Regulation of the exploration-exploitation trade-off captures long-term changes in rat behaviour

Cinotti, Coutureau, Khamassi*, Marchand*, Girard*

*contributed equally to this work.

Keywords: Meta-learning, decision-making, exploration-exploitation trade-off, dopamine, reinforcement learning

CONFLICT OF INTERESTS STATEMENT

The authors declare that they have no conflicts of interest.

ABSTRACT

In uncertain environments in which resources fluctuate continuously, animals must permanently decide whether to exploit what they currently believe to be their best option, or instead explore potential alternatives in case better opportunities are in fact available. 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 progressively 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 stabilization of the way they regulated the exploration-exploitation trade-off. We successfully captured these behavioural adaptations using a meta-learning model in which the exploration-exploitation trade-off is controlled by the animal's average reward rate.

INTRODUCTION

Faced with an uncertain environment in which resources fluctuate continuously, animals must permanently decide whether to exploit what they currently believe to be their best option, or instead explore potential alternatives in case better opportunities are in fact available. This trade-off between exploration and exploitation (Cohen, McClure and Yu, 2007) could itself be tuned to the circumstances: If the animal is currently experiencing a high reward rate, then it would seem in its best interest to keep exploiting its current strategy, whereas if the reward rate drops, this could be a signal that it is time to start exploring new strategies.

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 and 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 and 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 which the decision-making process has sampled, and conversely, if a learning deficiency makes actions less discriminable, then even a perfectly working decision-making mechanism will produce noisy outcomes. Therefore, the current predominance of theories on the regulation of learning should not close the door to a potential role of the regulation of exploration as a mechanism of adaptation to the environment.

Reinforcement learning (Sutton and Barto, 1998), a class of algorithms for learning what actions to take based on discrete outcomes in the form of rewards and punishment, is 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.

In a previous paper (Cinotti *et al.*, 2019), we indeed showed through careful modelling of rat behavioral adaptation to changing reward probabilities 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 the reward prediction errors which are crucial to reinforcement learning (Schultz, Dayan and Montague, 1997; Hart *et al.*, 2014). In addition, it has been postulated to carry information about uncertainty (Gilbertson and 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 (Khamassi *et al.*, 2011; Humphries, 2012).

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 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 structure, and whether this resulted in a stabilization of the way they regulated the exploration-exploitation trade-off during the post-training phase of the task. We successfully captured these behavioural adaptations using a meta-learning model (Schweighofer and Doya, 2003) in which the exploration-exploitation trade-off is controlled by the animal's average reward rate.

METHODS

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 x 26 x 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 x 30 cm deep x 35 cm high, Imetronic, Pessac, France), each located inside a sound and light-attenuatng wooden compartment (74 x 46 x 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 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 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 sessions of the probabilistic choice task, 20 sessions of six trial blocks each and four double sessions of 12 blocks each.

The experimental task (Figure 1) 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 unsignalled 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 (or four times in the last four double sessions).

Data analysis

In order to smoothen the appearance of block average performance and win-shift, trials were binned into groups of 4 trials. In the case of win-shift, the average was obtained by pooling across blocks all potential win-shift events belonging to a given bin (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 bin number within blocks. *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 2 c and d) for which we instead report that the interaction between risk and session was significant and compare average win-shift over all bins instead as shown in Figure 2 c and d. These same methods were used when analysing the different model simulations.

Model fitting

All models tested relied on a softmax action selection process which defines the trial-by-trial likelihood of the model (Daw, 2011):

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

Likelihood over the entire experiment is then defined as the product of these trial likelihood, and the log-likelihood as the sum of the log-likelihood of each trial. Models were optimised through maximisation of the log-likelihood using the built-in *fmincon* function in MATLAB which implements a gradient descent method to find the minimum of the negative log-likelihood. 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 five-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.

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 precise aims of these criteria (Lebarbier and Mary-huard, 2006).

Ultimately, models were judged by their ability to produce simulations similar to the original experimental data (Humphries and Gurney, 2007; Palminteri, Wyart and Koechlin, 2017). For each individual, we ran 100 simulations using the optimised set of parameters and the same block schedule as the one experienced by the subject. 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 whether they reproduced between session changes, and on their mean squared errors relative to the original average curves:

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

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

When studying the separate impacts of Q-values and regulated inverse temperature of the linear meta-learning model, we also ran constrained simulations in which the Q-values and β are updated at each trial according to the choices and outcomes made by the individual subjects and the optimised parameter values of that individual.

For the staggered model, we tested the effect of sessions on 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 by Ahmed BenSaïda), and the assumption of sphericity was also violated according to a Mauchly test ($p = 1.10^{-4}$).

Data and code availability

All code for analysis and modelling was written in MATLAB. Data and code are available from the corresponding author on reasonable request.

RESULTS

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 $3/16^{th}$ 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 2 a and b, we tracked the rats' average performance, 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: $p < 10^{-6}$) demonstrating that rats were unaware that a block change had occurred and were persisting with the previously best rewarded option. As rats learned which was 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% in 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 ($F(5,115) = 112.1$, $p<10^{-4}$), session($F(3,69)=29.4$, $p<10^{-4}$), and risk ($F(1,23)=98.4$, $p<10^{-4}$) as within-subjects factors; all possible combinations of these three main factors were also significant ($p<0.0463$). *Post hoc* Bonferroni 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 another 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 expected increased performance in low-risk blocks compared to the high-risk blocks, the results reveal a long-term improvement in performance.

Win-shift, an explorative strategy, also changed significantly throughout the experiment. It consists in the probability of changing lever, after being rewarded for a correct choice of the current best lever. As depicted in Figure 2 c and 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 (F(3,69)=6.1, p=0.0021) interaction 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.

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 solve this issue.

Variations of β

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:

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

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 which may cause undesirable volatility in a stochastic environment. In order to improve model fitting to the data, we added a forgetting mechanism as previously in Cinotti et al. (2019) through which the values of the two unchosen levers decrease towards 0, the initial value all levers were set at:

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

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:

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

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 others. Following our hypothesis that animals might regulate their exploration-exploitation trade-off, it is this parameter that will attract our interest.

A first possibility in explaining the long-term changes in behaviour is that the Q-values must first reach an average baseline value which is different from the initial value at the start of the experiment. To test this hypothesis, we optimised this forgetting Q-learning model with initial Q-values of 0 and carried over the Q-values from one session to the next, allowing for the gradual build-up of Q-values in between sessions. Using the optimised parameters 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 3 a and b, the simulated average performance and win-shift curves are very noticeably different from the corresponding experimental curves in 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 session

in complete contradiction with the experimental data. Similarly concerning win-shift, the simulated data completely fail to reproduce the effect of sessions present in the experimental data.

Having ruled out the possibility that these session effects are a simple effect of accumulated learning, we approached our hypothesis that the exploration-exploitation trade-off is regulated by optimising the same forgetting Q-learning model with the only difference that four different β values were used for each group of six sessions we arbitrarily divided the experiment into. We compared this model, which we call the staggered model, to the previous one using the Akaike (AIC) and Bayesian Information Criterion (BIC) reported in Tables 2 and 3. Despite its three extra parameters, we found that the staggered model had better scores on aggregate, i.e. when summing individual scores (total AIC = 178009 versus 178556 for the model with separate β values and the forgetting QL model respectively, total BIC = 178917 versus 179010). If we look into more details at Tables 2 and 3, only for one individual subject (rat 28) is the forgetting model the better fit according to the AIC, while for the BIC a majority (16) of subjects actually favour the forgetting model despite the total sum being smaller for the staggered model. We also ran simulations of this model and found that it was capable of qualitatively replicating experimental effects (Figure 4), a control analysis prescribed by the literature (Palminteri, Wyart and Koechlin, 2017; Wilson and Collins, 2019). Contrary to simulations of the simple forgetting Q-learning model, simulated performance of the staggered model was significantly different between sessions for later trials of both low- and high-risk blocks (Figure 4 a and b). In particular, performance in low-risk blocks of the first six sessions was significantly smaller than in sessions 12-18 and sessions 19-24. The simulated win-shift curves also fitted experimental data well with win-shift in the first six sessions being significantly higher than in subsequent sessions (Figure 4 c and d). In Table 4, we also report the mean squared errors which tells us how close the fit between the simulated and experimental curves is (see Methods) and find that for both performance and win-shift, mean-squared errors of the forgetting model are about twice as large as for the staggered model.

Thus, allowing the inverse temperature to adapt between sessions while keeping learning and forgetting rates fixed is sufficient to replicate the animals' improvements in performance and decrease in exploration. The evolution of individual inverse temperatures between sessions in Figure 4 e shows that there is a significant effect of sessions on the optimised values of the inverse temperature (Friedman ANOVA: p=0.0043). The optimised values of β in the first six sessions are indeed significantly smaller than for sessions 7-12 (p=0.0071) and sessions 19-24 (p=0,015). These changes correspond to increased exploitation, as expected from the decreased exploration obtained in simulations (Figure 4 c and d).

Meta-learning based on average reward rate model

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 hypothesized to integrate the reward prediction errors and thus represent an average reward rate. For these reasons, we were inspired in designing 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:

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

Because trial outcomes are either 1 or 0 and $\alpha_R < 1$, R_t is itself bounded between 0 and 1. Three different meta-learning variants were tested. The first is a simple linear function of the reward rate (Blackwell and Doya, 2023):

$$\beta_t = \beta_{min} + (\beta_{max} - \beta_{min}).R_t$$

Such a linear function has the advantage of being very simple, but sigmoid functions which converge slowly to finite limits could prove a better alternative which is why we designed two other sigmoid meta-learning variants. The first of these regulates β as follows:

$$\beta_t = \frac{\beta_{max}}{1 + e^{-k(R_t - 0.5)}}$$

Which describes a sigmoid function with 0 and βmax as limits at minus and plus infinity respectively, and a midpoint at $R_t = 0.5$. This variant has the same number of unknown parameters as the linear meta-learning model.

The second sigmoid variant is obtained by defining the limit at minus infinity as an extra unknow parameter:

$$\beta_t = \beta_{min} + \frac{\beta_{max} - \beta_{min}}{1 + e^{-k(R_t - 0.5)}}$$

An alternative to the meta-learning hypothesis is that rats were simply increasing their exploitation tendency over time irrespective of their performance (Moin Afshar *et al.*, 2020; Lloyd *et al.*, 2023). To test our hypothesis against this alternative, we also optimised two models in which β increases monotonically over time. Given the shape of the evolution of β in the staggered model (Figure 4 e), we designed a logarithmically increasing model and a geometrically increasing model. The inverse temperature for the logarithmic increase model is calculated as:

$$\beta_t = \beta_0 + k \log(a \times t)$$

With t the trial number, and $\beta 0$, k, and a three unknown parameters. As for the geometric increase, the inverse temperature is:

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

This model also 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 a at which it approaches this maximum. The crucial difference between these two models is that although both display the desired curved shape of increase, the former is a divergent function while the second converges to a known fixed value. The diverse mechanisms these models implement are shown in Figure 5.

All five models were optimised using the same procedure as previously described and were then compared together with the previous two models using the AIC (Table 2) and BIC (Table 3). According to the AIC, meta-learning models are better for 15 individuals the remaining 9 being better fit by a monotonically increasing model. Of the 15 individuals for which meta-learning is best, only 4 individuals favour the six-parameter variant of the sigmoid model. For no individual is the forgetting QL model the best model. For the BIC, 16 subjects are best fit by a meta-learning model, and only 5 by a monotonic increase. We also compared the seven models by running 100 simulations with optimised parameters then calculating the mean squared errors between average simulations and the experimental data (Table 4). For performance, the best fit was achieved with simulations of the linear meta-learning model, while the best fit of win-shift was instead by the two sigmoid variants, but crucially neither the geometric nor the logarithmic time-dependent models fitted the data better than

the linear meta-learning model. Overall, the linear meta-learning model has the lowest aggregate scores, both for the AIC and BIC, as well as the lowest mean-squared error for performance, which is why we selected this one for further study.

When simulated this model produced average performance and win-shift curves with significant session and risk effects as intended in Figure 6. Contrary to the forgetting Q-learning model and similarly to the staggered model, simulations of this model have improved performance and decreased exploration between sessions. Figures 7 and 8 which plot simulations for each risk and session group against the corresponding experimental data further illustrate the very good agreement between the simulated and experimental datasets.

Analysis of the effect of variations of the inverse temperature

One of the most interesting features of this novel meta-learning model is the separation of exploration from learning. This means there are potentially periods during which, despite similar Q-values, behaviour is noticeably different due to different values of the inverse temperature. We extracted from each trial the estimated Q-values and β_t by running simulations of the linear meta-learning model with optimised parameters constrained to the original experimental data, i.e. without letting the model generate its own actions, but instead letting it observe the actions and outcomes actually experienced by the subject. To disentangle the effect of Q-values and regulated inverse temperature, we synthesised the three Q-values estimated by the model into a single discriminability score that tells us how easy it is to recognise the biggest Q-value:

$$d_t = \frac{\max(Q_{i,t})}{\sum Q_{i,t}}$$

We then assessed the separate impacts of this discriminability score and of the current trial value of the inverse temperature using a logistic regression model fitting the success or failure to select what appears to be the lever with the highest Q-value, in other words to follow an exploitation strategy. We applied this analysis to all rats and to each rat separately. When applied to the population as a whole, we find as expected a highly significant (p = 0 according to the *fitglm* function of MATLAB) positive relationship between discriminability and the odds ratio of exploitation. In addition, there is also a very significant effect (p = 0) of the value of β on the odds ratio of exploitation with an estimated 9.6% increase in the odds ratio for each unit increase of β. With the exception of two rats for whom the effect of β was either non-significant or negative, similar results were obtained when the logistic model was fitted on a rat per rat basis. The lowest odds ratio increase per unit increase of beta was 3% and the largest 58%.

When applying a logistic regression model, there is a risk that predictors, in our case discriminability and β, are collinear. This would mean that we cannot distinguish the potential effect of these two predictors. To check for such an issue, we used Brian Lau's Collinearity Diagnostics Toolbox for MATLAB which calculates so-called condition indices. When this condition index is greater than 30, then a strong collinearity is possible. Applied to the population as a whole, the condition index for collinearity between discriminability and β was 2.7 indicative of only weak collinearity. When applied to rats separately, all condition indices ranged between 3.3 and 13.9, with only three individuals above the threshold of 10 which suggests only moderate collinearity. No individual condition index was above 30, the threshold for potentially strong collinearity.

To more concretely illustrate the effect of a varying inverse temperature, we ranked trials depending on their inferred discriminability binning trials according to whether discriminability was lower than 0.6, between 0.6 and 0.8, and higher than 0.8. By definition discriminability, which is the ratio of the highest Q-value over the sum of three Q-values, is greater than 0.3. For each individual, we further classified trials based on whether β was in the first, second or third tercile of the total range bounded by individual values of β_{min} and β_{max}. For each combination of β and discriminability, we could then calculate a probability of choosing the action with the highest Q-value based on the observed frequency of making these choices – something we could not do if we hadn't binned trials together – and determine the separate effects of these two factors as depicted in Figure 9. As expected, increasing discriminability increased the probability of exploiting the action with the highest Q-value. For a discriminability lower than 0.6, probability of exploitation was roughly 0.5, while it was over 0.6 for discriminabilities ranging between 0.6 and 0.8, and about 0.8 for the highest values of discriminability. In addition, the fluctuating value of β also had a noticeable effect on exploitation as in all cases, increasing β did indeed increase the probability of exploitation. These increases were statistically significant (Wilcoxon signed-rank tests with Bonferroni corrections) for intermediate and high values of discriminability as shown in Figure 9. In summary, in trials in which the inferred values of β according to the meta-learning model are high, there is an experimentally observable increase in exploitation while controlling for the effect of learning.

DISCUSSION

In this work, we compared the ability of different computational models to account for rats' progressive tuning of the exploration-exploitation trade-off 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' significantly performance 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 the exploration-exploitation trade-off was progressively learned and stabilized in adaptation to the task. Such behavioural tendencies cannot be captured by a standard reinforcement learning model. Instead, we found that a meta-learning model, which linearly tunes the inverse temperature parameter based on variations in the average reward rate, provided the best account of these long-term variations in rats' behaviour. We further confirmed these modelling results by model simulations and analyses. Importantly, we confirmed that the current trial value of the inverse temperature was predictive of the rats' tendency to deviate from the currently optimal lever, even when controlling for the level of discriminability between learned lever values. These results suggest that rats progressively tune their exploration-exploitation trade-off while learning the structure of new decision-making tasks.

We limited ourselves in this study to the hypothesis that meta-learning concerned the inverse temperature which regulates the exploration-exploitation trade-off. In making this choice we pursued a line of inquiry begun in (Humphries, 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 (Niv *et al.*, 2007; Hamid *et al.*, 2016), this leads to the idea that the reward rate

controls exploration through tonic dopamine levels. Another possibility which we did not present here is that it is the learning rate or the forgetting rate or a combination of these parameters that is being regulated over time. Such a complete analysis would require a huge number of optimisations, but we did in fact test different meta-learning models of the learning rate (data not shown). We found them to be very similar to the meta-learning models of the inverse temperature both in terms of the optimisation criteria AIC and BIC. 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 that there could be large overlap between the two signals. Taken together, these data point toward a larger class of meta-learning models in which uncertainty controls the learning rate and the reward rate the inverse temperature.

The idea that an increase in reward rate should cause an increase in exploitation 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 and 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 and Robbins, 2005; Redish, Jensen and Johnson, 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, Niv and Dayan, 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 a first phase in which an action reliably produces a reward, the accumulation of rewards causes an increase in inverse temperature alongside the increase of the value of that action. If the link between action and reward is altered, the now very strong tendency to 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 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 could explain a puzzling lack of effect of 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 and win-shift are different in high- and low-risk blocks, but simulations of a model with a single inverse temperature is entirely capable of producing this type of behaviour (data not shown) so that differences in the Q-values in the two types of blocks is a 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 advisable unless counterfactual checks are carried out, as we did in (Cinotti *et al.*, 2019). Maybe if the blocks were longer than 24 trials, then variations in exploration 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.

Overall, this work constitutes one of the rare attempts to account for rats' progressive adjustment of their exploration strategy while they are learning the structure of a new task. Because our method allowed to relate the random exploration rate in the model to the rats' probability to deviate from the currently optimal lever, independently from the discriminability between learned lever values, it provides a concrete means to identify exploration patterns in behaving rats. This contributes to a promising line of research which could help better understand why animals behave according to a precisely tuned exploration-exploitation trade-off in the post-training phases of decision-making tasks.

ACKNOWLEDGEMENTS

François Cinotti would like to express his gratitude to the organisers of a three-day writing retreat for university of Reading staff which allowed him valuable time to start writing this paper, as well as to his post-doctoral supervisors Professors Ingo Bojak and Jon Gibbins for letting him finish work on this project. 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.

REFERENCES

Behrens, T.E.J. *et al.* (2007) 'Learning the value of information in an uncertain world', *Nature Neuroscience*, 10(9), pp. 1214–1221. Available at: https://doi.org/10.1038/nn1954.

Blackwell, K.T. and 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), p. e1011385. Available at: https://doi.org/10.1371/journal.pcbi.1011385.

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

Cinotti, F. *et al.* (2019) 'Dopamine blockade impairs the exploration-exploitation trade-off in rats', *Scientific Reports*, 9(1). Available at: https://doi.org/10.1038/s41598-019-43245-z.

Cohen, J.D., McClure, S.M. and 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), pp. 933–942. Available at: https://doi.org/10.1098/rstb.2007.2098.

Daw, N.D. *et al.* (2006) 'Cortical substrates for exploratory decisions in humans.', *Nature*, 441(7095), pp. 876–9. Available at: https://doi.org/10.1038/nature04766.

Daw, N.D. (2011) 'Trial-by-trial data analysis using computational models: (Tutorial Review)', in M.R. Delgado, E.A. Phelps, and T.W. Robbins (eds) *Decision Making, Affect, and Learning: Attention and Performance XXIII*. Oxford University Press, p. 0. Available at: https://doi.org/10.1093/acprof:oso/9780199600434.003.0001.

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

Eisenegger, C. *et al.* (2014) 'Role of Dopamine D2 Receptors in Human Reinforcement Learning', *Neuropsychopharmacology*, 39(10), pp. 2366–2375. Available at: https://doi.org/10.1038/npp.2014.84.

Everitt, B.J. and Robbins, T.W. (2005) 'Neural systems of reinforcement for drug addiction: from actions to habits to compulsion', *Nature Neuroscience*, 8(11), pp. 1481–1489. Available at: https://doi.org/10.1038/nn1579.

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

Gilbertson, T. and Steele, D. (2021) 'Tonic dopamine, uncertainty and basal ganglia action selection', *Neuroscience*, 466, pp. 109–124. Available at: https://doi.org/10.1016/j.neuroscience.2021.05.010.

Hamid, A.A. *et al.* (2016) 'Mesolimbic dopamine signals the value of work', *Nature Neuroscience*, 19(1), pp. 117–126. Available at: https://doi.org/10.1038/nn.4173.

Hart, A.S. *et al.* (2014) 'Phasic Dopamine Release in the Rat Nucleus Accumbens Symmetrically Encodes a Reward Prediction Error Term', *The Journal of Neuroscience*, 34(3), pp. 698–704. Available at: https://doi.org/10.1523/JNEUROSCI.2489-13.2014.

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

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

Iigaya, K. *et al.* (2018) 'An effect of serotonergic stimulation on learning rates for rewards apparent after long intertrial intervals', *Nature Communications*, 9(1), p. 2477. Available at: https://doi.org/10.1038/s41467-018-04840-2.

Jepma, M. *et al.* (2016) 'Catecholaminergic Regulation of Learning Rate in a Dynamic Environment', *PLOS Computational Biology*. Edited by J.X. O'Reilly, 12(10), p. e1005171. Available at: 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, pp. 59–69. Available at: https://doi.org/10.1016/j.jmp.2015.03.006.

Khamassi, M. *et al.* (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.

Lebarbier, E. and Mary-huard, T. (2006) 'Une introduction au critère BIC : fondements théoriques et interprétation', *Journal de la société française de statistique*, 147(1), pp. 39–57.

Lloyd, A. *et al.* (2023) 'Understanding patch foraging strategies across development', *Trends in Cognitive Sciences*, p. S1364661323001729. Available at: https://doi.org/10.1016/j.tics.2023.07.004.

Moin Afshar, N. *et al.* (2020) 'Reinforcement Learning during Adolescence in Rats', *The Journal of Neuroscience*, 40(30), pp. 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), pp. 357–376. Available at: https://doi.org/10.1196/annals.1390.018.

Niv, Y. *et al.* (2007) 'Tonic dopamine: opportunity costs and the control of response vigor', *Psychopharmacology*, 191(3), pp. 507–520. Available at: https://doi.org/10.1007/s00213-006-0502-4.

Palminteri, S., Wyart, V. and Koechlin, E. (2017) 'The Importance of Falsification in Computational Cognitive Modeling', *Trends in Cognitive Sciences*, 21(6), pp. 425–433. Available at: https://doi.org/10.1016/J.TICS.2017.03.011.

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

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

Schultz, W., Dayan, P. and Montague, P.R. (1997) 'A Neural Substrate of Prediction and Reward', *Science*, 275(5306), pp. 1593–1599.

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

Sutton, R.S. and Barto, A.G. (1998) *Reinforcement learning : an introduction*. MIT Press.

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

TABLES

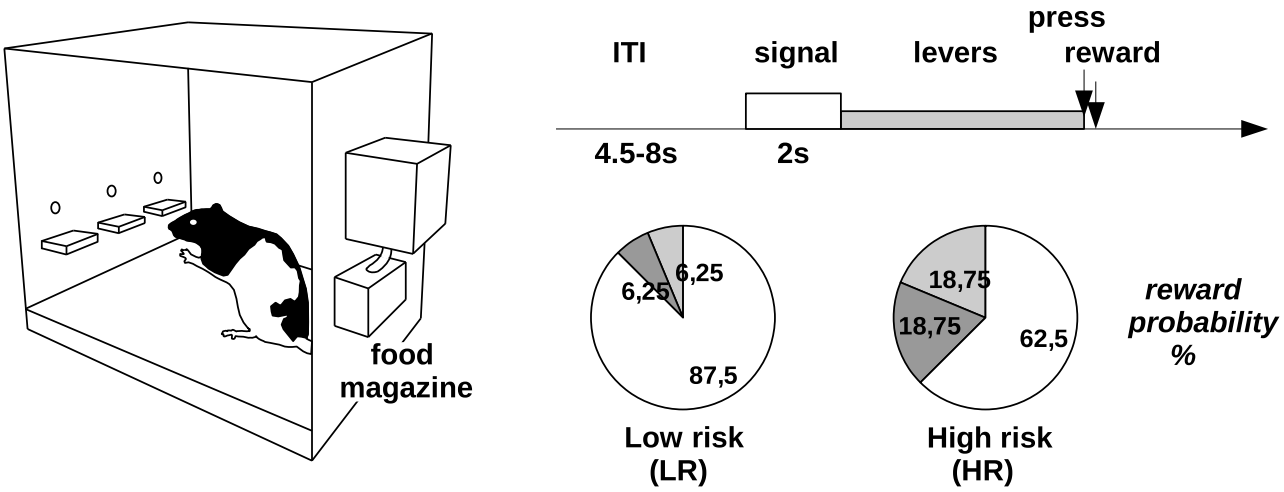

Figure 1: Outline of the experimental task, reproduced from Cinotti et al. 2019.

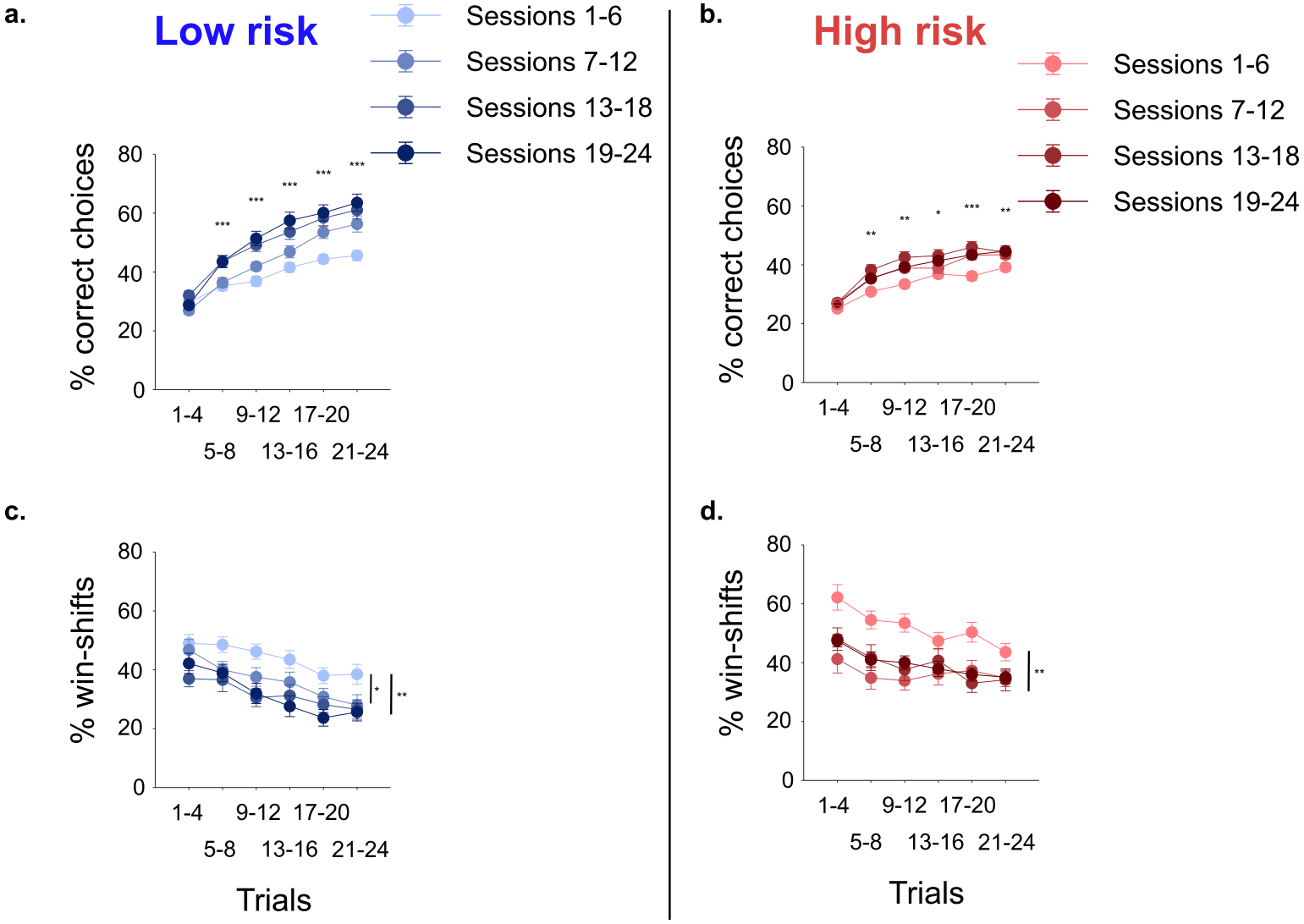

Figure 2: Between session changes in behaviour. **a.** & **b.** Mean performance +/- s.e.m. (n=24 subjects) increases between sessions in low- and high-risk blocks respectively. Stars indicate that for a given trial performance is significantly different between at least two groups of sessions. **c.** & **d.** Mean win-shift +/- s.e.m. decreases between sessions in low- and high-risk blocks respectively. Because of the lack of a significant interaction trial x sessions x risk, win-shift is not compared trial by trial as with performance but over all low- or high-risk blocks of different sessions. Significance levels as follows: * : $p < 0.05$; ** : $p < 0.01$, *** : $p < 0.001$.

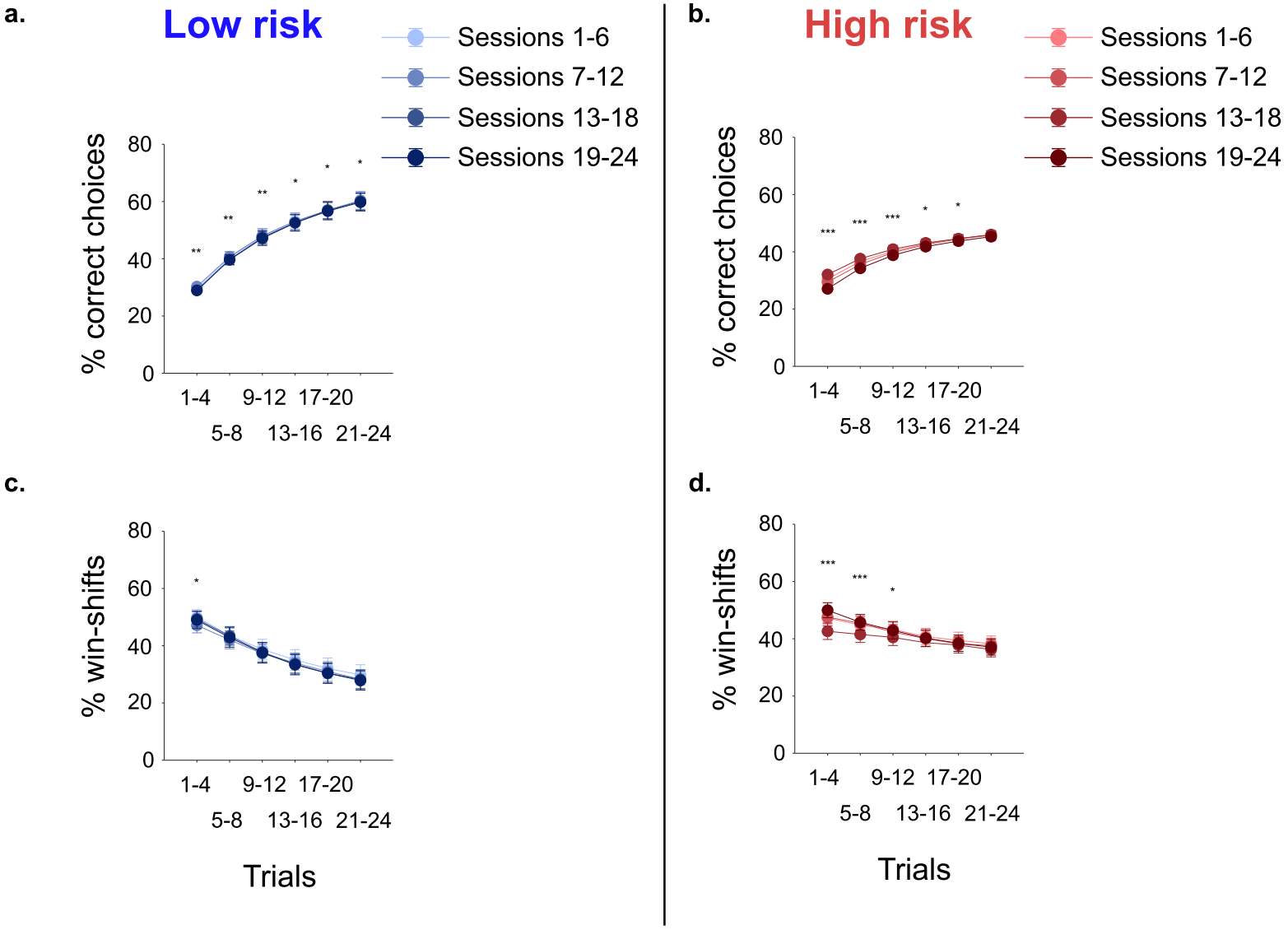

Figure 3: Simulations of the forgetting Q-learning model. **a.** & **b.** Mean performance +/- s.e.m. (n = 24) in low- and high- risk blocks respectively. **c.** & **d.** Mean win-shift +/- s.e.m. in low- and high-risk blocks respectively. Stars indicate that for a given trial, performance (or win-shift) is significantly different between at least two groups of sessions. Significance levels are as follows: * : $p < 0.05$; ** : $p < 0.01$; *** : $p < 0.001$.

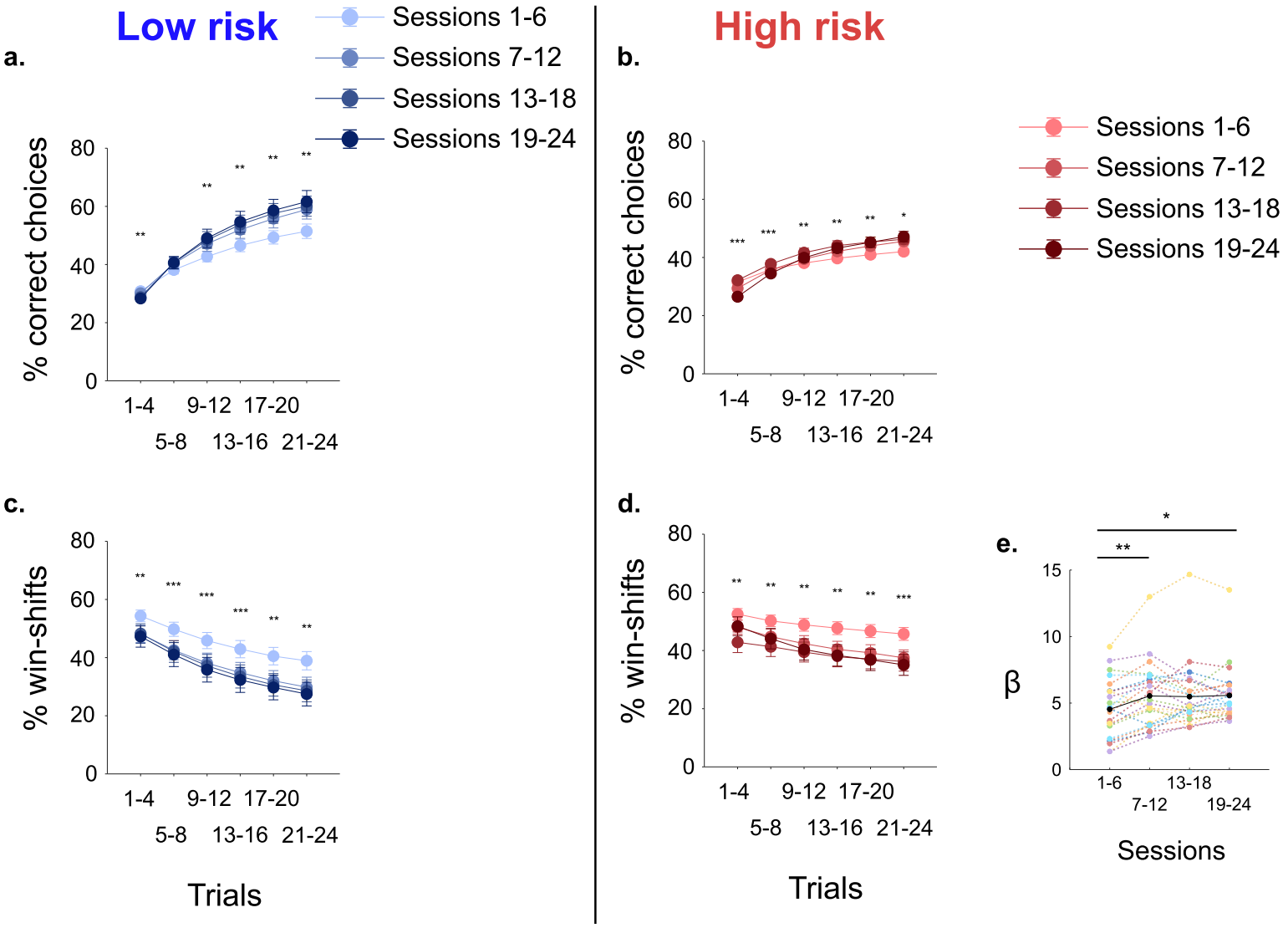

Figure 4: Simulations of the staggered model. **a.** & **b.** Mean performance +/- s.e.m. (n = 24) in low- and high- risk blocks respectively. **c.** & **d.** Mean win-shift +/- s.e.m. in low- and high-risk blocks respectively. Stars indicate that for a given trial, performance (or win-shift) is significantly different between at least two groups of sessions. **e.** Corresponding variations of the optimised values of the inverse temperature between sessions. Coloured lines represent variations for single individuals, the black line represents the mean. Significance levels are as follows: * : $p < 0.05$; ** : $p < 0.01$; *** : $p < 0.001$.

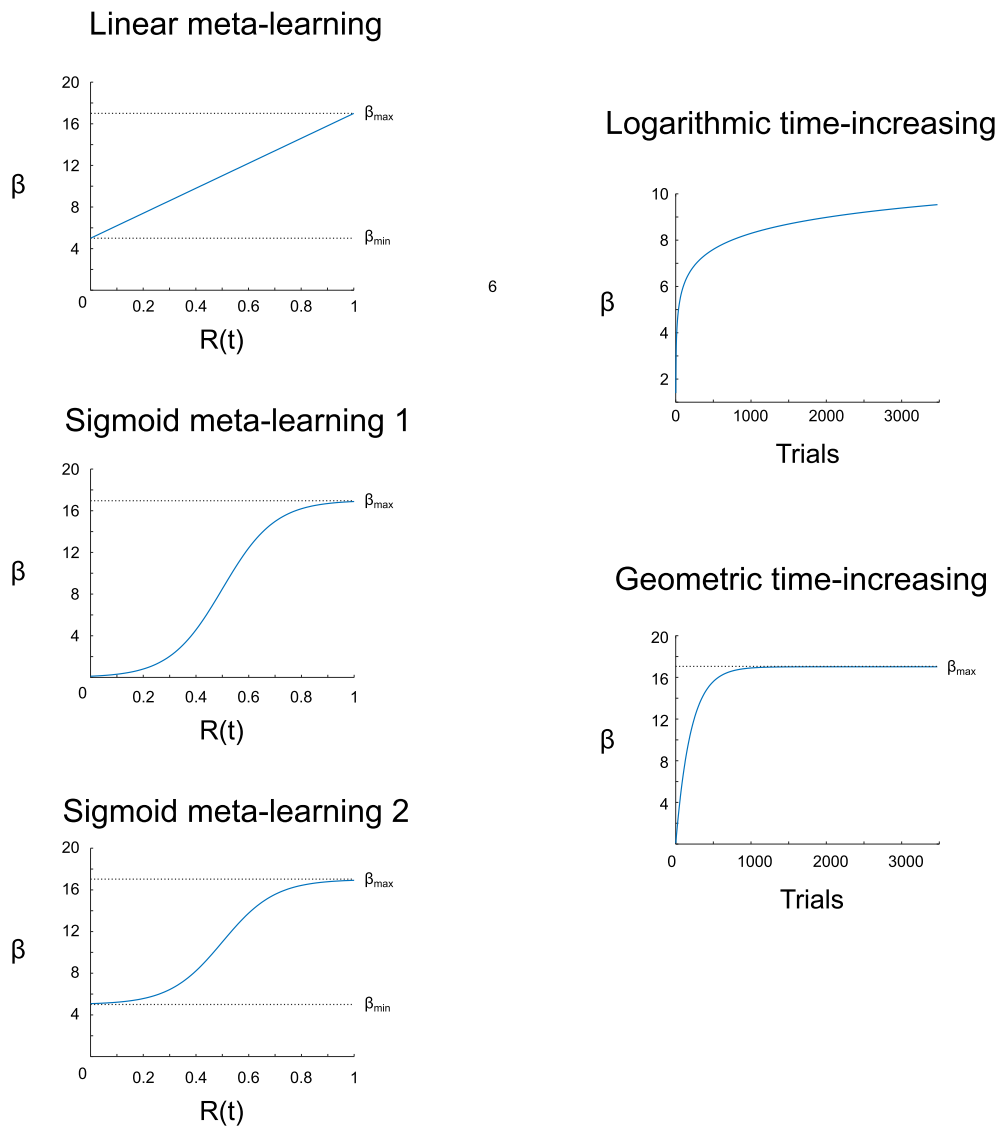

Figure 5: Regulation mechanisms of the inverse temperature in the different models.

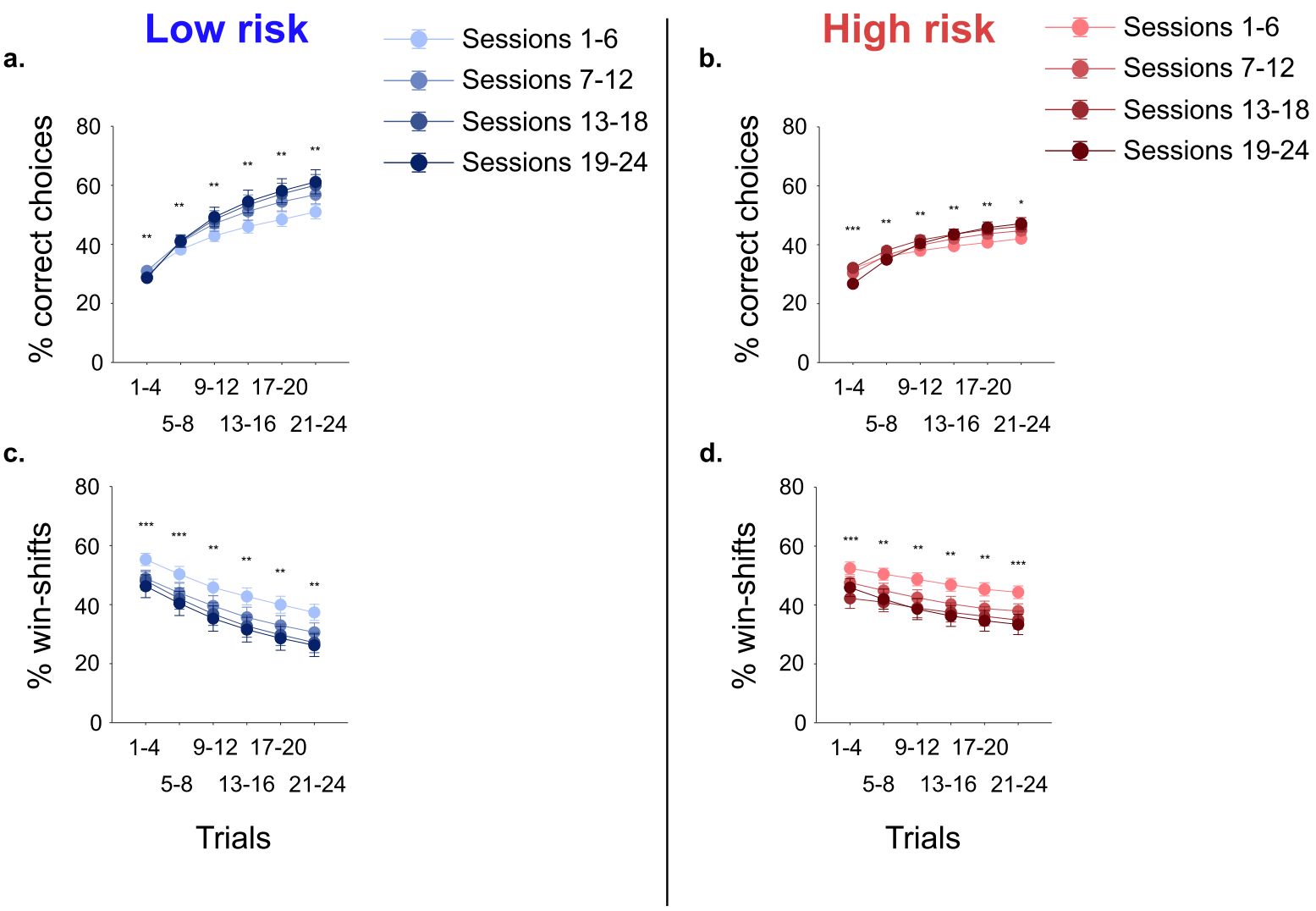

Figure 6: Simulations of the linear meta-learning model. **a.** & **b.** Mean performance +/- s.e.m. (n = 24) in low- and high- risk blocks respectively. **c.** & **d.** Mean win-shift +/- s.e.m. in low- and high-risk blocks respectively. Stars indicate that for a given trial, performance (or win-shift) is significantly different between at least two groups of sessions. Significance levels are as follows: * : $p < 0.05$; ** : $p < 0.01$; *** : $p < 0.001$.

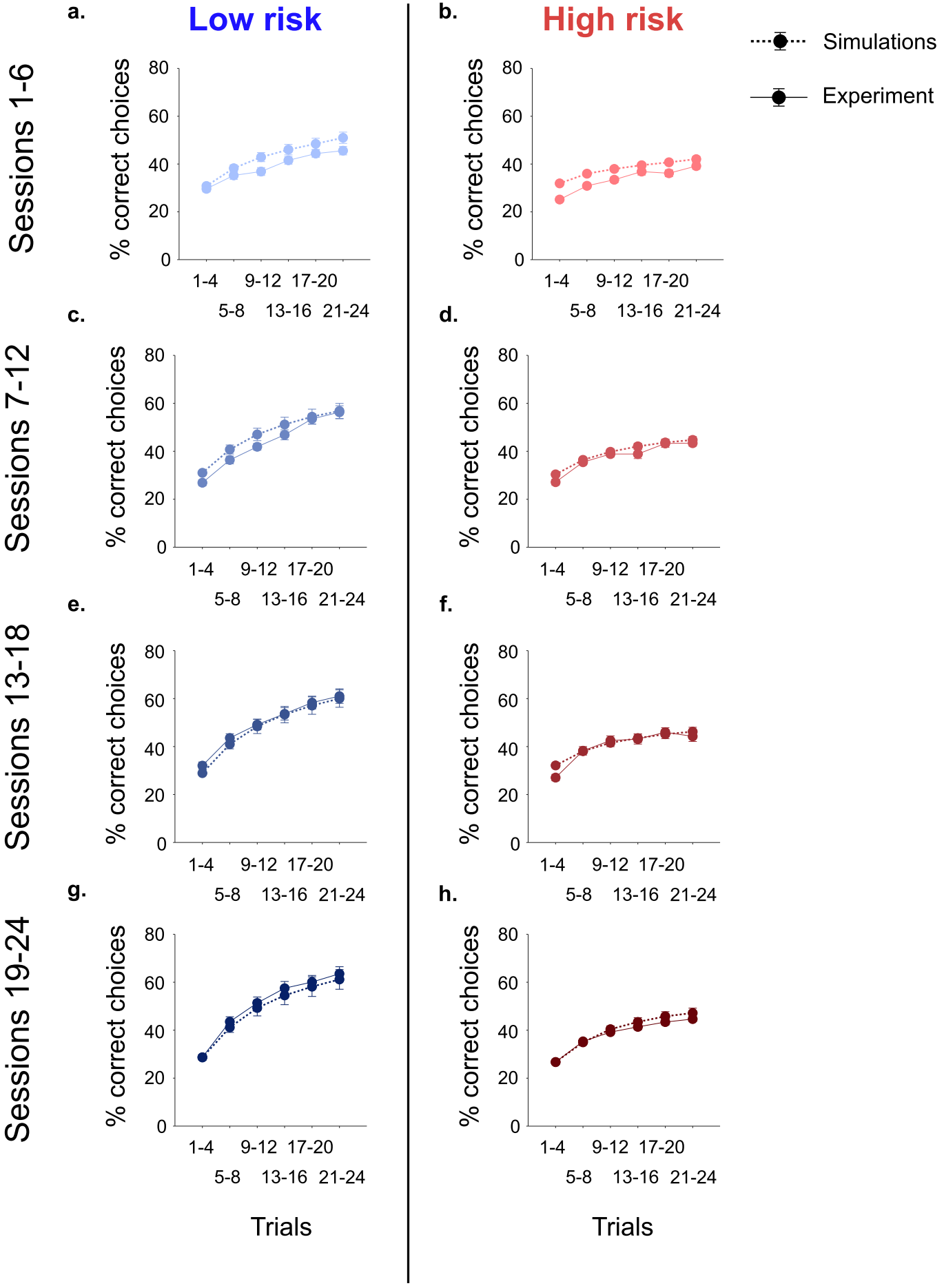

Figure 7: Fit of performance of simulations of the linear meta-learning model to the original experimental data. **a.**, **c.**, **e.**, & **g.** Performance in low-risk blocks. **b.**, **d.**, **f.**, & **h.** Performance in high-risk blocks.

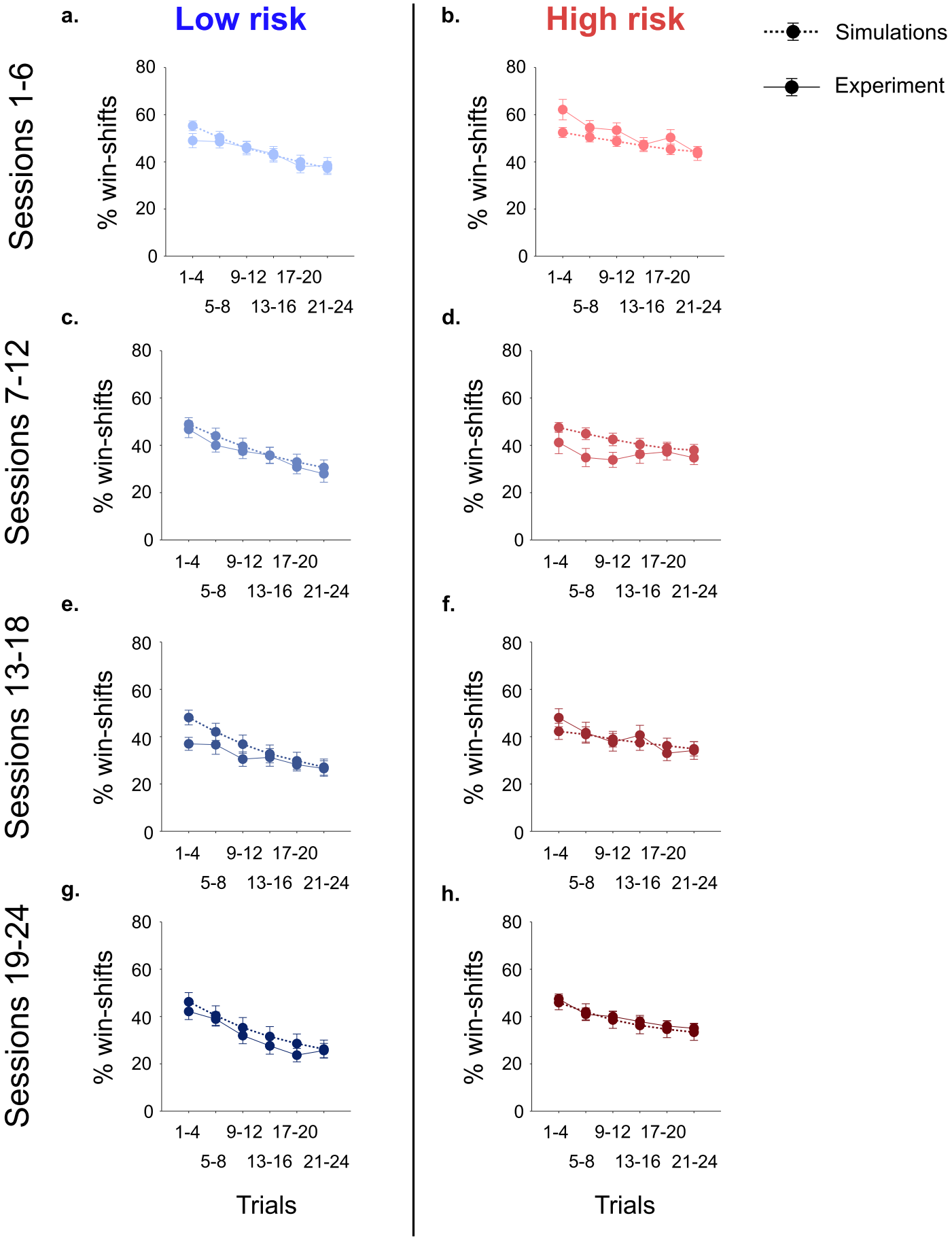

Figure 8: Fit of win-shift of simulations of the linear meta-learning model to the original experimental data. **a.**, **c.**, **e.**, & **g.** Win-shift in low-risk blocks. **b.**, **d.**, **f.**, & **h.** Win-shift in high-risk blocks.

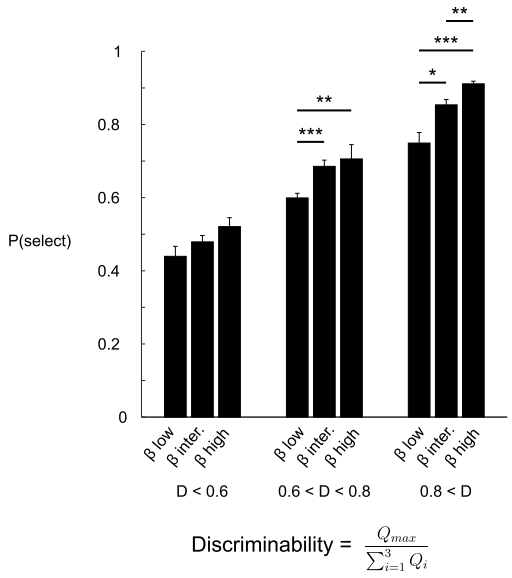

Figure 9: Effect of the inverse temperature on selection of the best action controlling for discriminability in constrained simulations of the linear meta-learning model. Significance levels (using a Wilcoxon signed-rank test with Bonferroni corrections) are as follows: * : p < 0.05, ** : p < 0.01, *** : p < 0.001.