Adaptive coordination of working-memory and reinforcement learning in non-human primates performing a trial-and-error problem solving task

Guillaume Viejo, Benoît Girard, Emmanuel Procyk, Mehdi Khamassi

▶ **To cite this version:**

HAL Id: hal-01624253
https://hal.science/hal-01624253

Submitted on 26 Oct 2017

Adaptive coordination of working-memory and reinforcement learning in non-human primates performing a trial-and-error problem solving task

Guillaume Viejo[a,b,], Benoît Girard[a,], Emmanuel Procyk[c], Mehdi Khamassi[a]

[a]*Sorbonne Universités, UPMC Univ Paris 06, CNRS, Institute of Intelligent Systems and Robotics (ISIR), F-75005 Paris, France*
[b]*Montreal Neurological Institute and Hospital, 3801 University Street, Montreal, Quebec, Canada,*
[c]*University of Lyon, Université Claude Bernard Lyon 1, INSERM, Stem Cell and Brain Research Institute U1208, Lyon, France*

Abstract

Accumulating evidence suggest that human behavior in trial-and-error learning tasks based on decisions between discrete actions may involve a combination of reinforcement learning (RL) and working-memory (WM). While the understanding of brain activity at stake in this type of tasks often involve the comparison with non-human primate neurophysiological results, it is not clear whether monkeys use similar combined RL and WM processes to solve these tasks. Here we analyzed the behavior of five monkeys with computational models combining RL and WM. Our model-based analysis approach enables to not only fit trial-by-trial choices but also transient slowdowns in reaction times, indicative of WM use. We found that the behavior of the five monkeys was better explained in terms of a combination of RL and WM despite inter-individual differences. The same coordination dynamics we used in a previous study in humans best explained the behavior of some monkeys while the behavior of others showed the opposite pattern, revealing a possible different dynamics of WM process. We further analyzed different variants of the tested models to open a discussion on how the long pretraining in these tasks may have favored particular coordination dynamics between RL and WM. This points towards either inter-species differences or protocol differences which could be further tested in humans.

---

1. Introduction

The use of computational models relying on the Reinforcement Learning (RL) theory [1] in decision-making tasks greatly contributed to a better understanding of dopamine reward signals in the brain [2], neural activities in other brain areas such as prefrontal cortex and basal ganglia [3, 4, 5, 6], as well as alterations of brain activity and behavior in different pathologies [7, 8]. The understanding of the brain mechanisms at stake has been facilitated by the replication of central results in rodents, humans and non-human primates, such as dopamine-related reward prediction error signals [9, 10, 11], action value encoding in the striatum [12, 13, 8], forgetting mechanisms of action values [14, 13, 15, 16], or even neural correlates of parallel model-based and model-free learning processes [17, 18, 19]. This enabled a transfer of knowledge between species and a more global understanding of possible neural architectures for the coordination of learning and decision-making processes.

However, it is not clear whether human and non-human primates always use similar learning and decision-making strategies in these types of tasks. In particular, Collins and Frank [20] recently showed that while most computational studies attempt to explain all aspects of behavior in terms of RL, human behavior in these types of tasks involve a combination of model-free reinforcement learning (MFRL) and working-memory (WM) processes. The evolution of human subjects' choices in their task was better explained as a weighted contribution of MFRL and WM in the decision process. Does monkey behavior show similar properties?

To answer this question, we propose to use the same model-based analysis approach that we recently employed in a human instrumental learning task [21]. In this study, we treated the WM component as a deliberative model-based system and a q-learning algorithm [22] as the MFRL system and tested different processes to dynamically determine the contribution of each system in the decision of each trial of a task. We moreover proposed a novel method to compare the ability of different models to not only fit subjects' choices but also the trial-by-trial evolution of their reaction times, hence providing a finer description of behavior.

We previously found with this computational method that humans adaptively combined MFRL and WM, spending more or less time searching in working memory depending on the uncertainty of the different trials [21]. Here we tested the same models, plus new variations of these models, on monkey behavioral data in a deterministic four forced-choice problem-solving task [23, 24, 15]. As for humans, we found that the behavior of the five monkeys was better explained in terms of a combination of MFRL and WM, rather than by one of these decision systems alone. We nevertheless found strong inter-individual differences, some being better captured by the same coordination dynamics than for humans, others showing the opposite pattern. We further analyze different variants of the tested models to open a discussion on how the long pretraining in these tasks may have favored particular coordination dynamics between reinforcement learning and working memory. This points towards either inter-species differences or protocol differences which could be further tested in humans.

2. Material and methods

2.1. Problem solving task

Five monkeys were trained to discover by trial and error the correct target out of four possible targets as shown in figure 1 [24, 15]. A typical problem starts with a search phase during which the animal does incorrect trials (INC) until he performs the first correct trial (CO1). Then, a repetition phase starts for a various number of trials (from 3 to 11) during which the animal should repeat the correct action. This varying number of repetitions prevents the animal from anticipating the end of the problem. In the following analysis, only the first three repetition trials were compared with the models since they constitute the maximal common denominator between all achieved problems. At the end of the repetition phase, a signal indicates the beginning of a new problem. In 90% of new problems, the rewarding target is different from the previous one. The successive events of a trial and a problem are represented in figure 1.

2.2. Theoretical models

Our hypothesis is that monkeys solve this task through a combination of working-memory (WM) and model-free reinforcement learning (MFRL), as humans appear to do in similar tasks [20, 21]. To test this hypothesis, we compared two different models representing different ways of coordinating
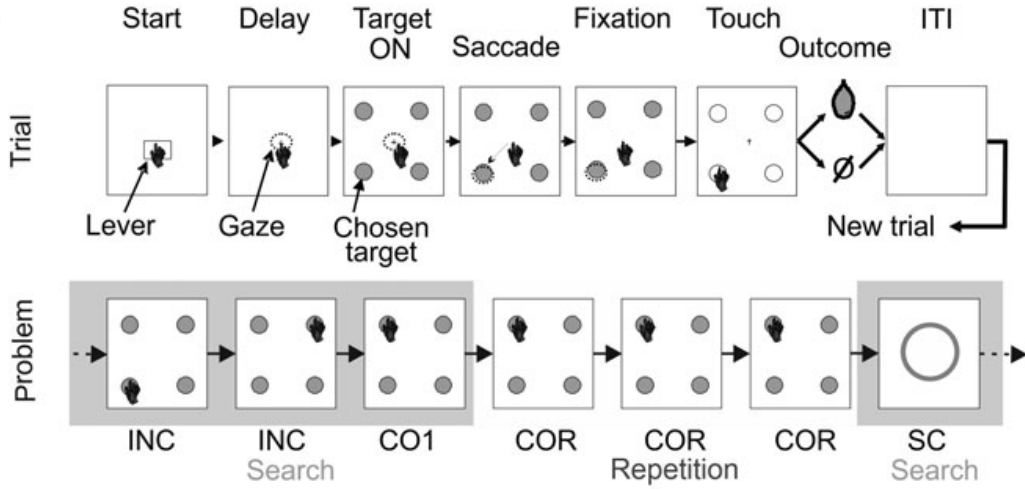
3

Figure 1: Trial-and-error Problem Solving Task. (Top) Successive events within a trial: the animal has to press a central lever and fixate his gaze on it until the trial starts; then targets appear, a go signal allows the animal to saccade towards its chosen target, then to touch the chosen target on the touch screen; finally, a reward is given or not, depending on the choice, and an inter-trial interval (ITI) is imposed. (Bottom) Successive trials within a problem. The animal performs a series of incorrect (INC) trials until finding the correct target and getting rewarded (first correct trial, CO1). Then the animal enters the repetition phase where it has to make at least three correct trials by repeating the choice of the same target. Finally, a signal-to-chance (SC) is presented indicating that the correct target is likely ($P = 0.9$) to change location.

WM and MFRL [20, 21], as well as models using either WM alone or MFRL alone to verify that they are not sufficient to explain the experimental data. We thus tested four computational models. Moreover, we tried different variations of these models to assess which particular computational mechanisms appear critical to explain the data.

With the exception of the q-learning algorithm [22], the models were first described in Viejo et al. [21] without the new variations presented here. The task is modeled as one state (for each trial, we only model the decision moment where the animal chooses between the four targets, and gets a feedback for this choice), four actions (one per target) and two possible values of rewards (1 for a correct trial, 0 for an error). The four models and their relations are described in figure 2.
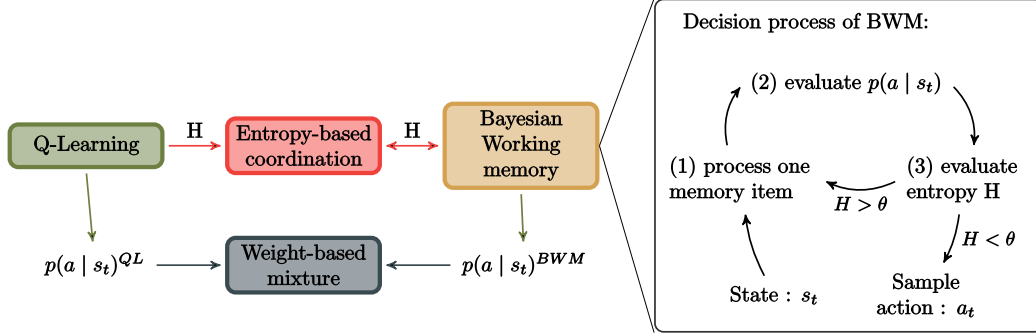
Figure 2: For all trials, the agent chooses an action with only one model (out of four). The deliberative behavior is represented by the Bayesian working memory model and the habitual behavior is represented by the q-learning. Between the two, the different models for interaction: Entropy-based coordination [21] and Weight-based mixture [21, 20]. The right panel shows the decision process of the working memory. In the first step, the probability distribution $p(a, r | t_{0 \to i})$ is computed. The second step evaluates action probabilities and the third step compute the entropy H from the action probabilities. The goal of the cycle is to reduce iteratively the entropy H.

2.2.1. Q-Learning (MFRL)

The q-learning model is a standard model-free algorithm in the reinforcement learning field [22]. It stores a table of q-values $Q(s, a)$ from which an action can quickly be drawn. After a transition to a new state s', the q-values are updated according to :

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma max'_a Q(s', a') - Q(s, a)] \tag{1}$$

The sampling of an action is made through a soft-max equation :

$$p(a|s) = \frac{e^{\beta Q(s,a)}}{e^{\sum_{b \in A} \beta Q(s,b)}} \tag{2}$$

At the initialization, the q-values are set to 0.0.

2.2.2. Bayesian Working memory (BWM)

The working memory model that we used chronologically stores the description (or memory item) of the events that occurred during past trials (i.e. chosen target, outcome). The number of memory items that can be stored is limited by a parameter N optimized for each monkey. Each time a new item enters memory, the item older than N trials is removed and memory decay is modeled by the convolution of the other items with a uniform distribution.

An element in memory contains the probability $p(a|t)$ of having performed an action in a certain past trial t and the probability $p(r|a,t)$ of having observed a certain reward r given an action a in a past trial t. During the decision process, those probability mass functions are first combined with Bayes rule, then summed (see Viejo et al. [21] for a full description of the equations). The sum of i memory items gives $p(a,r|t_{0 \to i})$ which is then reduced to $p(a|r,t_{0 \to i})$. The index $0 \to i$ indicates the number of memory items processed sequentially from the most recent one with an index of 0 to an oldest one with an index i. In this task, there are only two possible outcomes $r \in 0,1$ and only one action is rewarding. At the beginning of a problem, when only incorrect trials have yet been experienced, only non-rewarded actions have been stored and untried actions should thus be favored. On the contrary, the probability of choosing the only action associated with the reward should be maximal if already observed. This reasoning has been summarized in the following equation :

$$Q(a) = \frac{p(a|r=1, t_{0 \to i})}{p(a|r=0, t_{0 \to i})} \tag{3}$$

A simple normalization process allows the calculation of $p(a)^{BWM}$.

While previous working-memory models [20] process all items in memory each time memory is screened, here the main novelty consists in examining memory items one by one until the model is confident enough about which action to perform. This is modeled as a dynamical allocation of the number $t_{0 \to i}$ of memory items retrieved. The decision of evaluating the next memory item is dependent of the Shannon information entropy computed over the probability of action :

$$H = -\sum_a \left(p(a|t_{0 \to i}) \times log_2 p(a|t_{0 \to i}) \right) \tag{4}$$

If H is above a parameterizable threshold θ, the agent considers that it does not have enough information to decide and should thus evaluate the next element. If H is below θ, the model considers that enough information has been incorporated into the probability of actions in order to make a decision. If all memory items within the list have been screened, the model is forced to make a decision on which action to perform.

The number of memory items that can be processed for a given trial depends heavily on the history of past trials. If the correct action has been

made at the previous trial, the number of memory items retrieved will be very likely one. If the correct action has not yet been found, the number of memory items retrieved will most likely be large. This feature of the Bayesian working memory model ended up being the crucial aspect by which human subjects' reaction times (RT) could be explained in Viejo et al. [21]. Similarly, here the equation that relates the number of retrieved memory items with the simulated reaction times (sRT) is:

$$sRT(trial) = (log_2(i+1))^\sigma + H(p(a)) \tag{5}$$

The free parameter σ controls the proportion of the first part of the equation in sRT. This equation is used for all the other models (in the case of the q-learning, $log_2(i+1) = 0$, so that its reaction time only depends on the contrast between learned action values).

In [21] as well as in this study, we assumed that slower RT correspond to the use of working memory. The main justification comes from the literature studying the effects of working memory as part of cognitive control processes, which has emphasized that more cognitive control is reflected by an increase in RTs [25]. Some studies have further studied more specifically the link between RT variations and the balance between working memory and model-free reinforcement learning. For instance, in [26, 27], the authors studied a task with human subjects similar to the one presented in this draft. Subjects had to associate by trial-and-error each stimulus with one correct action out of 5. The authors showed that this deliberative process (i.e. remembering wrong actions to select untried actions in order to find the right one) was associated with slower reaction times. When modeling this task [21], we previously found that the working memory model presented in this draft, fitted best the behavior (choices and reaction times) during this deliberative process.

2.2.3. Weight-based mixture (MTB)

The weight-based mixture model [20] constitutes the first solution we tested for combining the Bayesian working memory model with the q-learning according to:

$$p(a) = (1 - w_t)p(a)^{QL} + w_t p(a)^{MTB} \tag{6}$$

The weight w_t evolves after each trial according to :

$$w_{t+1} = \frac{p(r_t|a_t)^{MTB}w_t}{p(r_t|a_t)^{MTB}w_t + p(r_t|a_t)^{QL}(1 - w_t)} \tag{7}$$

with $p(r_t|a_t)^{MTB}$ and $p(r_t|a_t)^{QL}$ being the relative likelihood that the corresponding model brought the reward. Thus, the weight evolves toward the most reliable strategy.

2.2.4. Entropy-based coordination

The entropy-based coordination, first proposed in [21], constitutes the second coordination solution that we tested. It explores the possibility of a closer interaction between the Bayesian working memory and the q-learning algorithm by conditioning the retrieval by the quantity of information contained in the working memory and the q-learning. The first point is to differentiate the entropies H^{QL} and H^{BWM} that can be computed for each strategy. The entropy H^{QL} is computed with a soft-max function. The second point is that H^{BWM} (being equal to $H^{max} = log_2(|Actions|)$ with $|Actions|$ the number of actions at the beginning of a trial) evolves inside a trial (hence reflecting a long inference process within the trial), while H^{QL} evolves between trials (hence reflecting a long learning process across trials). The two entropies are used to control the retrieval probability of the working memory with the following sigmoid equation:

$$p(retrieval|H^{BWM}, H^{QL}) = 1 - \frac{1}{1 + \lambda_1(n-i)exp^{-\lambda_2(2H^{max} - H_{0 \rightarrow i}^{BWM} - H^{QL})}} \quad (8)$$

with n the number of memory items present in the working memory list, i the number of memory items already retrieved and λ_1, λ_2 gain parameters. This model thus discards the hard decision threshold θ, used when the BWM operates in isolation. If the decision process is engaged, the q-values of each strategies are simply summed:

$$Q(a) = Q(a)_{0 \rightarrow i}^{BWM} + Q(a)^{QL} \quad (9)$$

2.2.5. Variations of the models

The second novelty of the present paper is that from the original version of the four previously described models, we tested a number of variations (see table 1). The aim is to examine which particular computational mechanisms of each model are critical to better capture the monkeys' behavior. The symbol $\varnothing$ indicates a model that is not concerned by the variation.

In a first variation, the discount factor $\gamma \in [0, 1[$ is optimized allowing the model to look in the future. In the original version of the q-learning, $\gamma = 0$ to account for the fact that transitions between states are randomized.

	BWM	**Q-L**	**Mixture**	**Coordination**
Variation 1	original model	original model	original model	original model
Variation 2	$\varnothing$	$\gamma \in [0, 1[$	$\gamma \in [0, 1[$	$\gamma \in [0, 1[$
Variation 3	$\varnothing$	$\gamma \in [0, 1[$ $\neg$ INIT(Q-L)	$\gamma \in [0, 1[$ $\neg$ INIT(Q-L)	$\gamma \in [0, 1[$ $\neg$ INIT(Q-L)
Variation 4	$\varnothing$	$\gamma \in [0, 1[$ $\neg$ INIT(Q-L) DECAY(Q-L)	$\gamma \in [0, 1[$ $\neg$ INIT(Q-L) DECAY(Q-L)	$\gamma \in [0, 1[$ $\neg$ INIT(Q-L) DECAY(Q-L)
Variation 5	ANT(BWM)	$\varnothing$	$\gamma \in [0, 1[$ $\neg$ INIT(Q-L) DECAY(Q-L) ANT(BWM)	$\gamma \in [0, 1[$ $\neg$ INIT(Q-L) DECAY(Q-L) ANT(BWM)
Variation 6	$\varnothing$	$\varnothing$	$\varnothing$	$\gamma \in [0, 1[$ $\neg$ INIT(Q-L) DECAY(Q-L) META-L
Variation 7	$\varnothing$	$\varnothing$	$\varnothing$ $\neg$ INIT(Q-L) DECAY(Q-L) THR(δ)	$\gamma \in [0, 1[$ $\neg$ INIT(Q-L) DECAY(Q-L) THR(δ)

Table 1: Variations of the Bayesian Working Memory model (BWM), the q-learning model (Q-L), the weighted mixture model and the entropy-based coordination model. The symbol $\varnothing$ designs the models that are not concerned by the tested variation. The variations with $\gamma \in [0, 1[$ allows the discounting factor γ to be optimized (set to 0 in the original version). With $\neg$ INIT(Q-L), the q-learning is not reinitialized at the beginning of a new block. The DECAY(QL) function allows the q-values to be forgotten given an optimized parameter of decay κ. With ANT(BWM), the working memory model evaluates the probability of action and the associated entropy during the outcome intervals in order to anticipate (thus gaining time) the decision for the next trial. This heuristic of anticipation is made only for the search phase. META-L is the meta-learning of average entropy value for each type of trial in order to bias the sigmoid function of the coordination model (see equation 11). THR(δ) conditions the encoding of a past trial in the working memory based on the error prediction δ calculated from q-learning (see equation 12).

Thus, there was no interest for the agent to learn the structure of states transition. In this framework, the task loops onto one state allowing the use of a discount factor.

Khamassi et al. [15] previously fitted various versions of the q-learning algorithm to this task. The most successful version considered that the information contained in the q-values were transmitted between problems i.e. the last rewarding action can bias the choice in a new search phase. This indicates some learning of the task structure by the monkeys: the rewarded target in the previous problem is very unlikely to be again rewarded in the present one. Therefore, we tested a second variation ($\neg$ INIT(Q-L)) where the q-learning values were not reinitialized. Nevertheless, it is likely that this strategy of non-reinitialization is more efficient if the values of q-learning are progressively erased along the task, as found in [15]. Thus, we tested a third variation of q-learning (DECAY(Q-L)) with a decay factor to mimic progressive forgetting. The values are modified at each trial according to:

$$Q(a) \leftarrow Q(a) + (1 - \kappa)(Q_0 - Q(a)) \tag{10}$$

with $Q_0 = 0$ and $0 \leq \kappa \leq 1$ the decay parameter.

For some cases (monkeys m and p, see figure 5), the reaction times show a specific pattern of diminishing during the search phase and re-increasing during the repetition phase, in contrast to what we previously observed in humans in a similar task [21]. This observation suggests that working memory is mostly involved during the repetition phase for these monkeys, since the use of working memory increases the reaction times. Given that the monkeys are trained for thousands of trials, it is very likely that the decision process during the search phase has been automatized. The first reason to hypothesize an automatization of choices during the search phase is the high number of trials that the monkeys did, leading the animals to learn an efficient search strategy. Their performance in the search and repetition phases is a good indicator of this fact. We hypothesize that the search strategy, as a rule, is a cognitive structure that has been learned and automatized such that it can be applied efficiently whatever the order of outcomes received during the search process. We have shown in previous publications that although the search is highly efficient, there is no fixed order in choices [28, 23].The second reason is the dynamic of reaction times during the search phase : for the monkeys p, s and g, the reactions times are decreasing as the incorrect actions are revealed. Thus, we supposed that this dynamics was best translated by an

automatic process. Yet, we did not found a fix individual idiosyncratic order of actions during the search phase. However, monkeys must operate a deliberative process in order to avoid doing repetitive errors which have a high opportunity cost since making a correct response is very likely during repetition. Thus, a possibility that we have explored is the anticipation of the action by the working memory (ANT(BWM)). During the update of the models by the reward and for the search phase only, a simple heuristic lets the working memory retrieve all the memory items (including the most recent one) in order to prepare the probability distribution of actions for the next trial. Then, the entropy $H(p(a))$ of q-values (being either working memory alone or the combination with q-learning) will be lower at the onset of a trial without the cognitive load that would normally come with the retrieval of previous incorrect actions. Since the animal cannot anticipate the end of a problem (and has no interest in doing so), this heuristic is not applied to the repetition phase.

To account for the over-training of monkeys, we tested the long-range learning of meta-parameters that can bias a model. We incorporated this idea in a new version of the entropy-based coordination model (META-L) with the addition of average entropy variables $\hat{H}_{type}^{BWM}$ and $\hat{H}_{type}^{QL}$ for each type of trial. Types can be trials during the search phase or trials during the repetition phase. Thus, the model learns a table that maps for each trial type the corresponding average entropies of the q-learning and the Bayesian working memory learned during thousands of trials performing the same the task. To average the entropies, the model is first tested in normal condition in order to store the distinct entropies computed at each trial. Finally, the sigmoid equation 8 becomes:

$$p(retrieval) = 1 - \frac{1}{1 + \lambda_1(n-i)exp^{-\lambda_2(2H^{max} - H_{0 \rightarrow i}^{BWM} - H^{QL} + \hat{H}_{type}^{BWM} - \hat{H}_{type}^{QL})}} \quad (11)$$

The addition of $\hat{H}^{BWM}$ will force trials with average high uncertainty of the working memory to decide faster. If $\hat{H}^{BWM}$ is high, the exponent $2H^{max} - H_{0 \rightarrow i}^{BWM} - H^{QL} + \hat{H}_{type}^{BWM} - \hat{H}_{type}^{QL}$ will increase and gives a lower $p(retrieval)$. On the contrary, the average entropy $\hat{H}^{QL}$ modifies the sigmoid equation in the same direction as H^{QL}. A low average uncertainty on the q-learning favors fast decision of the model and conversely.

The last modification incorporates the update of the working memory depending on the value δ_t of the temporal difference of the q-learning. The

relation between working memory and temporal difference has already been explored in various models of reinforcement learning [29, 30] to account for observations of the physiological effects of dopamine on the circuitry of the prefrontal cortex [31, 32, 33, 34]. Thus, we tightened the relation between the Bayesian working memory model and the q-learning algorithm (THR(δ)) for both models of interaction. The action of adding a new element in the working memory list is conditioned by:

$$[BWM] \leftarrow [p(a_t), p(r_t|a_t)] \ if \ \delta_t < \xi_1 \ or \ \delta_t > \xi_2 \tag{12}$$

For recall, the temporal difference is computed according to:

$$\delta_t = r_t + \gamma max_a Q(a) - Q(a) \tag{13}$$

with r_t the reward. In a nutshell, a large prediction error (being it positive or negative) induces an encoding of the last trial by the working memory since the q-learning has not converged. On the contrary, a small prediction error indicates a converging q-learning which can avoid a costly working memory update.

2.2.6. Parameters optimization

As in [21, 35, 36], the parameters optimization was made using the SFERES toolbox [37] that implements the standard NSGA-2 evolutionary algorithm. Each variation of each model was optimized separately for each monkey. The scores of a set of parameters are the ability to maximize the likelihood on every step (correct and incorrect trials) that the model do the same actions as the monkey and minimize the mean-square error on representative steps between the average monkey reaction times and the simulated reaction times. The representative steps are defined as the trials inside the search phase and the following three repetitive trials separated according to the length of the search phase. Only the 0 to 4 errors blocks were considered in order to compute the representative steps. The SFERES toolbox outputs the set of parameters that maximizes both the fit to choice and the fit to reaction times under the form of a Pareto front (see figure 6). For each considered computational model, the Pareto front shows the solutions (i.e. the parameter sets) which are either not dominated by any other solution on at least one dimension, or not dominated by any other solution on at least one weighted combination of dimensions. For instance, a solution which gives the best fit on reaction times will be part of the Pareto front for the considered

model even if its fit to choices is not the best. Similarly, a solution which gives the best mean fit on choices and reaction times will be part of the Pareto front even if there exists other solutions which are better at fitting either choice only or reaction times only.

The methodology for selecting the optimal solution in a multi-dimensional optimization then relies on the use of an aggregation function. Such a function allows for the combination of numerical values $x_1, ..., x_m$ into one value to rank all possible solutions. We used the Chebyshev distance in order to aggregate the normalized fitness functions (i.e. fit to choices and fit to reaction time) in one single value [38]. The aggregation function for each x solution is defined as:

$$t(x) = \max_{i \in M} \lambda_i \frac{\alpha_i - x_i}{\alpha_i - \beta_i} + \epsilon \sum_{i=1}^{m} \lambda_i \frac{\alpha_i - x_i}{\alpha_i - \beta_i} \tag{14}$$

with $\lambda \in \mathbb{R}_+^M$ a weight vector in order to bias the ranking if one fitness function is more important than the other (in our case $\lambda = 0.5$ for both), $\alpha_i = sup_{x \in \mathbb{X}}$ the optimal point and $\beta_i = inf_{x \in \mathbb{X}} x_i$ the worst combination of scores (although called the Nadir).

3. Results

3.1. Simple models of the interaction between working memory and reinforcement learning

3.1.1. Fit to choices only

We first evaluated the ability of each original model to replicate only the choices made by the monkeys. In figure 3, we compared the output of the optimization process using the Bayesian Information Criterion (BIC). Models were penalized proportionally to the number of free parameters. Nevertheless, the more complex dual models (entropy-based coordination or weight-based mixture) show better fits compared to simple models (Bayesian working memory or q-learning). Thus, this first analysis confirms that dual models made of the interaction between working memory and reinforcement learning are better than simple models at capturing the monkeys' behavior in this task.

Palminteri et al. [39] convincingly argued that model comparison without model simulation is not sufficient. In order to assess the validity of the fit according to BIC, we thus simulated the best model for each monkey. The
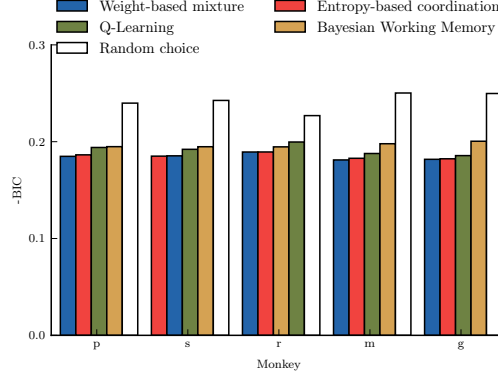
Figure 3: Bayesian Information Criterion (BIC) score for each monkey and each model and compared to a random decision model (white bar). Overall, the dual models (entropy-based coordination or weight-based mixture) have a lower BIC score despite the fact that they have more free parameters (3 for q-learning, 4 for bayesian working memory, 7 for weight-based mixture and 8 for entropy-based coordination).

set of best models is composed of the weight-based mixture models for 4 monkeys and the entropy-based coordination model for 1 monkey. During the simulation, only the list of problems (i.e. a list of indexes of the correct action) made by the monkey was used for transitioning between problems. The model was free to make its own choices and we repeated the experiment 1000 times. To display the performances, blocs (made of n trials of search phase plus 3 trials of repetition phase) were grouped for averaging according to the number of errors n-1 made during the search phase. For each group of blocs (defined by their number of errors n-1 during the search phase), the mean number of positive outcomes gained for each trial of the repetition phase (trials 1, 2 and 3 during which the animal should repeat the correct action) are then averaged to give the performances in repetition as shown in figure 4.

First, we found that the level of performance was roughly captured by most models. The performances of the monkeys were really high in repetition and all models reached this level of performance. Second, the striking observation is the inverse relation between the performance of the monkeys and the performance of the models. On average, monkeys have a decreasing performance between the first repetition and the third repetition trial. The models shows the opposite pattern with an increase of performance mostly due to the fact that dual models keep adding information about past trials
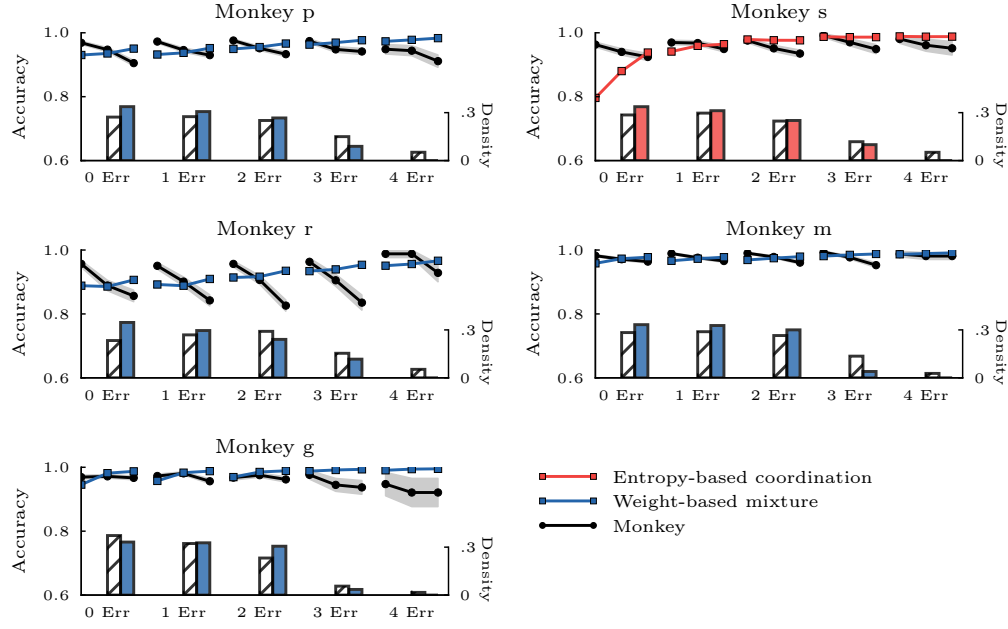
Figure 4: Mean performances (mean ± sem) during the first three repetition trials for each monkey (black circles) and each best fitted model according to the BIC criterion (squares). Performances in repetition phase are averaged over problems with the same number of errors (from 0 to 4 errors). Each model is simulated 1000 times with the same chain of problems (same correct actions) than the corresponding monkey. Along the accuracy, the density of problem types (number of errors) is represented with bars for the monkey (dashed bar) and the model (full bar).

in the working memory list. In figure 4, the bar density of each problem type (number of errors) is represented for each monkey and each best fitted model. We found that the problems count is slightly different between monkeys and models. When comparing the density of each type of problems, monkeys found the right action with less errors than the models (see density bars in figure 4).

These analyses suggest that an optimization of the model parameters in order to fit more aspects of monkeys' behavior, such as choices and reaction times, would be appropriate here.

3.1.2. Simulation of choices and reaction times

In figure 5, we examined the fit of both choices and reaction times. Once again, only dual models were selected by the method of multi-criteria optimization tested in Viejo et al. [21]. In addition to choices, we compared monkeys' reaction times with models' reaction times over representative steps. Reaction times are averaged over all problems of the same length (i.e. with the same number of errors). To the exception of monkey g, we found that the fit of reaction times was poorly performed by the original models (i.e. without the variations listed in table 1). Monkeys p and s showed progressively decreasing reaction times during the search phase, which were better fitted with the entropy-based coordination. Differently, monkey m showed constant reaction times during the search phase and a net increase during the repetition phase. Best fitted by the weight-based mixture, the model didn't manage to reproduce the global dynamics.

Overall, the original models of weight-based mixture and entropy-based coordination performed poorly in reproducing choices and reaction times in this trial-and-error search task with monkeys. In fact, the dynamics of the reaction times are much more diverse than the possibility that was given to those models for human behavioral data in [21]. In this original study, mean reaction times increased during the search phase and it was modeled by increments into the number of processed items within the working memory model in order to remember past incorrect actions (in order to avoid them). During the repetition phase, reaction times decreased and this was explained by the progressive shift to an habitual behavior (modeled using the q-learning) and thus a decrease in reaction times (since working memory was less used). In this task, the most common pattern is the opposite with decreasing reaction times during the search phase (monkeys p, s and g) and larger reaction times during repetition phase (monkeys m, p, s and g). To
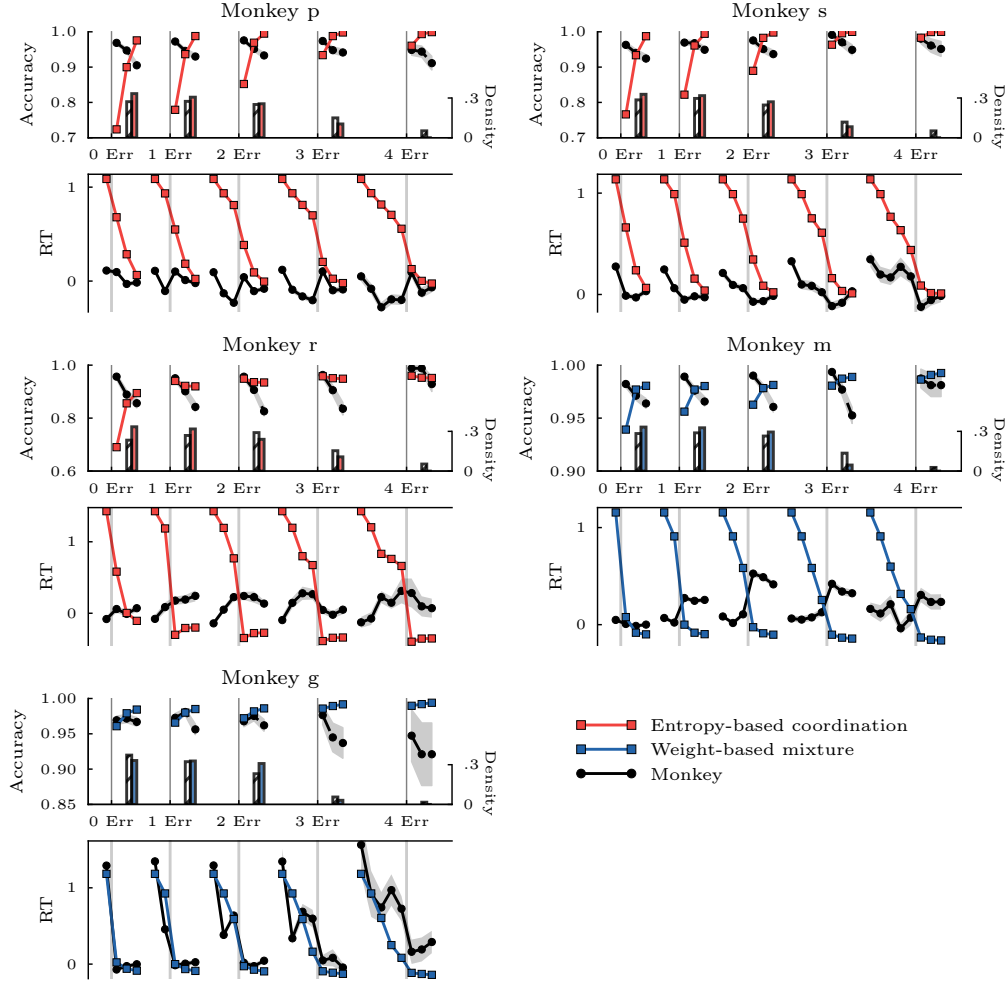
Figure 5: For each monkey, the upper panel shows the mean performances (mean ± sem) and the lower panel shows the centered reaction times for the same trials. Vertical gray lines indicates the transition from search phase to repetition phase. Dotted lines shows the monkey's behavior and squared lines shows the best fitted model behavior. Similar to figure 4, performances and reaction times are averaged over problems with the same number of errors (from 0 to 4 errors). Each model is simulated 1000 times with the same chain of problems (same correct actions) than the corresponding monkey. Along the performance, the density of problem types (number of errors) is represented with bars for the monkey (dashed bar) and the model (full bar).

conclude, none of the fitted models so far displayed the ability to have first a decrease then an increase in reaction times. It is thus this ability that we tried to capture by testing more complex variations of the models.

3.2. Towards more complex models of interaction between working memory and reinforcement learning

The results of the optimization process for the proposed variations of the initial model is shown in figure 6 for each monkey. The Pareto fronts (first column) show a domination of dual models once again confirming our first hypothesis that monkey behavior in this task can be better explained in terms of combination of working-memory and reinforcement learning processes rather than on each one alone. Moreover, for all monkeys we found that the Pareto fronts of weight-based mixture and entropy-based coordination were overlapping : no model took advantage over the other one. For the sake of clarity, the density of model variations inside each Pareto front only is represented (column 2). The last column shows the output of the Chebyshev function i.e. the aggregation of the fit to choice score and the fit to reaction times score for both the weight-based mixture and entropy-based coordination solutions. Since no dual models was definitively taking over, we decided to select and test the best solution according to the Chebyshev ranking (i.e. minimum value) for both models. For the entropy-based coordination, the best trade-off between fit to choice and fit to reaction times assigned the variation 5 ($\gamma \in [0, 1[, \neg INIT(Q-L), DECAY(Q-L), ANT(BWM)$)) to monkeys p, s and g and the variation 7 ($\gamma \in [0, 1[, \neg INIT(Q-L), DECAY(Q-L), THR(\delta)$)) to monkeys r and m. For the weight-based mixture, the best trade-off assigned the variation 2 ($\gamma \in [0, 1[$) to monkey g, the variation 5 to monkeys p and r and the variation 7 to monkeys m and s. As shown in figure 6, those variations are effectively over-represented in the Pareto fronts of each monkey. The overlaps between the two models suggest with higher confidence that the following computational mechanisms are important to explain monkeys' behavior: a positive discount factor ($\gamma \in [0, 1[$) for all monkeys; forgetting without reset of action values ($\neg INIT(Q - L), DECAY(Q - L)$) for all monkeys except monkey g for which only the entropy-based coordination was selected with this mechanism; anticipation of the next trial during the search phase ($ANT(BWM)$) for monkey p (but sometimes also selected for monkeys g, r and s); modulation of storage in working-memory based on the sign and magnitude of reward prediction errors ($THR(\delta)$) for monkey m (but sometimes also selected for monkeys r and s). This confirms our prior
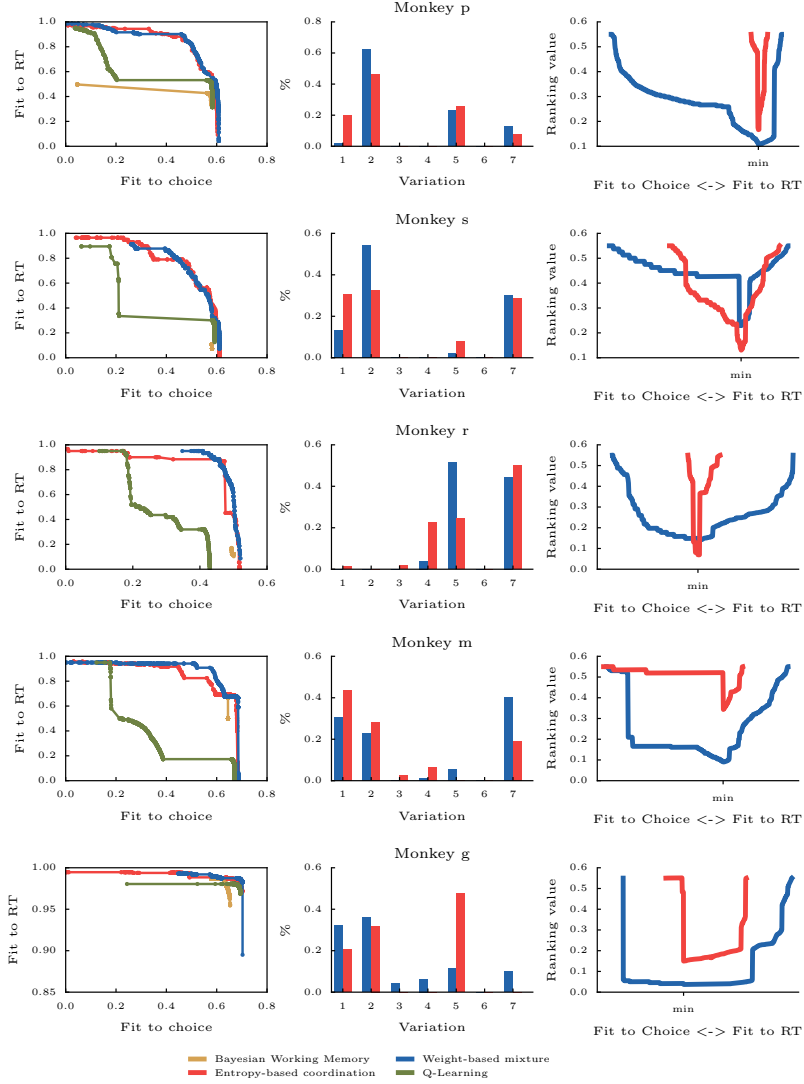
Figure 6: Results of the optimization process for each monkey. The first column shows the pareto front for the weight-based mixture, the entropy-based coordination, the bayesian working memory and the q-learning. The second column shows the count for each variations within the weight-based mixture and entropy-based coordination model of parametersets selected within the pareto front. The third column displays the output of the Chebyshev aggregation function that converts the two dimension points of the Pareto fronts into a single value allowing points to be ranked.

use of action value forgetting mechanism for the subset of data associated with neurophysiological recordings in monkeys m and p [15]. This nevertheless suggests a non-null discount factor in contrast to our prior work in both monkeys and humans in this type of tasks [15, 21], which will be further discussed later on.

3.2.1. Simulation of choices and reaction times

We tested each dual model's group of solutions as shown in figure 7 for entropy-based coordination and figure 8 for weight-based mixture. Overall, we found that the main caveat of the original models was corrected: reaction times could increase or decrease along the representative steps thus improving the fit to monkeys' reaction times.

As usual when working on a multi-dimensional problem, the improvements on one dimension can lead to a degradation of the fit in another dimension. An instance of this issue is shown for the fit to choice translated into the performance of the simulated model in repetition. For all dual models (figure 7 and 8) to the exception of the monkey g with weight-based mixture, the performances in repetition were lower. For the reaction times, we observed improvements with our innovations. Contrary to the original models, our new models displayed various reaction times dynamics such as a decrease during the search phase and an increase during the repetition phase, in contrast to the pattern of reaction times that we previously found in humans [21].

3.2.2. Contribution of models

We then examined the internal dynamics of each best model in figure 9 for entropy-based coordination and figure 10 for weight-based mixture. While the relative contribution of Bayesian working memory and q-learning in weight-based mixture is easily measurable through the weight w_t, the relative contribution is less identifiable in entropy-based coordination. Thus, we decided to plot only the most relevant variable for each variation in order to decipher the dynamics of the model.

For variation 5, the most relevant innovation is the anticipation of the next trial during the search phase. The whole content of working memory is recalled during the update phase after the agent receives a negative outcome during the search phase. This pre-retrieval modifies the probability of action for the Bayesian working memory (and by extension H^{BWM}) at the onset of a trial in the search phase which allows shorter reaction times. Thus,
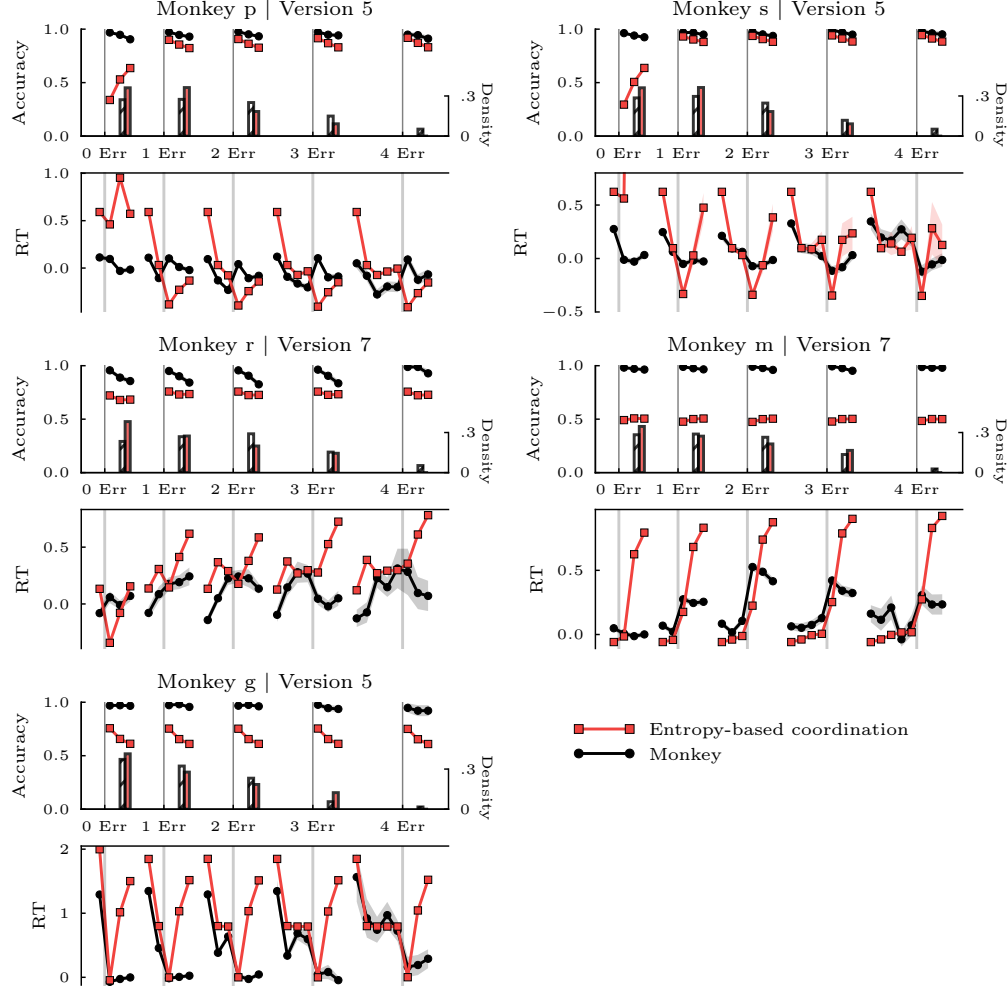
Figure 7: Best simulated behavior for entropy-based coordination models. For each monkey, the upper panel shows the mean performances (mean ± sem) and the lower panel shows the centered reaction times (mean ± sem) for the same trials. Vertical gray lines indicates the transition from search phase to repetition phase. Dotted lines shows the monkey's behavior and squared lines shows the best fitted entropy-based coordination behavior. Versions were selected with a trade-off amongst solutions composing the Pareto front of this model as shown in figure 6. The version for each model is displayed next to the monkey name. Each model is simulated 1000 times with the same chain of problems (same correct actions) than the corresponding monkey. Along the performance, the density of problem types (number of errors) is represented with bars for the monkey (dashed bar) and the model (full bar).
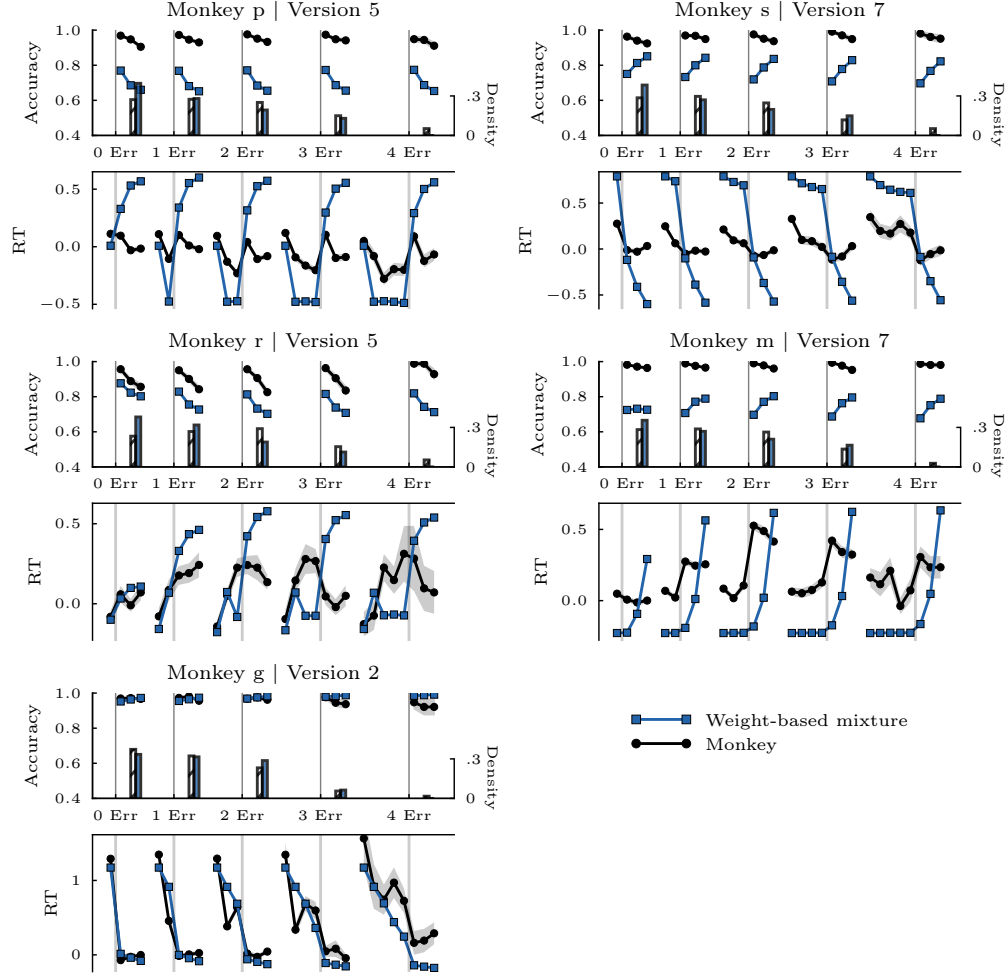
Figure 8: Best simulated behavior for weight-based mixture models. For each monkey, the upper panel shows the mean performances (mean ± sem) and the lower panel shows the centered reaction times (mean ± sem) for the same trials. Vertical gray lines indicates the transition from search phase to repetition phase. Dotted lines shows the monkey's behavior and squared lines shows the best fitted weight-based mixture behavior. Versions were selected with a trade-off amongst solutions composing the Pareto front of this model as shown in figure 6. The version for each model is displayed next to the monkey name. Each model is simulated 1000 times with the same chain of problems (same correct actions) than the corresponding monkey. Along the performance, the density of problem types (number of errors) is represented with bars for the monkey (dashed bar) and the model (full bar).

we looked at the number of retrieved memory items (star lines in figures 9 and 10). For weight-based mixture, the relation was binary (monkeys r and p, figure 10) : no memory items were retrieved right before the decision in the search phase and exactly one memory item (representing the last trial) was retrieved during the repetition phase. Besides, the mean weight w_t for each representative step (normal lines in figure 10) indicated the domination of the Bayesian working memory probability of action for the final decision. For the entropy-based coordination with the same variation, we found an intermediate level between no retrieval and constant retrieval. For monkeys p and s best explained by this model (figure 9), the average number of retrieved items is 0.5 for the third to fifth trials within the search phase meaning that the model is more uncertain when anticipating the future trial. When items accumulate in the working memory list after 2 incorrect trials, the model has 50% chance of retrieving the last item. Nevertheless, the entropy-based coordination is also able to produce the same pattern of binary retrieval during the search phase as shown by monkey g's fitted model. During the repetition phase, the model had the same behavior than the weight-based mixture with around 100% chance of retrieving the last item. To summarize, we found here a possible explanation for the fast reaction times during the search phase (compared to the repetition phase) for monkeys p, s and m: the specific use of working-memory during repetition in order to ensure correct performance by anticipating the next trial. This constitutes an important prediction by the model whose neural correlates could be explored experimentally.

The second most used variation through dual models is variation 7 with the conditioning of the update of working memory by the reward prediction error δ_t computed from the q-learning. Within each panel with this variation in figure 9 and 10, we computed the probability of update by averaging the number of times an item was integrated into the working memory list (dotted lines). For all fitted models using this new rule, we found that the probability of update was maximal at the end of the search phase i.e. when the first correct response is delivered to the agent. Thus, the fitted models of weight-based mixture (monkeys r and m) and entropy-based coordination (monkeys s and m) encoded positive outcomes inside the working memory only when the prediction errors were positive. In fact, the parameter ξ_2 that controls the upper bound was set between 0 and 0.5 for all fitted models (the parameter ξ_2 can take value in the range $[0, 20]$ during the optimization process) and the reward prediction error $\delta = 1$ at the first positive outcome.

23

Strikingly, none of the fitted models had a lower bound ξ_1 that allowed the encoding of a trial leading to a negative outcome. This makes sense since the retrieval of a memory item about a negative outcome brings less information and thus reduces less uncertainty than memory items about a positive outcome, as we previously explained in [21], making the storage of memory items about negative outcomes less beneficial than positive ones. Because of this blocking of negative outcomes to enter working memory, the number of retrieved memory items is null during the search phase and only increases during the repetition phase for all models. Similarly to the previously described variation 5, this selectivity of encoding in working memory is what allowed the reaction times to be lower in the search phase compared to the repetition phase. Lastly, the mean weight $w_t \approx 1.0$ in figure 10 for monkeys m and s indicates a reduction of the role of the q-learning to a simple prediction error signaler.

Finally, monkey g was best fitted by the second variation of the weight-based mixture. The behavior, especially the reaction times, was already very well fitted with the original model. Here we found that the optimization of the q-learning with $\gamma \in [0, 1[$ is the best fitted model according to the optimization process but the improvement of fit is minimal. Besides, the dynamic of the model shows a preference for the q-learning with a low w_t. Still, the working memory contributes to the final decision by constantly remembering the last trial.

4. Discussion

In this paper, we expanded and tested new models of coordination between working memory and reinforcement learning that were originally proposed in Viejo et al. [21] in order to explain monkeys' behavior (choices and reaction times). During a succession of problems (defined by the correct action), monkeys had to find one correct target amongst four. When the correct target was found, animals repeated the correct action for a various number of trials (to prevent anticipation of the end of a problem). The first round of optimization with the original models [21] proved that a combination of working memory and reinforcement learning were better at explaining choices and reaction times than just working memory or reinforcement learning alone, which was the main hypothesis developed in this paper.

The hypothesis that distinct memory modules co-exist in the brain is supported by a range of lesion data in human [40, 41, 42, 43], in monkeys [44, 45]
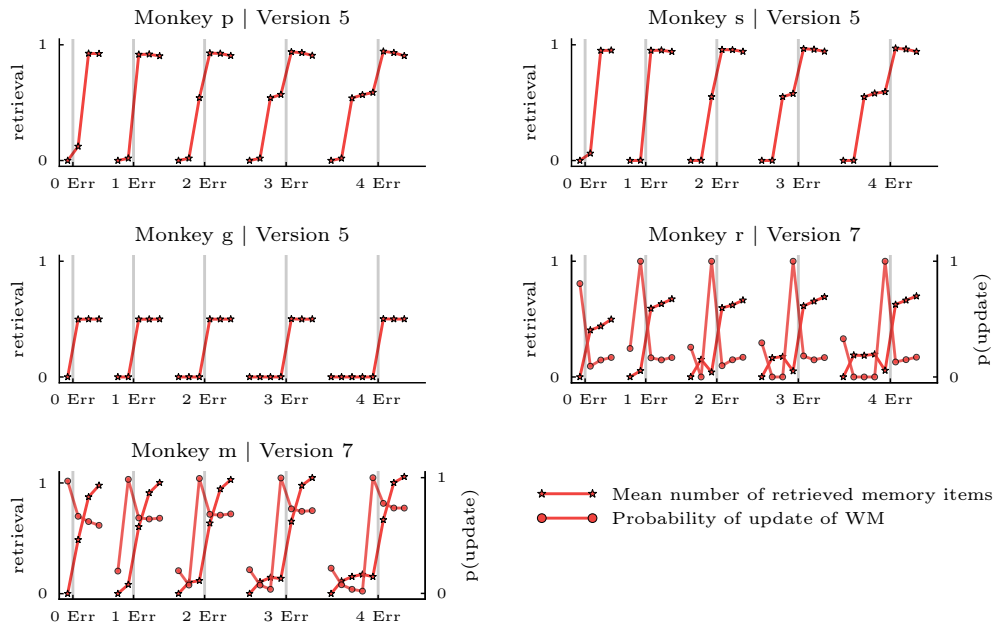
24

Figure 9: Variables of entropy-based coordination model for each best fitted model. The star lines show the mean number of retrieved memory items. For monkeys best fitted with the version 7, the dotted lines show the probability of update of the working memory depending of the reward prediction error δ of the q-learning.
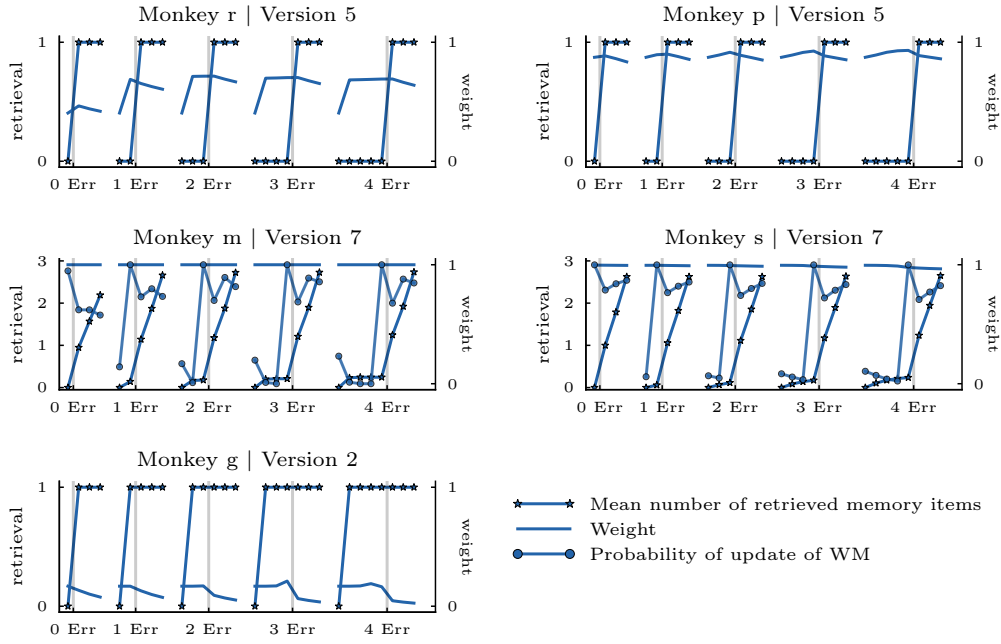
Figure 10: Variables of weight-based mixture model for each best fitted model. The star lines show the mean number of retrieved memory items. The straight lines show the mean weight w_t. For monkeys best fitted with the version 7, the dotted lines show the probability of update of the working memory depending of the reward prediction error δ of the q-learning.

26

and in rodents [46, 47, 48]. Instrumental conditioning studies brought light on the interaction between distinct memory modules by deciphering the transfer of control that occurs between the early stage of learning (i.e when behavior is considered goal-directed) and late stage of learning (i.e when behavior is considered habitual). Lesion studies showed that different sets of brain areas supported those two stages of learning [49, 50, 51, 52]. Thus, modeling studies used two different algorithms (respectively model-based learning algorithm and model-free reinforcement learning algorithm) for the two stage of learning with a transfer of control from model-based to model-free [53, 54]. Overall, the mapping between the reinforcement learning algorithm and brain activity during habitual behavior has been well described (see [55]) . Evidences for the mapping between model-based learning algorithm and brain activity during goal-directed behavior are scarcer. But evidence concerning specific neural substrates and properties for working-memory processes are supported by a vast literature (e.g. [31, 56, 57]. Here we do not assume a correspondence between working-memory processes and model-based learning. We simply consider that (1) WM belongs to a wider prefrontal cortex system dedicated to cognitive control, dedicated to inhibiting routine behaviors in response to environmental changes, and that (2) coordination mechanisms between model-based and model-free RL may be similar to coordination mechanisms between WM and model-free RL. In support of this, model-based and WM involve common prefrontal cortex regions [58, 59], regions such as the OFC being considered to encode the outcome of action and goal-directed action-outcome contingencies in working-memory [60]. Model-based processes actually do require working-memory when sequentially inferring the outcome of multiple actions within a cognitive graph [61]. Thus, there is a lot of possibilities in the combination and process of interaction of memory modules. The particular approach that we used to systematically compare different models of interaction between working memory and reinforcement learning using both choices and reaction times for each subject, is, in our sense, the best way to explore all the possibilities within the field of memory systems modeling.

Originally developed to fit the behavior of humans in a visuo-motor association task [26, 27], the models proved to be transferable to non-human primates. We then proceeded to improve the original models with different versions guided by the particular pattern of reaction times that we observed in monkeys: reaction times were lower for some trials during the search phase compared to the repetition phase. This observation opposes human's behav-

ior for which the models were originally developed. The general hypothesis of the interaction process in humans was stated as followed: working memory is used during the search phase by remembering previous trials in order to avoid the selection of incorrect actions inducing an increase of reaction times as errors accumulate, and the q-learning gradually suppresses the use of working memory during the repetition phase as it converges toward the optimal decision with the accumulation of positive outcomes inducing faster reaction times. In sharp contrast, the general tendency of monkeys' reaction times was to accelerate during the search phase and to slightly slow down during the repetition phase. Thus, we made the hypothesis that working memory retrieval was not the main strategy that was used during the search phase or that it was used differently in combination with the reinforcement learning strategy. Oddly, the requirement of both strategies goes again the simplicity of the task : monkeys need to remember only the last correct action in order to succeed. While the task calls only for a working memory strategy, we found that a model-free reinforcement learning strategy was required. This result suggest that model-free might operate as a default strategy in the brain as previously proposed in [15]. In [6], the authors reported a similar dual model of working memory and model-free reinforcement learning strategies in order to explain the choices of monkeys confronted to a biased matching pennies game against a computer opponent. Similarly, monkeys complemented the model-free reinforcement learning algorithm with a more flexible strategies that was best reproduced with a working memory model. Nevertheless, they didn't explored the various combinations that a dual-strategy offers as in this study.

In order to easily bias the use of working memory within a dual-strategy model, we first tested variations of q-learning with small changes: optimization of the discount factor, no initialization of q-values between problems and decay of q-values. In a second round of innovations, we tested more complex variations of working memory: (1) anticipation of the next trial during the search phase by preparing the probability of action, (2) meta-learning of mean entropies for suppressing the use of working memory when uncertainty is high on average, (3) encoding of past trials inside working memory conditioned by the reward prediction error from q-learning. Overall, we found that the anticipation of the next trial during the search phase and limiting the encoding of past trials were the best innovations to improve the fit to monkeys' behavior.

By analyzing the dynamics generated by simulating the best fitted mod-

els, we found that the anticipation of the next trial prevented in most dual models the retrieval process during the search phase and favored the retrieval of exactly one memory item (describing the last trial) during the repetition phase. The same process is at play when limiting the encoding of memory items (only the correct outcomes were included in the working memory list). Thus, it is very likely that the best theoretical model would incorporate the fact that working memory is somehow inhibited during the search phase and replaced by a more automatic behavior that can be different from the q-learning such as meta-learning of average uncertainty. We tested this approach by computing a table of mean entropies for each trial type and used this average uncertainties to bias the probability of retrieval of the entropy-based coordination models. Yet, this approach did not produce the best fit for explaining choices and reaction times.

This problem solving task has been studied in a series of articles using monkeys [15, 24, 62, 63] but also with humans in functional magnetic resonance imaging (fMRI) [64] and in electroencephalography (EEG) with 5 actions instead of 4 [65]. In both cases, the authors tried to correlate a reward prediction error $RPE = r_{obtained} - p_{correct} \cdot r_{expected}$ with the cerebral activity recorded during the search phase. In fMRI, activations in the dorsal anterior cingulate cortex (midcingulate cortex), the frontal insular cortex, the striatum, the retrosplenial cortex and the middle dorsolateral prefrontal cortex correlated with a positive RPE. In other words : a high RPE means a lower expectation of reward and this is associated with a high cerebral activity. More interestingly, this correlation disappears for the negative RPE. In EEG, the authors analyzed the event-related potentials (ERP) when the subjects receive the outcome. Contrary to the results in fMRI, ERPs correlated with positive and negative RPE within the frontal regions. Besides, an ERP also appeared for the start signal of a new problem indicating a possible process of monitoring the structure of the task and not only positive and negative outcomes. Those experimental results tend to validate our second best fitted models, that condition the update of working memory by the reward prediction error δ. Similar to the results in EEG, positive and negative reward prediction errors are used during the encoding phase of the task. Then, we found that only the positive outcomes were to be encoded within the working memory list which would resurged through neural activity (detected by fMRI) during the decision phase as a post-marker of the filtering during the update phase by the reward prediction error.

The opposite reaction times patterns that we observed here for the mon-

key behavioral data [15] compared to the human data [27] could be seen as an indication of inter-species differences in learning and decision-making strategies. When fitting our dual models to human data we previously found that working-memory was important to prevent the repetition of errors during search [21]. Here the model-based analyses suggest that working-memory in the five studied monkeys is important to ensure the repetition of correct response once the correct target has been discovered, while working-memory processes may present a cost-benefit ratio too low during search (retrieving a memory item about a negative outcome is less informative because it only tells which target not to select, while a positive outcome directly tells which target to select). An alternative explanation to the inter-species differences may be that the long pretraining phase that monkeys underwent for this task may have enabled them to learn more aspects of the task structure and hence to restrict their use of working-memory. We have tried to capture this phenomenon in two different ways here: 1) using meta-learning where the model learns the average uncertainty that results from deliberative processes at experienced type of trials (this enables the model to automatically learn that using working-memory during the search phase does not reduce much the uncertainty), and 2) using trial anticipation where the model retrieves a memory item in preparation of the decision at the next trial to ensure correct response repetition. The meta-learning variation of the tested models was never selected as best model. Nevertheless, the trial anticipation variation was consistently selected as best model for monkey p, and sometimes also for monkeys g, r and s. An important experimental prediction of these model variations is that humans undergoing the same long pretraining phase for this task would be able to decipher the task structure and thus to show the same opposite reaction times patterns than those observed in humans without pretraining in the task of Brovelli et al. [27]. Similarly, humans being given detailed instructions about the task could also extract sufficient information about task structure to display the same opposite reaction time patterns. A perspective of this work would be to apply the same model-based analysis to the human data gathered in the same problem-solving task by Sallet et al. [65]. The specific design of this task may have favored particular coordination dynamics between reinforcement learning and working memory.

The study of this problem solving task with non-human primates [15, 24, 62, 63] and human subjects [64, 65] clearly shows cerebral activity associated with the evaluation, encoding and monitoring of uncertainty associated with

decisions. Simple models of coordination between working memory and rein-forcement learning or working memory alone do not have this ability as they just encode the description of a trial. Meta-learning or anticipation as tested in this paper could thus bridge the gap between dual strategies models and high-level cognitive models as the fitted models indicate. To conclude, those models would be perfect to look for new computational variables that can be used for correlation with neuronal activity and to elucidate the processes taking place in the underlying brain structures.

[1] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, volume 1, MIT press Cambridge, 1998.

[2] K. M. Diederen, H. Ziauddeen, M. D. Vestergaard, T. Spencer, W. Schultz, P. C. Fletcher, Dopamine modulates adaptive prediction error coding in the human midbrain and striatum, Journal of Neuroscience 37 (2017) 1708–1720.

[3] J. P. ODoherty, Reward representations and reward-related learning in the human brain: insights from neuroimaging, Current opinion in neurobiology 14 (2004) 769–776.

[4] P. Dayan, N. D. Daw, Decision theory, reinforcement learning, and the brain, Cognitive, Affective, & Behavioral Neuroscience 8 (2008) 429–453.

[5] M. Ito, K. Doya, Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit, Current opinion in neurobiology 21 (2011) 368–373.

[6] H. Seo, X. Cai, C. H. Donahue, D. Lee, Neural correlates of strategic reasoning during competitive games, Science 346 (2014) 340–343.

[7] Q. J. Huys, T. V. Maia, M. J. Frank, Computational psychiatry as a bridge from neuroscience to clinical applications, Nature neuroscience 19 (2016) 404–413.

[8] S. Palminteri, M. Pessiglione, Opponent brain systems for reward and punishment learning: causal evidence from drug and lesion studies in humans, Decision Neuroscience. Academic Press, London, UK (2017).

[9] O. Hikosaka, E. Bromberg-Martin, S. Hong, M. Matsumoto, New insights on the subcortical representation of reward, Current opinion in neurobiology 18 (2008) 203–208.

[10] J. O. Gan, M. E. Walton, P. E. Phillips, Dissociable cost and benefit encoding of future rewards by mesolimbic dopamine, Nature neuroscience 13 (2010) 25–27.

[11] S. Palminteri, M. Khamassi, M. Joffily, G. Coricelli, Contextual modulation of value signals in reward and punishment learning, Nature communications 6 (2015).

[12] K. Samejima, Y. Ueda, K. Doya, M. Kimura, Representation of action-specific reward values in the striatum, Science 310 (2005) 1337–1340.

[13] M. Ito, K. Doya, Validation of decision-making models and analysis of decision variables in the rat basal ganglia, Journal of Neuroscience 29 (2009) 9861–9874.

[14] D. J. Barraclough, M. L. Conroy, D. Lee, Prefrontal cortex and decision making in a mixed-strategy game, Nature neuroscience 7 (2004) 404–410.

[15] M. Khamassi, R. Quilodran, P. Enel, P. Dominey, E. Procyk, Behavioral regulation and the modulation of information coding in the lateral prefrontal and cingulate cortex, Cereb Cortex (2015).

[16] Y. Niv, R. Daniel, A. Geana, S. J. Gershman, Y. C. Leong, A. Radulescu, R. C. Wilson, Reinforcement learning in multidimensional environments relies on attention mechanisms, Journal of Neuroscience 35 (2015) 8145–8157.

[17] A. Johnson, A. D. Redish, Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point, Journal of Neuroscience 27 (2007) 12176–12189.

[18] J. Gläscher, N. Daw, P. Dayan, J. O'Doherty, States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning, Neuron 66 (2010) 585–595.

[19] S. W. Kennerley, T. E. Behrens, J. D. Wallis, Double dissociation of value computations in orbitofrontal and anterior cingulate neurons, Nature neuroscience 14 (2011) 1581–1589.

[20] A. G. Collins, M. J. Frank, How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis, European Journal of Neuroscience 35 (2012) 1024–1035.

[21] G. Viejo, M. Khamassi, A. Brovelli, B. Girard, Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning, Frontiers in behavioral neuroscience 9 (2015).

[22] C. J. C. H. Watkins, Learning from delayed rewards, Ph.D. thesis, University of Cambridge England, 1989.

[23] E. Procyk, P. S. Goldman-Rakic, Modulation of dorsolateral prefrontal delay activity during self-organized behavior, The Journal of neuroscience 26 (2006) 11313–11323.

[24] R. Quilodran, M. Rothe, E. Procyk, Behavioral shifts and action valuation in the anterior cingulate cortex, Neuron 57 (2008) 314–325.

[25] J. D. Cohen, G. Aston-Jones, M. S. Gilzenrat, A systems-level perspective on attention and cognitive control: Guided activation, adaptive gating, conflict monitoring, and exploitation versus exploration. (2004).

[26] A. Brovelli, N. Laksiri, B. Nazarian, M. Meunier, D. Boussaoud, Understanding the neural computations of arbitrary visuomotor learning through fmri and associative learning theory, Cerebral Cortex 18 (2008) 1485–1495.

[27] A. Brovelli, B. Nazarian, M. Meunier, D. Boussaoud, Differential roles of caudate nucleus and putamen during instrumental learning, NeuroImage 57 (2011) 1580–1590.

[28] P. Enel, E. Procyk, R. Quilodran, P. F. Dominey, Reservoir computing properties of neural dynamics in prefrontal cortex, PLoS computational biology 12 (2016) e1004967.

[29] M. T. Todd, Y. Niv, J. D. Cohen, Learning to use working memory in partially observable environments through dopaminergic reinforcement, in: Advances in neural information processing systems, pp. 1689–1696.

[30] N. P. Rougier, D. C. Noelle, T. S. Braver, J. D. Cohen, R. C. O'Reilly, Prefrontal cortex and flexible cognitive control: Rules without symbols, Proceedings of the National Academy of Sciences of the United States of America 102 (2005) 7338–7343.

[31] P. S. Goldman-Rakic, Cellular basis of working memory, Neuron 14 (1995) 477–485.

[32] M. Pessiglione, V. Czernecki, B. Pillon, B. Dubois, M. Schüpbach, Y. Agid, L. Tremblay, An effect of dopamine depletion on decision-making: the temporal coupling of deliberation and execution, Journal of cognitive neuroscience 17 (2005) 1886–1896.

[33] R. Cools, M. D'Esposito, Inverted-u–shaped dopamine actions on human working memory and cognitive control, Biological psychiatry 69 (2011) e113–e125.

[34] S. B. Floresco, A. G. Phillips, Delay-dependent modulation of memory retrieval by infusion of a dopamine d1 agonist into the rat medial prefrontal cortex., Behavioral neuroscience 115 (2001) 934.

[35] F. Lesaint, O. Sigaud, S. Flagel, T. Robinson, M. Khamassi, Modelling individual differences in the form of pavlovian conditioned approach responses: A dual learning systems approach with factored representations, PLoS Comput Biol 10 (2014) e1003466.

[36] J. Liénard, B. Girard, A biologically constrained model of the whole basal ganglia addressing the paradoxes of connections and selection, J Comput Neurosci 36 (2014) 445–468.

[37] J.-B. Mouret, S. Doncieux, Sferes v2: Evolvin' in the multi-core world, in: WCCI 2010 IEEE World Congress on Computational Intelligence, Congress on Evolutionary Computation (CEC), Ieee, 2010, pp. 4079–4086.

[38] A. Wierzbicki, On the completeness and constructiveness of parametric characterizations to vector optimization problems, OR Spektrum 8 (1986) 73–87.

[39] S. Palminteri, V. Wyart, E. Koechlin, Computational cognitive neuroscience: Model fitting should not replace model simulation., bioRxiv (2016) 079798.

[40] W. B. Scoville, B. Milner, Loss of recent memory after bilateral hippocampal lesions, Journal of neurology, neurosurgery, and psychiatry 20 (1957) 11.

[41] S. Corkin, Acquisition of motor skill after bilateral medial temporal-lobe excision, Neuropsychologia 6 (1968) 255–265.

[42] S. Corkin, Lasting consequences of bilateral medial temporal lobectomy: Clinical course and experimental findings in hm, in: Seminars in Neurology, volume 4, © 1984 by Thieme Medical Publishers, Inc., pp. 249–259.

[43] B. J. Knowlton, J. A. Mangels, L. R. Squire, A neostriatal habit learning system in humans, Science 273 (1996) 1399.

[44] M. Mishkin, Memory in monkeys severely impaired by combined but not by separate removal of amygdala and hippocampus, Nature 273 (1978) 297–298.

[45] L. R. Squire, S. Zola-Morgan, The medial temporal lobe memory system, Science 253 (1991) 1380.

[46] R. Sutherland, R. McDonald, Hippocampus, amygdala, and memory deficits in rats, Behavioural brain research 37 (1990) 57–79.

[47] R. J. McDonald, N. M. White, A triple dissociation of memory systems: hippocampus, amygdala, and dorsal striatum., Behavioral neuroscience 107 (1993) 3.

[48] M. G. Packard, R. Hirsh, N. M. White, Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: evidence for multiple memory systems, Journal of Neuroscience 9 (1989) 1465–1472.

[49] M. G. Packard, J. L. McGaugh, Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning, Neurobiology of learning and memory 65 (1996) 65–72.

[50] E. Coutureau, S. Killcross, Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats, Behavioural brain research 146 (2003) 167–174.

[51] S. Killcross, E. Coutureau, Coordination of actions and habits in the medial prefrontal cortex of rats, Cerebral cortex 13 (2003) 400–408.

[52] H. H. Yin, B. J. Knowlton, The role of the basal ganglia in habit formation, Nature reviews. Neuroscience 7 (2006) 464.

[53] N. D. Daw, Y. Niv, P. Dayan, Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control, Nature neuroscience 8 (2005) 1704–1711.

[54] M. Keramati, A. Dezfouli, P. Piray, Speed/accuracy trade-off between the habitual and the goal-directed processes, PLoS computational biology 7 (2011) e1002055.

[55] Y. Niv, Reinforcement learning in the brain, Journal of Mathematical Psychology 53 (2009) 139–154.

[56] M. D. Esposito, J. A. Detre, D. C. Alsop, R. K. Shin, et al., The neural basis of the central executive system of working memory, Nature 378 (1995) 279.

[57] C. Ranganath, M. X. Cohen, C. Dam, M. D'Esposito, Inferior temporal, prefrontal, and hippocampal contributions to visual working memory maintenance and associative memory retrieval, Journal of Neuroscience 24 (2004) 3917–3925.

[58] M. G. Stokes, M. Kusunoki, N. Sigala, H. Nili, D. Gaffan, J. Duncan, Dynamic coding for cognitive control in prefrontal cortex, Neuron 78 (2013) 364–375.

[59] B. W. Balleine, J. P. O'doherty, Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action, Neuropsychopharmacology 35 (2010) 48.

[60] M. J. Frank, E. D. Claus, Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal., Psychological review 113 (2006) 300.

[61] R. C. Wilson, Y. K. Takahashi, G. Schoenbaum, Y. Niv, Orbitofrontal cortex as a cognitive map of task space, Neuron 81 (2014) 267–279.

[62] E. Procyk, Y. Tanaka, J.-P. Joseph, Anterior cingulate activity during routine and non-routine sequential behaviors in macaques, Nature neuroscience 3 (2000) 502–508.

[63] M. Rothé, R. Quilodran, J. Sallet, E. Procyk, Coordination of high gamma activity in anterior cingulate and lateral prefrontal cortical areas during adaptation, The Journal of Neuroscience 31 (2011) 11110–11117.

[64] C. Amiez, J. Sallet, E. Procyk, M. Petrides, Modulation of feedback related activity in the rostral anterior cingulate cortex during trial and error exploration, Neuroimage 63 (2012) 1078–1090.

[65] J. Sallet, N. Camille, E. Procyk, Modulation of feedback-related negativity during trial-and-error exploration and encoding of behavioral shifts, Frontiers in neuroscience 7 (2013) 209.