A computational model of integration between reinforcement learning and task monitoring in the prefrontal cortex

Mehdi Khamassi, Rene Quilodran, Pierre Enel, Emmanuel Procyk, and Peter F. Dominey

INSERM U846 SBRI, Bron, France
correspondance: mehdi.khamassi@inserm.fr

Résumé Taking inspiration from neural principles of decision-making is of particular interest to help improve adaptivity of artificial systems. Research at the crossroads of neuroscience and artificial intelligence in the last decade has helped understanding how the brain organizes reinforcement learning (RL) processes (the adaptation of decisions based on feedback from the environment). The current challenge is now to understand how the brain flexibly regulates parameters of RL such as the exploration rate based on the task structure, which is called meta-learning ([1] : Doya, 2002). Here, we propose a computational mechanism of exploration regulation based on real neurophysiological and behavioral data recorded in monkey prefrontal cortex during a visuo-motor task involving a clear distinction between exploratory and exploitative actions. We first fit trial-by-trial choices made by the monkeys with an analytical reinforcement learning model. We find that the model which has the highest likelihood of predicting monkeys' choices reveals different exploration rates at different task phases. In addition, the optimized model has a very high learning rate, and a reset of action values associated to a cue used in the task to signal condition changes. Beyond classical RL mechanisms, these results suggest that the monkey brain extracted task regularities to tune learning parameters in a task-appropriate way. We finally use these principles to develop a neural network model extending a previous cortico-striatal loop model. In our prefrontal cortex component, prediction error signals are extracted to produce feedback categorization signals. The latter are used to boost exploration after errors, and to attenuate it during exploitation, ensuring a lock on the currently rewarded choice. This model performs the task like monkeys, and provides a set of experimental predictions to be tested by future neurophysiological recordings.

1 Introduction

Exploring the environment while searching for resources requires both the ability to generate novel behaviors and to organize them for an optimal efficiency. Besides, these behaviors should be regulated and interrupted when the goals of exploration have been reached : a transition towards a behavioral state

called exploitation should then be implemented. Previous results on neural bases of these functions in the frontal cortex showed crucial mechanisms that could participate both to reinforcement learning processes [2] and to the auto-regulation of exploration-exploitation behaviors [3]. Several computational and theoretical models have been proposed to describe the collaborative functions of the anterior cingulate cortex (ACC) and the dorsolateral prefrontal cortex (DLPFC) - both belonging to the prefrontal cortex - in adaptive cognition [4, 5, 6]. Most models are based on the hypothesized role for ACC in performance monitoring based on feedbacks and of DLPFC in decision-making. In exploration, challenging, or conflicting situations the output from ACC would trigger increased control by the DLPFC. Besides, several electrophysiological data in non human primates suggest that modulation of this control within the ACC-DLPFC system are subserved by mechanisms that could be modelled with the reinforcement learning (RL) framework [2, 7, 8]. However, it is not clear how these mechanisms integrate within these neural structures, and interact with subcortical structures to produce coherent decision-making under explore-exploit trade-off.

Here we propose a new computational model to formalize these frontal cortical mechanisms. Our model integrates mechanisms based on the reinforcement learning framework and mechanisms of feedback categorization - relevant for task-monitoring - in order to produce a decision-making system consistent with behavioral and electrophysiological data reported in monkeys. We first employ the reinforcement learning framework to reproduce monkeys exploration-exploitation behaviors in a visuo-motor task. In a second step, we extract the main principles of this analysis to implement a neural-network model of fronto-striatal loops in learning through reinforcement to adaptively switch between exploration and exploitation. This model enabled to reproduce monkeys behavior and to draw experimental predictions on the single-unit activity that should occur in ACC and DLPFC during the same task.

2 Problem-solving task (PST)

We first use behavioral data recorded in 2 monkeys during 278 sessions (7656 problems $\equiv$ 44219 trials) of a visuo-motor problem-solving task that alternates exploration and exploitation periods (see Fig.1A). In this task, monkeys have to find which of four possible targets on a screen is associated with reward. The task is organized as a sequence of problems. For each problem, one of the targets is the correct choice. Each problem is organized in two succesive groups of trials ; starting with search trials where the animal explores the set of targets until finding the rewarded one ; Once the correct target is found, a repetition period is imposed so that the animal repeats the correct response at least three times. Finally, a cue is presented on the screen and indicates the end of the current problem and the beginning of a new one. Data used here were recorded during electrophysiological experiments, after animals had experienced a pre-training stage. Thus, monkeys are particularly overtrained and optimal on this

task. Monkey choice, trial correctness and problem number are extracted and constitute the training data for the reinforcement learning model.

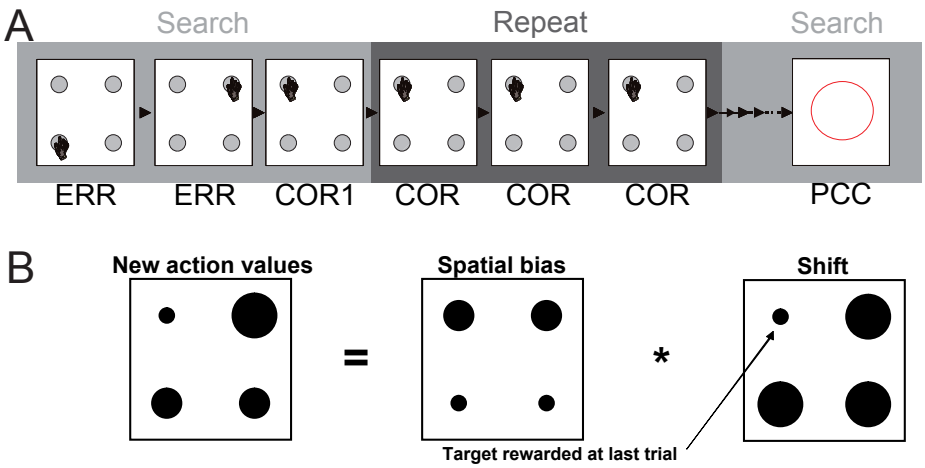

Figure 1. Monkeys had to find by trial and error which target, presented in a set of four, was rewarded. A) Monkeys performed a set of trials where they chose different targets until the solution was discovered (search period). Each block of trials (or problem) contained a search period and a repetition period during which the correct response was repeated at least three times. A Problem-Changing Cue (PCC) is presented on the screen to indicate the beginning of a new problem. B) Action value reset in the model at the beginning of each new problem.

3 Behavior analysis with the Reinforcement Learning framework

3.1 Theoretical model description

We use the reinforcement learning framework as a model of the way monkeys learn to choose appropriate targets by trial-and-error [9]. The main assumption in such framework is that monkeys try to maximize the amount of reward they will get during the task. This framework assumes that animals keep estimated action values (called Q-values) for each target (i.e. $Q(UL)$, $Q(LL)$, $Q(UR)$ and $Q(LR)$). It also assumes that monkeys decide which action to perform depending on these values, and update these values based on feedbacks (i.e. the presence/absence of reward) at the end of each trial. We used a Boltzmann softmax rule for action selection. The probability of choosing an action a (either UL, LL, UR or LR) is given by

$$P(a) = \frac{exp(\beta Q(a))}{\sum_b exp(\beta Q(b))} \quad (1)$$

where β is an exploration rate ($\beta \geq 0$). In short, when β is low (close to 0), the contrast between action values is decreased, thus increasing the probability to select a non optimal action (exploration). When β is high, the contrast is high and decision-making becomes more greedy. We differently use β_S and β_R parameters on *search* and *repetition* trials so as to allow different shapes of the Boltzmann function on these two periods. In other words, β_S and β_R were used as two distinct free parameters to see if they would converge on different values, hence indicating meta-learning through the use of two different exploration rates by the animal.

At the end of each trial, the action value is updated by comparing the presence/absence of reward r with the value expected from the performed action according to the following equation

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r - Q(a)) \tag{2}$$

where α is the learning rate of the model ($0 \leq \alpha \leq 1$). Similarly to previous work [2], we generalize reinforcement learning to also update each non chosen action b according to the following equation

$$Q(b) \leftarrow (1 - \kappa) \cdot Q(b) \tag{3}$$

where κ is a forgetting rate ($0 \leq \kappa \leq 1$).

Finally, we add an action-value reset at the beginning of each new problem, when a *PCC* cue appears on the screen. This is based on the observation that monkeys almost never select the previously rewarded target, and have individual spatial biases in their exploration pattern : they often start exploration by choosing the same preferred target (see Fig.1B).

3.2 Simulation of the RL model on monkey behavioral data

The reinforcement learning model is simulated on monkey data, that is, at each trial, the model chooses a target, we store this choice, then we look at the choice made by the animal, and the model learns as if it had made the same choice. At the next trial, the model makes a new choice, and so on. At the end, we compare the sequence of choices made by the model with the monkey's choices. With this method, the model learns based on the same experience as the monkey. Thus the choice made at trial t becomes comparable to the animal's choice at the same trial because it follows the same trial history $\{1...t-1\}$. For each behavioral session, we optimize the model by finding the set of parameters that provides the highest likelihood of fitting monkeys choices. This optimization leads to an average likelihood of 0.6537 per session corresponding to 77% of the trials where the model predicted the choice the monkeys actually made. Fig.2 shows simulation results on a sample of 100 trials for 1 monkey.

Interestingly, we find that the distribution of each session's β_S used to set the exploration rate during *search* periods is significantly lower than the distribution of β_R used for *repetition* periods (ANOVA test, $p < 0.001$). The mean β_S equals 5.0 while the mean β_R equals 6.8. This reveals a higher exploration

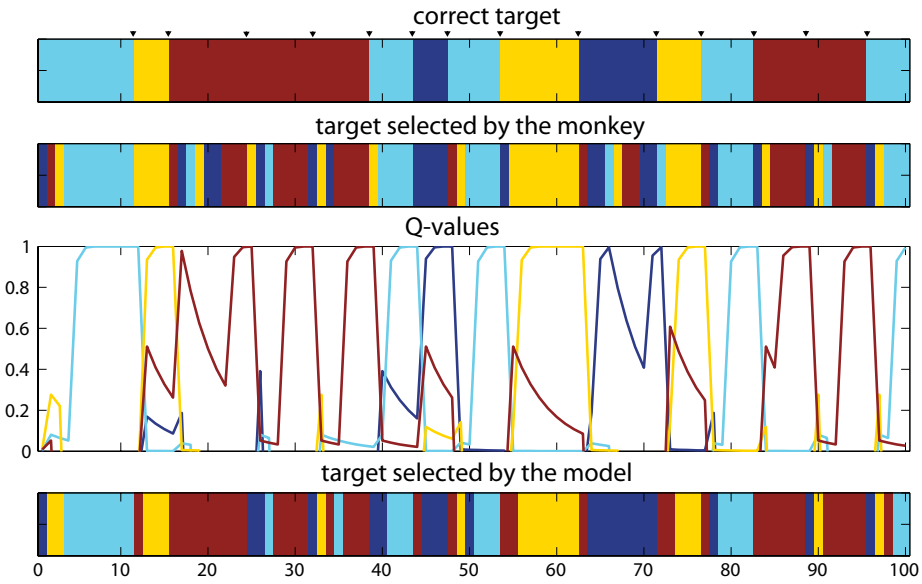

Figure 2. Simulation of the reinforcement learning model on 100 trials. Each color is associated with a different target (UL, LL, UR, LR). The top line denotes the problem sequence experienced by both the monkey and the model. Black triangles indicate cued problem changes. The second line shows the monkey's choice at each trial. Curves show the temporal evolution of action values in the model. Non selected target have their value decrease according to a forgetting process. These curves also show the action value reset at the beginning of each problem, the decrease of incorrect selected targets value, and the increase of the correct targets value once selected by the animal. The bottom of the figure shows choices made by the model based on these values.

rate in monkeys' choices during *search* periods. In addition, we found an average learning rate around 0.9 for the two monkeys and a smaller forgetting rate (mean : 0.45). This suggests that reinforcement learning mechanisms in the monkey brain are regulated by parameters that were learned from the task structure. In contrast, raw reinforcement learning algorithms such as Q-learning usually employs a single fixed β value, and need to make errors before abandoning the optimal action and starting a new exploration phase. In the next section, we extract these principles to propose a neural-network model integrating such reinforcement learning and task monitoring mechanisms.

4 Neural network model

4.1 Reinforcement learning processes

We propose a neural network model in order to propose a computational hypothesis concerning the modular organization of these processes within cortical

networks. Our model extends previous models of cortico-striatal loops which are known to be crucial neural substrates for reward-based learning and decision-making [10, 11]. The principle novelty here is to have the integration of reinforcement learning and task monitoring within the ACC-DLPFC system that produces explore-exploit behaviors. In our neural network, dopaminergic (DA) neurons from the Ventral Tegmental Area (VTA) compute a reward prediction error following equation 2, consistently with DAs supposed role in reinforcement learning [12]. DA signals are used to update action values encoded within the ACC, consistently with previous work [7]. These action values are then sent to DLPFC which makes decision of the target to choose and biases action selection in the striatum. Similarly to classical basal ganglia models but not detailed here, appropriate action in the striatum is selected by desinhibiting the thalamus through the substantia nigra pars reticulata [13, 14]. Finally, the thalamus projects to the motor cortex which drives behavioral output, and which sends efferent copies to the ACC in order to update only relevant action through reinforcement learning (fig.3).

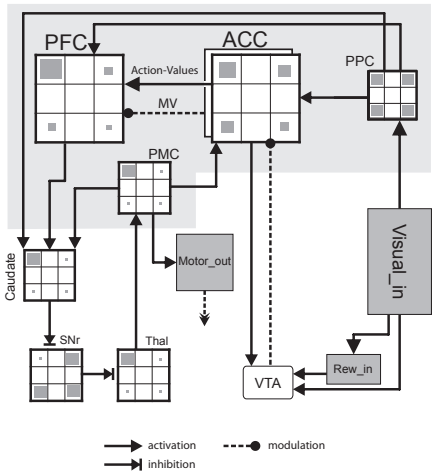

Figure 3. Neural network model. Visual input (targets presented on the screen) is processed by the parietal cortex and sent to the ACC-DLPFC system (colored in grey) which performs reinforcement learning (RL) to rapidly adapt choices by trial-and-error during the search period. A specific visual signal is used to indicate reward delivery, representing juice obtained by monkeys. According to RL principles, this visual signal is translated by VTA into a reinforcement signal which changes action values within the ACC. In parallel, this reinforcement signal is used to regulate the level of exploration with MV.

With such organization, the system is yet purely dedicated to reinforcement learning. In order to add task monitoring mechanisms, we take inspiration from additional results measured in the PST task. In [8], reaction times were observed to decrease gradually after errors during the search period, to raise sharply after the first correct trial, and to remain high during repetition (fig.4A-B). The exact opposite pattern was observed at the level of the average activity measured in DLPFC neurons ([15] ; fig.4C). These patterns suggest an integration of feedbacks used to update a control or attentional level, and a state change of the system from exploration to exploitation. This resembles the vigilance level employed in [16]'s theoretical model to modulate the level of activation of a glo-

bal workspace in charge of solving the task. In the next paragraph, we provide a computational hypothesis on the way the ACC could evaluate such kind of vigilance level to modulate the level of control and exploration in DLPFC.

4.2 Integrating task monitoring signals within the neural network

In order to regulate exploration based on feedbacks obtained from the environment, we add to our ACC component a second population of neurons dedicated to feedback categorization as described in the monkey ACC in the same task [8]. In our model, these neurons receive the same dopaminergic reward prediction error signals as ACC action value neurons. The difference resides in the influence of such DA signals on feedback categorization neurons. The latter either are inhibited by DA signals and thus produce responses to errors (ERR) or are excited by DA signals and thus produce responses to correct trials (COR). The high learning rate used in the model to fit behavioral data in section 3 results in a strong response of COR neurons to the first reward and in a smaller response to subsequent ones. This produces neurons responding to the first correct trials (COR1) as observed by [8]. Fig.5 shows a simulation of these neurons response patterns. COR and ERR signals are then used to update a modulatory variable (MV) according to the following equation :

$$MV \leftarrow MV + \alpha^+ \cdot \delta_+ + \alpha^- \cdot \delta_- \tag{4}$$

Where δ_+ and δ_- represent the response of correct and error neurons respectively, while α^+ and α^- are synaptic weights set to $-\frac{5}{2}$ and $\frac{1}{4}$ for the considered task. MV is constrained between 0 and 1. This equation makes MV be :

- sharply decreased after a correct trial (COR) ;
- increased after an error (ERR) ;
- increased after presentation of the problem changing cue (PCC). Although we did not yet study how the model works during pretraining phases of this task (*i.e.* habituation phases preceding electrophysiological recordings), we observed that before the animal learns what the PCC means, the presentation of this cue is very often followed by an error - because the animal persists in repeating the same choice while the problem has changed. Thus we consider here that the PCC has been associated to an error during the pretraining phase and consequently produces an increase of MV each time it occurs during the task.

Importantly, MV is used to modulate the exploration rate and the gain in the DLPFC. The first function is assured in the following manner :

$$\beta_t = \frac{\omega_1}{(1 + exp(\omega_2 * (1 - MV_t) + \omega_3))} \tag{5}$$

Where ω_1, ω_2 and ω_3 are parameters respectively equal to 10, -6 and 1. Such function is a sigmoid which inverses the tendency of MV (see fig.5) and transforms a value between 0 and 1 (for MV) into a value between 0 and 10 (for β) according to table 1.

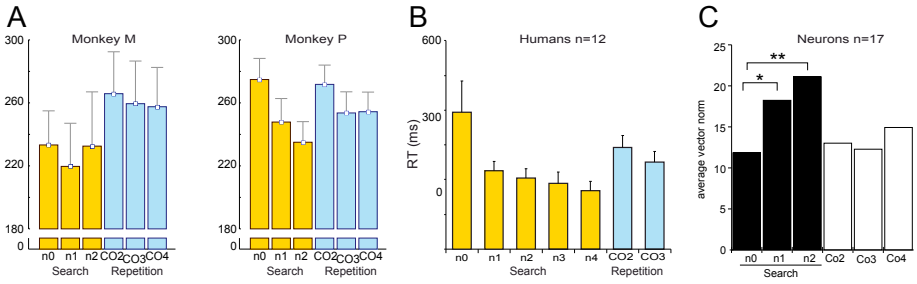

Figure 4. (A-B) Reaction times during the PST task show a progressive decrease along the search period, and a sharp change during repetition. Adapted from [8]. C) Average activity in the dorsolateral prefrontal cortex show a similar (inversed) pattern. Adapted from [15].

MV	0.00	0.25	0.50	0.75	1.00
β	9.9	9.7	8.8	6.2	2.7

Table 1. MV effect on β following equation (5) with a = 10, b = -6, c = 4.4

The second function is assured by weighting DLPFCs activity by multiplying it by MV (which is always inferior or equal to 1). As a consequence, a low MV produces a high β (almost no exploration) and a low DLPFC activity so as to focus and lock the DLPFC on performing the action with the highest value; whereas a high MV produces a low β (higher stochasticity in decision-making, thus more exploration) and a high activity in DLPFC so as to enable the performance of non optimal actions.

The model can perform the task like monkeys, alternating between search and repetition phases. Fig.5 shows the activation of different neurons in the model during a sample simulation.

5 Discussion and conclusion

We implemented a reinforcement learning model that can monitor exploration-exploitation trade-off in a monkey visuo-motor task. The model helped us formally describe monkey behavior in a task involving clear distinction between *search* and *repetition* trials. In addition, the model is based on existing anatomical and physiological properties of the monkey brain. Properties of MV modulation in our model are consistent with data in human and in animal showing a higher involvement of ACC-DLPFC when the task is demanding or when it involves conflict resolution [17]. Moreover, our results are consistent with previous electrophysiological work suggesting a role of the ACC-DLPFC system in performance monitoring [5], and in reinforcement learning [2, 7]. Our model enables to draw a list of experimental predictions that have to be tested

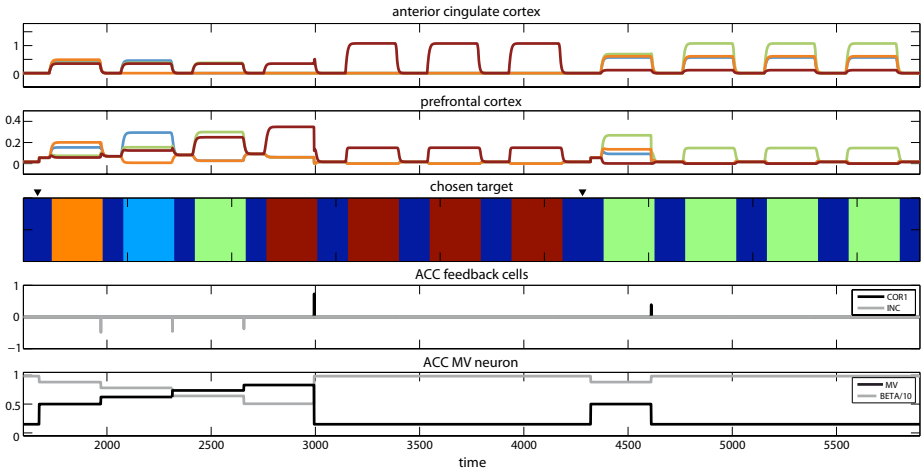

Figure 5. Neural network model simulation during 2 consecutive problems. Black triangle indicate cued problem changes.

by simultaneously recording Anterior Cingulate Cortex (ACC) and dorsolateral Prefrontal Cortex (DLPFC) neurons in this task :

1. There should exist MV neurons in ACC. Such MV neurons would have a particular profile of activity : progressive increase of activity during the search period, drop of activity after the first correct response, activity remaining low during the repetition period (as shown on fig.5).
2. MV modulation should effect only on DLPFC action value neurons and not on ACC action value neurons. In the model, we made the choice to keep original action values (that is, not altered by the MV modulation) in the ACC so as to have part of the system properly perform the reinforcement learning algorithm without perturbation, so as to ensure convergence.
3. There should be a higher global spatial selectivity - which reflects the degree to which neurons discriminate choices of spatial targets on the touch screen [15] - in DLPFC than in ACC due to the decision-making process based on the softmax function (which increases contrasts between action values when β is high).
4. There should be an increase of spatial selectivity in DLPFC but not in ACC during the repetition period. Such increase of spatial selectivity in DLPFC neurons in the model is due to the modulatory effect of MV on the β parameter used in the softmax function.

Performance of the neural-network model enables a robotics arm to reproduce monkey behavior in front of a touch screen. Such a pluridisciplinary approach provides tools both for a better understanding of neural mechanisms of decision making and for the design of artificial systems that can autonomously extract regularities from the environment and interpret various types of feed-

backs (rewards, feedbacks from humans, etc...) based on these regularities to appropriately adapt their own behaviors.

Future work will consist in modelling how RL parameters are progressively set during familiarization with the environment. Such goal can be achieved by using the model to predict day-by-day behavior observed during monkey pretraining. This will help us understand the dynamics of meta-learning which enable animals in this task to autonomously learn that a high learning rate is relevant and that clear transition between exploration and exploitation are required - based on the extracted structure of task.

Acknowledgments This work was supported by the French National Research Agency (ANR Amorces) and the European Community Contract FP7-231267 (EU Organic Project).

Références

[1] Doya, K. : Metalearning and neuromodulation. Neural Netw **15**(4-6) (2002) 495–506

[2] Barraclough, D., Conroy, M., Lee, D. : Prefrontal cortex and decision making in a mixed-strategy game. Nat Neurosci **7**(4) (2004) 404–10

[3] Procyk, E., Tanaka, Y., Joseph, J. : Anterior cingulate activity during routine and non-routine sequential behaviors in macaques. Nat Neurosci **3**(5) (2000) 502–8

[4] Aston-Jones, G., Cohen, J. : Adaptive gain and the role of the locus coeruleus-norepinephrine system in optimal performance. J Comp Neurol **493**(1) (2005) 99–110

[5] Brown, J., Braver, T. : Learned predictions of error likelihood in the anterior cingulate cortex. Science **307** (2005) 1118–21

[6] Dosenbach, ., Visscher, K., Palmer, E., F., M., Wenger, K., Kang, H., Burgund, E., Grimes, A., Schlaggar, B., Peterson, S. : A core system for the implementation of task sets. Neuron **50** (2006) 799–812

[7] Matsumoto, M., Matsumoto, K., Abe, H., Tanaka, K. : Medial prefrontal cell activity signaling prediction errors of action values. Nat Neurosci **10** (2007) 647–56

[8] Quilodran, R., Rothe, M., Procyk, E. : Behavioral shifts and action valuation in the anterior cingulate cortex. Neuron **57**(2) (2008) 314–25

[9] Sutton, R., Barto, A. : Reinforcement Learning : An Introduction. MIT Press, Cambridge, MA (1998)

[10] Dominey, P., Arbib, M., Joseph, J. : A model of corticostriatal plasticity for learning oculomotor associations and sequences. Journal of Cognitive Neuroscience **7**(3) (1995) 311–336

[11] Khamassi, M., Martinet, L., Guillot, A. : Combining self-organizing maps with mixture of epxerts : Application to an Actor-Critic model of reinforcement learning in the basal ganglia. In : Proceedings of the 9th International Conference on the Simulation of Adaptive Behavior (SAB), Rome, Italy, Springer-Verlag (2006) 394–405

[12] Schultz, W., Dayan, P., Montague, P. : A neural substrate of prediction and reward. Science **275**(5306) (1997) 1593–9

[13] Gurney, K., Prescott, T., Redgrave, P. : A computational model of action selection in the basal ganglia. I. A new functional anatomy. Biol Cybern **84**(6) (2001) 401–10

[14] Girard, B., Cuzin, V., Guillot, A., Gurney, K., Prescott, T. : A basal ganglia inspired model of action selection evaluated in a robotic survival task. Journal of Integrative Neuroscience **2**(2) (2003) 179–200

[15] Procyk, E., Goldman-Rakic, P. : Modulation of dorsolateral prefrontal delay activity during self-organized behavior. J Neurosci **26**(44) (2006) 11313–23

[16] Dehaene, S., Changeux, J. : A neuronal model of a global workspace in effortful cognitive tasks. Proc Natl Acad Sci U S A **95** (1998) 14529–34

[17] Cohen, J., Aston-Jones, G., Gilzenut, S. : A systems-level perspective on attention and cognitive control. In Posner, M., ed. : Cognitive Neuroscience of Attention. Guilford Publications (2004) 71–90