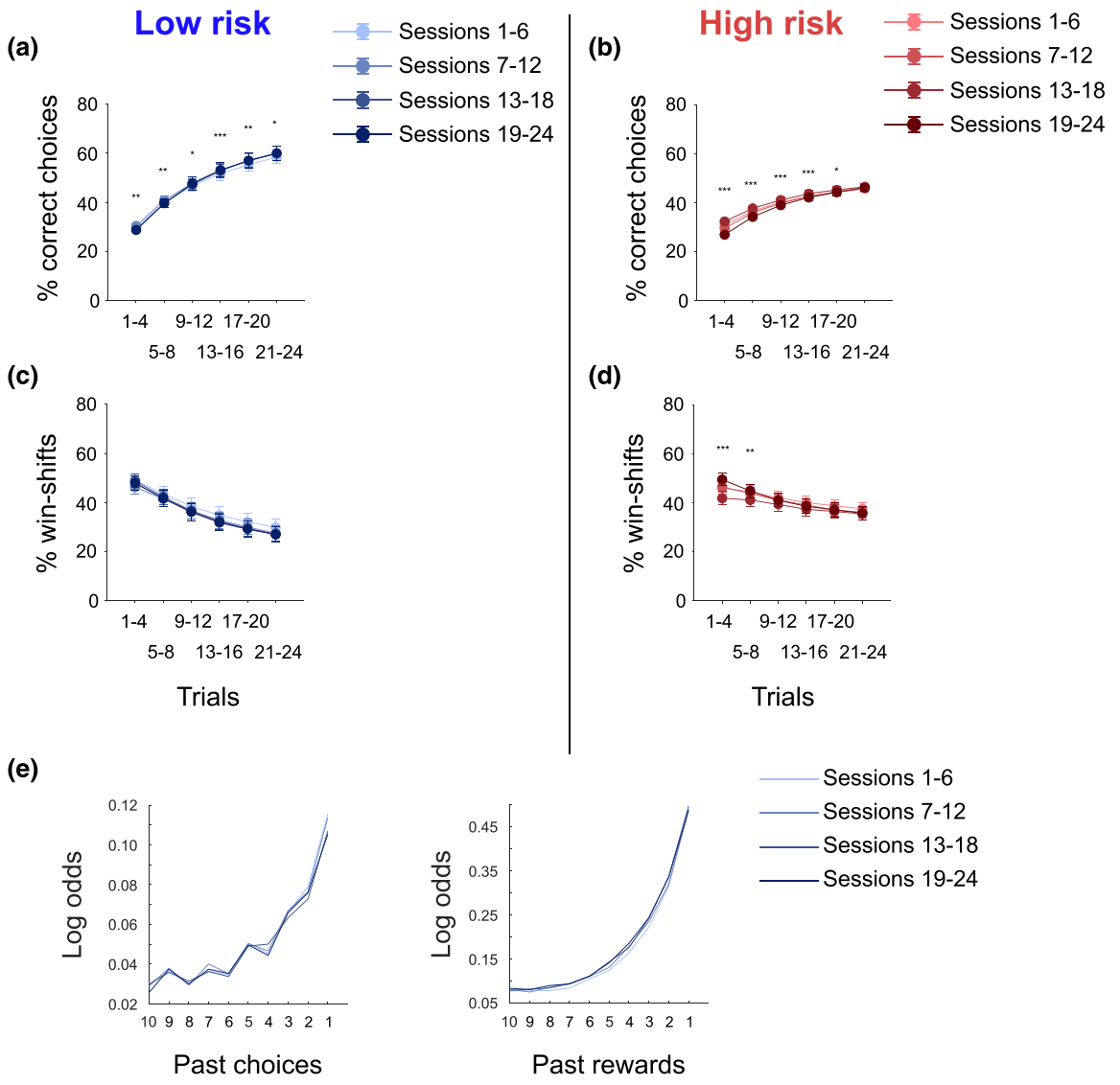

FIGURE 11 Simulations of the α_2 meta-learning model. (a,b) Mean performance ± s.e.m. (n = 24 average simulations) in low- and high-risk blocks, respectively. (c,d) Mean win-shift ± s.e.m. in low- and high-risk blocks, respectively. Stars indicate a significant difference between at least two groups of sessions for a given bin of four trials. (e) Average logistic regression weights for effects of past choices and rewards on current trial. Significance levels as follows: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

rate, we found that the model with meta-learning of the learning rate could not be discarded. While the meta-learning model of the inverse temperature led to better AIC and BIC scores, the post-optimisation simulations revealed that meta-learning of the learning rate led to better MSE fits (Figure 12). In terms of simulations, we found no way to separate the two models. However, regulation of the learning rate has previously been linked to task volatility (Behrens et al., 2007) or uncertainty (Jepma et al., 2016) rather than the reward rate, and might depend on a different neurotransmitter than dopamine such as serotonin (Iigaya et al., 2018) or noradrenaline (Jepma et al., 2016). The difficulty we encountered in separating meta-learning on learning rate or inverse temperature may be due to the fact that online estimation of uncertainty, like the reward rate, is dependent on the past history of rewards, so there could be a large overlap between the two signals. Taken together, these data point towards a larger class of meta-learning models in which uncertainty controls the learning rate and the reward rate the inverse temperature.

Interestingly, we also tested Thompson Sampling (TS) models and showed that they failed to capture rat behaviour in our task. While this class of models have recently been successfully applied to stationary tasks with normal reward probability distributions (Gershman, 2018), the non-stationarity of our task, combined with binomial reward probability distributions,

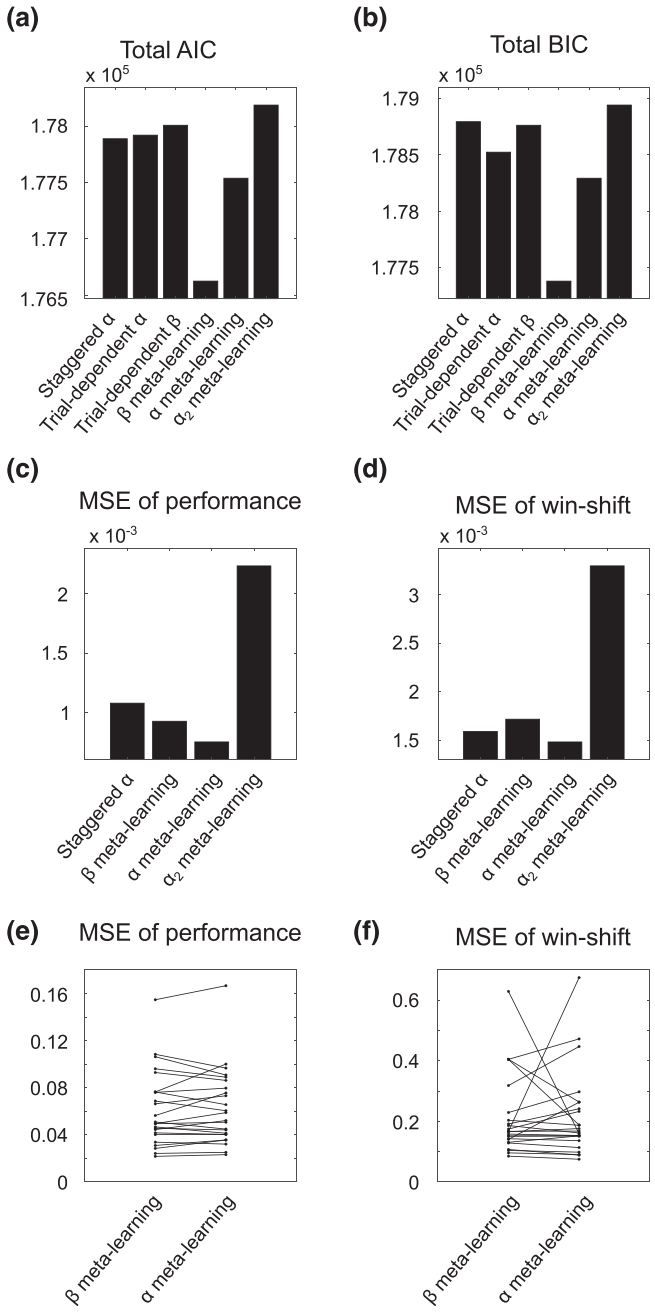

FIGURE 12 Comparison of the trial-dependent and meta-learning models using (a) the Akaike Information Criterion (AIC) summed over all subjects, (b) the Bayesian Information Criterion (BIC) summed over all subjects, (c) the Mean Squared Error (MSE) of average simulated performance with respect to experimental performance, and (d) the MSE of average simulated win-shift with respect to experimental win-shift. (e) Distribution of individual MSEs for performance of α and β meta-learning models. No significant difference was found. (f) Distribution of individual MSEs for win-shift of α and β meta-learning models. No significant difference was found.

required extensions of classical TS models. We tested a sliding-window TS model and a dynamic TS model, and showed that they either learned too fast or converged to too high performance plateaus, depending on the employed parameters.

The idea that an increase in reward rate should cause a change in reinforcement learning parameters could have important implications in another field of decision-making, the transition from goal-directed to habitual behaviour. Goal-directed behaviour is characterised by flexibility, the ease with which an organism adjusts behaviour when its goal is manipulated (Robinson & Berridge, 2013). On the other hand, animals display habitual behaviour when they repeat previously reinforced actions, even when these actions are no longer rewarded or are even punished. This is particularly relevant for the study of addiction, which could, partly, be explained by habitual modes of behaviour struggling for control with higher-level goal-directed decision-making (Everitt & Robbins, 2005; Redish et al., 2008). The computational account for these two types of behaviour usually hinges on assigning habitual behaviour to a slower model-free learning process such as the Q-learning algorithm, and goal-directed behaviour to a model-based learning algorithm in which the organism relies on a representation of the task or environment structure to guide its actions (Daw et al., 2005). The transition from goal-directed to habitual behaviour could be explained as a reduction in computational complexity when a certain level of performance is achieved. The meta-learning model we presented offers another possible and complementary explanation. In the first phase in which an action reliably produces a reward, the accumulation of rewards causes either an increase in the inverse temperature or a decrease in the learning rate, alongside the increase of the value of that action. If the link between action and reward is altered, the now very strong tendency to stabilise learning and exploitation will cause the animal to persevere longer in repeating that action despite its falling value. This is because, as shown through the slow inter-session effect on behaviour contrasted with the fast and efficient evolution of behaviour within blocks, the dynamics of the inverse temperature and learning rate are potentially much slower than those of Q-values. Hence, an action could see its Q-value fall dramatically, and still be selected. Of course, this increased perseverance should occur only as long as the Q-value of the previously rewarded action remains above any alternative actions, the inverse temperature blindly favouring whichever action currently has the highest value.

A slow evolution of the inverse temperature or learning rate could explain a puzzling lack of effect on the risk level of blocks. As the reward rate is lower in high-risk blocks, we would expect this to have an effect on exploration in addition to that on learning. Indeed, performance

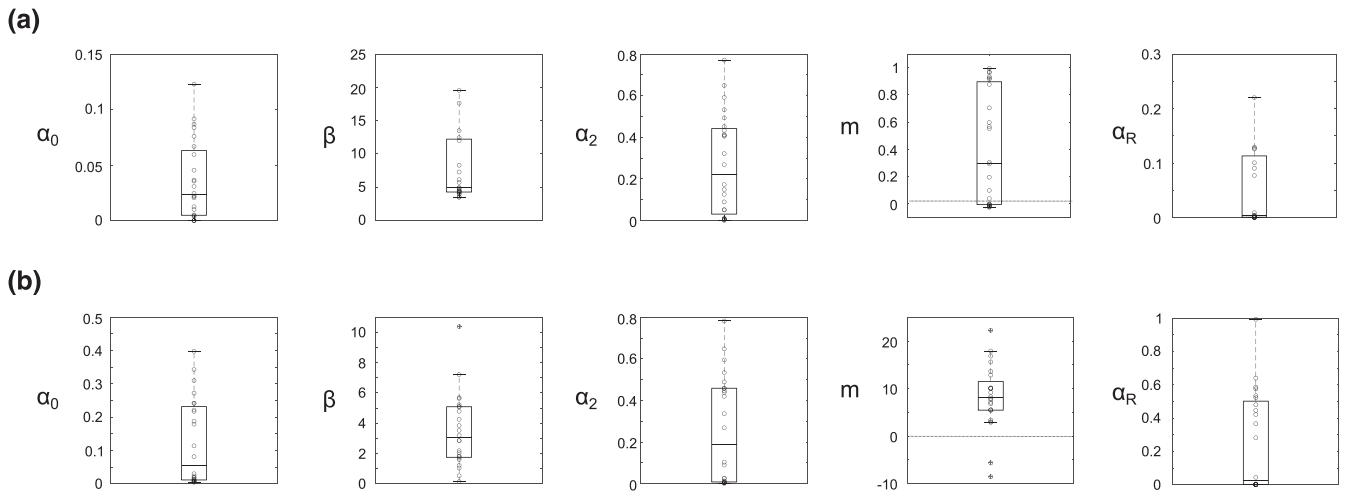

FIGURE 13 Parameter distributions of the (a) α meta-learning model and (b) the β meta-learning model. Dots represent single individuals, box plots show the median value, interquartile range and extreme values excluding outliers (red crosses).

and win-shift are different in high- and low-risk blocks, but simulations of a model with fixed inverse temperature and learning rate are entirely capable of producing this type of behaviour (Figure 3), so that differences in the Q-values in the two types of blocks are sufficient explanation. Furthermore, we also optimised a model with separate inverse temperatures for high- and low-risk blocks, a strategy previously used by Eisenegger et al. (2014) to compare human populations with different type-2 dopamine receptors, and did find significantly higher optimised values in low-risk blocks, consistent with increased exploitation as predicted by the model (analyses not shown). However, in a counterfactual test where we optimised the same model with separate inverse temperatures based on block risk level on data simulated with a model using a single inverse temperature, we also found significant differences meaning that such methods are not to be blindly trusted without carrying out appropriate counterfactual checks. Maybe if the blocks were longer than 24 trials, then variations in exploration between risk conditions could unambiguously be detected. To also distinguish meta-learning from time-related increases in exploitation, a possible experimental design would be to alternate low-risk and high-risk periods for greater amounts of trials, perhaps even entire sessions. We could then perhaps detect changes in behaviour following long periods of low reward rates corresponding to a predicted decrease in exploitation, which would contradict the effect of time only.

In this paper, we propose that meta-learning – here restricted to learning of meta-parameters (Khamassi et al., 2011; Wang, 2021) – is an adaptive mechanism that enables flexibility in variable environments. Counter this proposal, Findling et al. (2019) and Findling et al. (2020) showed that computational noise in the estimation of Q-values paired with argmax selection mechanisms can explain a majority of non-greedy choices in human decision-making tasks, which are usually classified as exploratory decisions and which are in fact, according to this hypothesis, due to learning errors. Crucially, they propose that the noise introduced in the estimation of Q-values be proportional to the RPEs. Thus, an unstable environment in which RPEs fluctuate greatly causes greater reversals in Q-values and produces more seemingly explorative behaviour. While such a model could explain variations between environments with different reward probabilities, such as between the high-risk and low-risk blocks of our experiment, it is hard to see how they could explain long-term changes between sessions, which all contain the same number of high-risk and low-risk blocks and thus have the same average volatility. A potential explanation might be that the relationship between RPEs and the variance of the computational noise is evolving over time, which could be the subject of future research.

In our β meta-learning model, we restricted ourselves to the hypothesis that it is a random or undirected exploration that might be regulated between sessions. An alternative form of exploration consists of deliberately sampling choices with high uncertainty about their value. This is called directed exploration and can be modelled with an augmentation of Q-values with an uncertainty bonus (Velentzas et al., 2017; Wilson et al., 2014). Interestingly, a study by Gershman and Tzovaras (2018) links both directed and random exploration to dopamine, with higher dopamine in the prefrontal cortex associated with a stronger bias towards uncertain actions, and, contrary to our expectation, higher dopamine in the striatum associated with less random exploration.

In the initial paper of Humphries et al. (2012), the mechanism by which dopamine might regulate the exploration-exploitation trade-off was situated in the basal ganglia, in which tonic dopamine modulates the contrast between activities related to alternative actions represented within parallel channels. Alternatively or complementarily, several lines of evidence point to the prefrontal cortex as responsible for exploratory decisions (Frank et al., 2009). Maybe this is limited to the previously mentioned directed exploration but Hattori et al. (2023) found that inhibition of plasticity in the orbitofrontal cortex in mice impaired between session improvements in performance on a probabilistic reversal task. These experimental results were mirrored by a deep reinforcement learning model with meta-learning capabilities in which a fast reinforcement learning component controls trial-by-trial decisions, while a slow reinforcement learning critic changes connection weights in between sessions. It is also possible that these meta-learning capabilities are not dependent on dopamine but on another neuromodulator like noradrenaline (Doya, 2002).

Overall, this work constitutes one of the rare attempts to account for rats' progressive adjustment of their reinforcement learning parameters while they are learning the structure of a new task. This contributes to a promising line of research, which could help better understand why animals behave according to a precisely tuned trade-off between exploration and exploitation, or between learning fast or slow, in the post-training phases of decision-making tasks.

AUTHOR CONTRIBUTIONS

Francois Cinotti: Formal analysis; investigation; methodology; software; visualization; writing—original draft; writing—review and editing. **Etienne Coutureau:** Conceptualization; funding acquisition; methodology; supervision; writing—review and editing. **Mehdi Khamassi:** Conceptualization; formal analysis; funding acquisition; investigation; methodology; project administration; software; supervision; writing—original draft; writing—review and editing. **Alain Marchand:** Conceptualization; investigation; methodology; writing—review and editing. **Benoît Girard:** Conceptualization; formal analysis; investigation; methodology; supervision; writing—review and editing.

ACKNOWLEDGEMENTS

This work was partially supported by the French Agence Nationale de la Recherche (ANR) "Learning Under Uncertainty" Project under reference ANR-11-BSV4-006 and "Neurobehavioral assessment of a computational model of reward learning" CRCNS 2015 Project under reference ANR-15-NEUC-0001.

Virginie Fresno carried out the experimental study.

The preparation of this paper was supported through a writing retreat funded by the Agriculture, Food and Health research theme at the University of Reading.

CONFLICT OF INTERESTS STATEMENT

The authors declare that they have no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/ejn.16449.

DATA AVAILABILITY STATEMENT

All code for analysis and modelling was written in MATLAB and is available on Github (https://github.com/frct/Metalearning) together with the original data.

ORCID

François Cinotti https://orcid.org/0000-0003-2921-0901
Etienne Coutureau https://orcid.org/0000-0001-6695-020X
Mehdi Khamassi https://orcid.org/0000-0002-2515-1046
Alain R. Marchand https://orcid.org/0000-0002-0231-5562
Benoît Girard https://orcid.org/0000-0002-8117-7064

REFERENCES

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. Available at:. https://doi.org/10.1038/nn1954

Blackwell, K. T., & Doya, K. (2023). Enhancing reinforcement learning models by including direct and indirect pathways improves performance on striatal dependent tasks. *PLOS Computational Biology*. Edited by M.B. Cai, *19*(8) Available at:, e1011385. https://doi.org/10.1371/journal.pcbi.1011385

Cazé, R. D., & Van Der Meer, M. A. A. (2013). Adaptive properties of differential learning rates for positive and negative outcomes. *Biological Cybernetics*, *107*(6), 711–719. Available at:. https://doi.org/10.1007/s00422-013-0571-5

Cinotti, F., Fresno, V., Aklil, N., Coutureau, E., Girard, B., Marchand, A. R., & Khamassi, M. (2019). Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Scientific Reports*, *9*(1) Available at:, 6770. https://doi.org/10.1038/s41598-019-43245-z

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 933–942. Available at:. https://doi.org/10.1098/rstb.2007.2098

Daw, N. D. (2011). Trial-by-trial data analysis using computational models: (Tutorial review). In M. R. Delgado, E. A. Phelps, & T. W. Robbins (Eds.), *Decision making, affect, and learning:*

Attention and performance XXIII (pp. 3–38). Oxford University Press. Available at:. https://doi.org/10.1093/acprof:oso/9780199600434.003.0001

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. Available at:. https://doi.org/10.1038/nn1560

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. Available at:. https://doi.org/10.1038/nature04766

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks: the Official Journal of the International Neural Network Society*, *15*(4–6), 495–506. Available at:. https://doi.org/10.1016/s0893-6080(02)00044-8

Eisenegger, C., Naef, M., Linssen, A., Clark, L., Gandamaneni, P. K., Müller, U., & Robbins, T. W. (2014). Role of dopamine D2 receptors in human reinforcement learning. *Neuropsychopharmacology*, *39*(10), 2366–2375. Available at:. https://doi.org/10.1038/npp.2014.84

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience*, *8*(11), 1481–1489. Available at:. https://doi.org/10.1038/nn1579

Findling, C., Chopin, N., & Koechlin, E. (2020). Imprecise neural computations as a source of adaptive behaviour in volatile environments. *Nature Human Behaviour*, *5*(1), 99–112. Available at:. https://doi.org/10.1038/s41562-020-00971-z

Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2019). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature Neuroscience*, *22*(12), 2066–2077. Available at:. https://doi.org/10.1038/s41593-019-0518-9

Findling, C., & Wyart, V. (2021). Computation noise in human learning and decision-making: Origin, impact, function. *Current Opinion in Behavioral Sciences*, *38*, 124–132. Available at:. https://doi.org/10.1016/j.cobeha.2021.02.018

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, *12*(8), 1062–1068. Available at:. https://doi.org/10.1038/nn.2342

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*(August 2017), 34–42. Available at:. https://doi.org/10.1016/j.cognition.2017.12.014

Gershman, S. J., & Tzovaras, B. G. (2018). Dopaminergic genes are associated with both directed and random exploration. *Neuropsychologia*, *120*(July), 97–104. Available at:. https://doi.org/10.1016/j.neuropsychologia.2018.10.009

Gilbertson, T., & Steele, D. (2021). Tonic dopamine, uncertainty and basal ganglia action selection. *Neuroscience*, *466*, 109–124. Available at:. https://doi.org/10.1016/j.neuroscience.2021.05.010

Gupta, N., Granmo, O.-C., & Agrawala, A. (2011). Thompson sampling for dynamic multi-armed bandits. In *in 2011 10th International Conference on Machine Learning and Applications and Workshops. 2011 Tenth International Conference on Machine Learning and Applications (ICMLA)* (pp. 484–489. Available at:). IEEE. https://doi.org/10.1109/ICMLA.2011.144

Hamid, A. A., Pettibone, J. R., Mabrouk, O. S., Hetrick, V. L., Schmidt, R., Vander Weele, C. M., Kennedy, R. T., Aragona, B. J., & Berke, J. D. (2016). Mesolimbic dopamine signals the value of work. *Nature Neuroscience*, *19*(1), 117–126. Available at:. https://doi.org/10.1038/nn.4173

Hart, A. S., Rutledge, R. B., Glimcher, P. W., & Phillips, P. E. M. (2014). Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *The Journal of Neuroscience*, *34*(3), 698–704. Available at:. https://doi.org/10.1523/JNEUROSCI.2489-13.2014

Hattori, R., Hedrick, N. G., Jain, A., Chen, S., You, H., Hattori, M., Choi, J. H., Lim, B. K., Yasuda, R., & Komiyama, T. (2023). Meta-reinforcement learning via orbitofrontal cortex. *Nature Neuroscience*, *26*(12), 2182–2191. Available at:. https://doi.org/10.1038/s41593-023-01485-3

Humphries, M., Khamassi, M., & Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in Neuroscience* Available at:, *6*, 9. https://doi.org/10.3389/fnins.2012.00009

Humphries, M. D., & Gurney, K. (2007). A means to an end: Validating models by fitting experimental data. *Neurocomputing*, *70*(10–12), 1892–1896. Available at:. https://doi.org/10.1016/j.neucom.2006.10.061

Iigaya, K., Fonseca, M. S., Murakami, M., Mainen, Z. F., & Dayan, P. (2018). An effect of serotonergic stimulation on learning rates for rewards apparent after long intertrial intervals. *Nature Communications*, *9*(1), 2477. Available at:. https://doi.org/10.1038/s41467-018-04840-2

Jepma, M., Murphy, P. R., Nassar, M. R., Rangel-Gomez, M., Meeter, M., & Nieuwenhuis, S. (2016). Catecholaminergic regulation of learning rate in a dynamic environment. *PLOS Computational Biology*. Edited by J.X. O'Reilly, *12*(10) Available at:, e1005171. https://doi.org/10.1371/journal.pcbi.1005171

Katahira, K. (2015). The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior. *Journal of Mathematical Psychology*, *66*, 59–69. Available at:. https://doi.org/10.1016/j.jmp.2015.03.006

Khamassi, M., Wilson, C. R. E., Rothé, M., Quilodran, R., Dominey, P. F., & Procyk, E. (2011). Meta-learning, cognitive control, and physiological interactions between medial and lateral prefrontal cortex. In R. Mars, et al. (Eds.), *Neural bases of motivational and cognitive control*. MIT Press. https://doi.org/10.7551/mitpress/8791.003.0025

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, *84*(3), 555–579. Available at:. https://doi.org/10.1901/jeab.2005.110-04

Lebarbier, E., & Mary-huard, T. (2006). Une introduction au critère BIC: fondements théoriques et interprétation. *Journal de la Société Française de Statistique*, *147*(1), 39–57.

Lloyd, A., Viding, E., McKay, R., & Furl, N. (2023). Understanding patch foraging strategies across development. *Trends in Cognitive Sciences*, *27*(11), S1364661323001729. Available at:. https://doi.org/10.1016/j.tics.2023.07.004

Moin Afshar, N., Keip, A. J., Taylor, J. R., Lee, D., & Groman, S. M. (2020). Reinforcement learning during adolescence in rats. *The Journal of Neuroscience*, *40*(30), 5857–5870. Available at:. https://doi.org/10.1523/JNEUROSCI.0910-20.2020

Niv, Y. (2007). Cost, benefit, tonic, phasic: What do response rates tell us about dopamine and motivation? *Annals of the New York Academy of Sciences*, *1104*(1), 357–376. Available at:. https://doi.org/10.1196/annals.1390.018

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology*, *191*(3), 507–520. Available at:. https://doi.org/10.1007/s00213-006-0502-4

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. Available at:. https://doi.org/10.1016/J.TICS.2017.03.011

Redish, A. D., Jensen, S., & Johnson, A. (2008). A unified framework for addiction: Vulnerabilities in the decision process. *Behavioral and Brain Sciences*, *31*(4), 415–437. Available at:. https://doi.org/10.1017/S0140525X0800472X

Robinson, M. J. F., & Berridge, K. C. (2013). Instant transformation of learned repulsion into motivational "wanting". *Current Biology*, *23*(4), 282–289. Available at:. https://doi.org/10.1016/j.cub.2013.01.016

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schweighofer, N., & Doya, K. (2003). Meta-learning in reinforcement learning. *Neural Networks*, *16*(1), 5–9. Available at:. https://doi.org/10.1016/S0893-6080(02)00228-9

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 9) (p. 1054). MIT Press. https://doi.org/10.1109/TNN.1998.712192

Trovo, F., Paladino, S., Restelli, M., & Gatti, N. (2020). Sliding-window Thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, *68*, 311–364. Available at:. https://doi.org/10.1613/jair.1.11407

Velentzas, G., Tzafestas, C., & Khamassi, M. (2017) 'Bridging computational neuroscience and machine learning on non-stationary multi-armed bandits'. Available at: https://doi.org/10.1101/117598

Wang, J. X. (2021). Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, *38*, 90–95. Available at:. https://doi.org/10.1016/j.cobeha.2021.01.002

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547. Available at:. https://doi.org/10.7554/eLife.49547

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of Experimental Psychology. General*, *143*(6), 2074–2081. Available at:. https://doi.org/10.1037/a0038199

How to cite this article: Cinotti, F., Coutureau, E., Khamassi, M., Marchand, A. R., & Girard, B. (2024). Regulation of reinforcement learning parameters captures long-term changes in rat behaviour. *European Journal of Neuroscience*, 1–22. https://doi.org/10.1111/ejn.16449