Reinforcement learning model in probalistically rewarded task

Pierre Enel, Mehdi Khamassi, Emmanuel Procyk, Peter Dominey

REINFORCEMENT LEARNING MODEL IN PROBABILISTICALLY REWARDED TASK

Pierre Enel, Mehdi Khamassi, Emmanuel Procyk and Peter F. Dominey
INSERM U846 Stem Cell and Brain Research Institute
18 avenue Doyen Lépine 69675 BRON cedex
France
{pierre.enel,mehdi.khamassi,emmanuel.procyk,peter.dominey}@inserm.fr

ABSTRACT

Adapting resource seeking behavior is of primary importance in survival. Then, balancing exploration and exploitation of discovered resources is at the core of adaptation to the environment. The reinforcement learning theoretical framework has been elaborated to formalize such reward seeking behavior. Biologically plausible models based on this algorithm have flourished recently. Among them, a neural network model was developed to investigate the functions of the anterior cingulate cortex (ACC) and the dorsolateral prefrontal cortex (DLPFC) involved in action valuation and action selection, respectively [1]. This model proposes a method to regulate dynamically the exploration inspired by literature on meta-learning in order to solve dynamically the exploration/ exploitation trade-off [2]. This model performed well in a deterministic problem solving task (PST). Our goal was to demonstrate that the model is generalizable to a more ecological PST with probabilistically dispensed rewards. The model was tested with its preset learning rate / exploration rate / initial action values and then optimized with search of the parameters space. The initial values of model's parameters proved to be good however not optimal for the new task. Interestingly, the model's performance is very dependent on the initial action values.

KEY WORDS
neural simulation, computational model, anterior cingulate cortex, prefrontal cortex, decision making, exploration

1. Introduction

In the context of evolution, an organism needs to explore its environment to find new resources in order to survive. However, endless exploration may lead to fewer resources than exploiting the knowledge of the environment acquired through exploration. Thus, exploitation and exploration constitute a trade-off that needs to be balanced so that the organism's behavior is optimal. The now well known framework of reinforcement learning (RL) has been developed to formalize the decision making process taking place in this context [3]. Furthermore, several studies have shown evidence for the neural substrate of RL, supporting the plausibility that RL may take place in the brain [4,5]. Hypotheses on the role of the noradrenaline projections and its contribution to RL processes have emphasized the regulation of the exploration-exploitation trade-off [6]. In addition, research on human and non-human primates have shown activity related to action values and reinforcement signals in the anterior cingulate cortex (ACC) and dorsolateral prefrontal cortex (DLPFC) under varying conditions of the exploration-exploitation trade-off [7-10]. These regions may thus play a key role in the regulation of explore-exploit behavior strategy. ACC would encode the reward value associated to an action – referred to as action values. In addition, the literature on ACC suggests that this area may extract regularities of the environment to adjust the speed of adaptation [11]. In contrast, DLPFC may be involved in the action selection. Computational models have been developed to further investigate the implication of different brain areas known to play a role in the RL process [12-14] but few have focused, on the two areas mentioned above.

We focus here on a model developed to analyze electrophysiological data of the ACC and DLPFC obtained from behaving monkeys [1,10] (figure 1). In addition to implement these two areas, the model investigated the presumed role of the ACC in regulating the exploration rate. Our investigations led to the implementation of a meta-parameter MV (for modulatory variable) that solved the exploration/ exploitation trade-off in accordance with literature on meta-learning [2]. The meta-parameter dynamically solve the exploration/exploitation trade-off by modulating the exploration rate β of the model.. The model performed well on a PST where rewards are distributed in a deterministic fashion [1].

Here we wanted to test whether the model could be generalized to a more ecological task (used in previous monkey experiments) where rewards are probabilistically distributed [15]. In this task, monkeys had to choose among two targets, each rewarded with a different probability. In terms of machine learning, this task is referred to as a 2-armed bandit problem. Monkeys were not systematically given the same reward for a given target, and hence had to make their choice based on several feedbacks rather than just one as in the deterministic PST task. We first optimized the model on its parameters except β that was regulated by the meta-parameter MV. We then optimized the model

on the exploration rate β (without MV) in order to see if it could yield an optimal solution without auto-regulation.

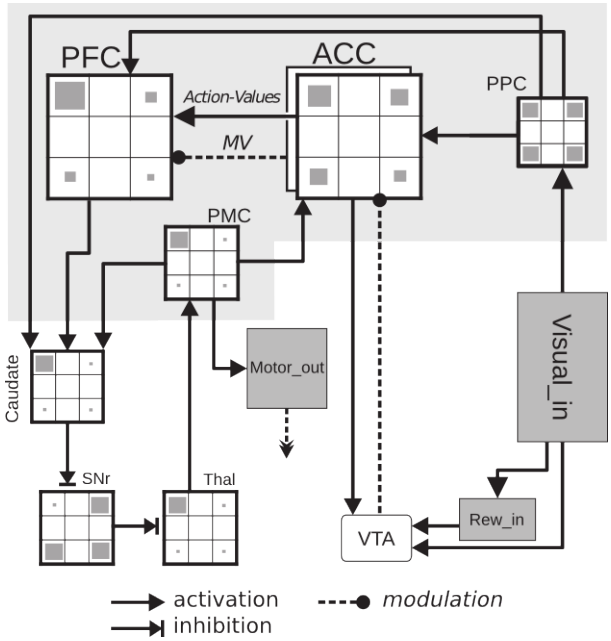

Figure 1. Khamassi et al. neural network model. Anterior cingulate cortex (ACC) sends the action values to the prefrontal cortex (PFC), which in turn sends its output to the caudate nucleus. The caudate module is the first module of the cortico-striatal loop. Output of the loop is the premotor cortex (PMC). Computed in ACC, MV modulates the selection of action taking place in PFC. The ventral tegmental area (VTA) sends the reward feedback to the ACC to update the action value linked to the last action performed. PPC: posterior parietal cortex; SNr: substantia nigra pars reticula; Thal: thalamus.

We hypothesized that the model would be able to generalize to this new task, and that the stochastic nature of the task would lead the optimized model to employ different parameter values than for the deterministic task (i.e. a lower learning rate, a smoother transition between exploration and exploitation).

2. Materials and Methods

2.1 Deterministic PST

The model described below has been implemented for a PST developed by Quilodran et al. [10] to investigate neuronal activity changes elicited by a shift from exploration to exploitation phases. In this PST a monkey explored among four targets to determine which one is rewarded. Rewards were given in a deterministic way in the sense that the monkeys either received a fixed amount of fruit juice or nothing. With this task, the exploration phase continues until the monkey finds the first reward. Then it could repeat the same action: touching the last chosen target for a variable number of rewarded trials (exploitation). At the end of exploitation a Signal to Change (SC) indicated the beginning of a new exploration.

2.2 Reinforcement learning algorithm

The Khamassi et al. [1] model was based on the RL framework and was used to formalize the decision making process of behaving monkeys. In this framework it is assumed that the monkey tries to maximize the amount of reward it gets while performing the problem solving task (PST). The specific RL algorithm that was used was the Q-learning algorithm which associates reward value to actions. These associations between action and reward were stored in Q-values ($Q(a_i)$, where 'a_i' is the action i). As described below, the actions corresponded to responses to each of four targets on a touch-screen. These action-values were references used by the algorithm to choose which action to perform and were updated with the feedback associated with each action. The discrepancy between the actual and the expected reward was defined as the reward prediction error (RPE) computed as following:

$$RPE \leftarrow Q(a_i) - r \qquad (1)$$

where r is the reward. It was then applied in the action-value update formula:

$$Q(a_i) \leftarrow Q(a_i) + (\alpha . RPE) \qquad (2)$$

α is the learning rate, modulating the impact of the last RPE on the new computed Q-value (α was bounded to the interval [0 ; 1]). Then the action choice was selected with probabilities computed from the action-values with the Boltzman softmax rule:

$$P(a_i) = \frac{\exp(\beta . Q(a_i))}{\sum_j \exp(\beta . Q(a_j))} \qquad (3)$$

where beta regulates the exploration rate. A small beta (close to 0) leads to a very similar probability for each action, while a high beta increases the difference between the highest action value and the others. In other words, the Q-value of the last selected action was updated with the difference between the actual reward and this Q-value (eq. 1 & 2). Out of the Q-values, including the one that was just updated, the Boltzman softmax rule computed a probability of selection for each action (eq. 3). These probabilities are then used to select a final action to perform.

The modulatory variable (MV) is a meta-parameter that regulates β. It was implemented after observing that ACC neurons' activity in Quilodran PST increased during the exploration phase and sharply decreased at the onset of exploitation phase.

$$\beta = \frac{\omega_1}{\left(1 + \exp(\omega_2 . (1 - MV) + \omega_3)\right)}$$

with $\omega_1 = 10$, $\omega_2 = -6$ and $\omega_3 = 1$. Figure 2 shows the relation between β and MV.

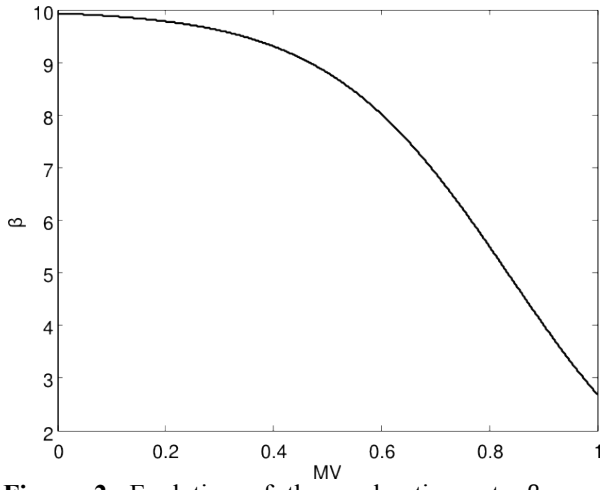

Figure 2. Evolution of the exploration rate β as a function of the modulatory variable MV. The higher MV the lower β.

MV is updated with the RPE:

$$MV \leftarrow MV + \begin{cases} \alpha_+ . RPE & if\ RPE > 0 \\ \alpha_- . RPE & if\ RPE < 0 \end{cases} \quad (4)$$

with $\alpha_+ = 2{,}5$ and $\alpha_- = 0{,}25$. MV was bounded to the interval [0 ; 1]. This last equation updated the modulatory variable to slowly increase the exploration rate when low rewards were given but sharply decrease the exploration rate if a high reward is given, leading to an exploitative behavior. The parameters α_+ and α_- have been chosen so as to fit the behavior of monkeys performing the deterministic task and to have a MV variable that reproduce some properties of ACC neurons that may modulate the exploration. They are similar to parameters used to regulate a "vigilance" level in Dehaene et al. [16]. Their level increases after errors and decreases after correct responses. They used separate learning rate for errors and correct responses so as to allow a non-symetric dynamics. In our case, monkey behavior and neural activities observed in the PST task show a progressive integration of errors, and a sharp decrease after a correct response (marking a shift from exploration to exploitation). Hence, the choice of values we used for α_+ and α_-.

Since the beginning of each new problem is cued by a Signal to Change (SC) and since monkeys are pretrained, they show a flexible change in their selected target after an SC. Thus, when an SC is presented, we reset Q-values to a default value Q_{init} which constitutes an additional parameter of the model. This parameter has been shown as critical to predict monkeys' choices in the deterministic PST task [1].

As a consequence of the task structure, the model optimized for the task ends up with a very high learning rate (average 0.9), and an MV mechanism which makes the exploration rate β fluctuating between 5 and 10 (thus being more exploratory during the search period than during the repetition, although the beta value of 5 denotes no pure randomness, consistently with the apparent controlled exploration that was observed in monkeys) [1].

2.3 Neural network model

The RL algorithm described above was implemented in a neural network model using the Neural Simulation Language (NSL) software. With this software, the simulated neurons' activity was represented as an average firing rate. Neurons integrated inputs over time as leaky integrators. NSL was used primarily to modularize the model's computations so as to match the involved brain structures in the monkey's task. Our model extends previous models involving the basal ganglia loop [17]. Although we do not pretend to precisely model the basal ganglia here, our model involves a cortico-basal ganglia loop which serves to select only 1 action at a given time, inspired by more detailed basal ganglia models [18]. Hence, as shown in figure 1, the basal ganglia involves an input nucleus (the caudate) – whose activity is considered in the literature as representing competing actions for action selection [19] – an output nucleus (the Substantia Nigra Pars Reticulata, Snr) with a tonic baseline of activity which exerts an inhibition on the thalamus. Thus, this system performs an action selection based on desinhibition of the chosen action in the thalamus so as to allow execution of the corresponding behavior [20]. In this model, a module representing the ACC encodes and updates action values which are sent to another module representing DLPFC which selects the action to perform. The selection process is based on the Boltzmann softmax function (equation 3) that strengthened the differences between the Q-values. The DLPFC module was part of the cortico-striatal loop involved in the final action selection.

ACC and DLPFC are assumed to be at the core of the RL and task monitoring processes needed for the PST tasks used in our study. One key point of this model is the implementation of MV in the ACC module consistently with this region's presumed role in modulating parameters of RL. In accordance with hypothesized role of mesencephalic dopamine neurons in reinforcement learning, the neural network modeled dopaminergic inputs from the ventral tegmental area (VTA) to the ACC as a reward prediction error according to equation 1. It was formalized in the model with a VTA module that projected to the ACC. Then, the neurons encoding MV and the Q-values in the ACC were updated with the RPE computed with VTA input (equation 2 and 4).

2.4 Stochastic PST

In the present study, we tried to generalize the above mentioned model by testing it on a more probabilistically rewarded PST developed by Amiez et al. [15]. In this task monkeys were facing a touch screen and had to find which one of two targets had the best rewarding rate. The reward probabilities were as follow: target 'A' was rewarded by 1.2 ml of juice 70% of the trials and by 0.4 ml the rest of the time; conversely target 'B' was rewarded 0.4 ml in 70% of the trials and 1.2 ml the last 30% trials (see table 1). A problem comprised a search – or exploratory – period and a

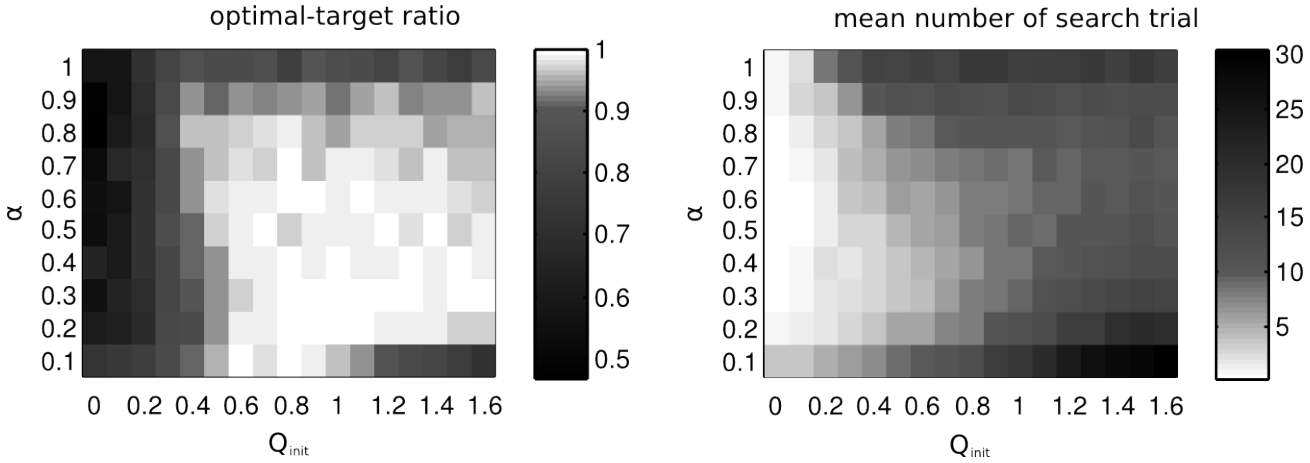

Figure 2. Relation between the initial Q-value Q_{init} and the learning rate α with exploration rate modulated by MV and the two indicators of performance. Lighter gray represent high performances.

repetition – or exploitation – period. There was no sharp change between exploration and exploitation phases but trials were categorized as repetition trials a posteriori: the monkey had to choose the same target for five consecutive trials followed by selection of the same target for the next five trials or five of the next six trials.

amount of juice dispensed as reward	Target A	Target B
1.2 ml	70%	30%
0.4 ml	30%	70%

Table 1. Reward probabilities for target A and B in stochastic task [15]. Monkeys must find which target was more rewarded. Here, target A is the optimal target.

At the end of the repetition period a new problem started. However, if after 50 trials the monkey had not entered the repetition phase a new problem started. The exact same behavioral protocol and behavioral measures were used to evaluate the model's performance in the task.

2.5 Methods

The rewards were implemented as numerical values (1.2 and 0.4) and distributed in accordance with the PST probabilities. The reward probabilities of the two targets were independent, so the choice of one target did not influence the reward probabilities of the other target in subsequent trials. We were principally interested in the behavioral results of this model i.e. the responses chosen by the model. Then, the performance of the model was defined as the mean number of search trials, and the ratio of trials where optimal target 'A' is chosen over the total number of trials. These are referred to as the indicators of performance. In the original Amiez experiment, the two monkeys found the optimal target in 98% and 94.5% of the problems. The search phase lasted in average 6.4 ± 5.6 and 5.6 ± 6.9 trials respectively.

We first tested the model with the parameter set used for the deterministic task: $\alpha = 0.9$, $\beta = 5$ and $Q_{init} = 0.4$. In order to adapt the model to the stochastic task three parameters were adjusted. The learning rate α, the exploration rate β and the initial Q-value Q_{init}. We explored α (from 0.1 to 1 with rate 0.1) and Q_{init} (from 0 to 1.2 with rate 0.1) with a β regulated by MV. In the same way the Q-values are reset at the start of each problem, MV was reset to 0.5 at the start of each problem as defined by default in the Quilodran PST. To further investigate the influence of these three parameters on model's performances, we optimized the parameter β as well (from 0 to 100 with rate 0.5).

3. Results

A naive test on the stochastic task with the optimal parameters used with the deterministic PST ($\alpha = 0.9$, $\beta = 5.2$, $Q_{init} = 0.4$) elicited a mean number of search trials of 13.3 ± 12.3 with optimal-target ratio 87% which represents poor performances compared to monkeys' performances (see methods).

The adaptation of the parameters with an exploration rate β regulated by the modulatory variable MV was a success. Roughly, the optimal α is between 0.4 and 0.6, and the optimal Q_{init} between 0.6 and 0.8 (figure 3.). With $\alpha = 0.5$ and $Q_{init} = 0.6$ the mean number of search trial is 5.5 ± 6.2 and the optimal-target ratio is 99% which is similar to the monkeys' performances.

The optimization including the exploration rate showed that parameters α and β both had relatively comparable effects across performance indicators. α and β described a rather stable performance space as long as β was not too low ($\beta > 5$) and α was between 0.2 and 0.9 (see fig. 3 A and B). In the stochastic task, the regulation of β by MV elicits values close to 10, the highest values possible for β, hence corresponding to the values where β is optimal for this PST.

Further analysis showed that the two indicators of performance had opposite tendencies with respect to the initial Q-values. As shown in figure 3 C, low initial action values elicited few optimal-target choices but short search phases. Conversely, high initial action values induced a high percentage of optimal response choices but a too lengthy search period. However, an average initial Q-value can balance these two effects so to have a relatively good performance with the two

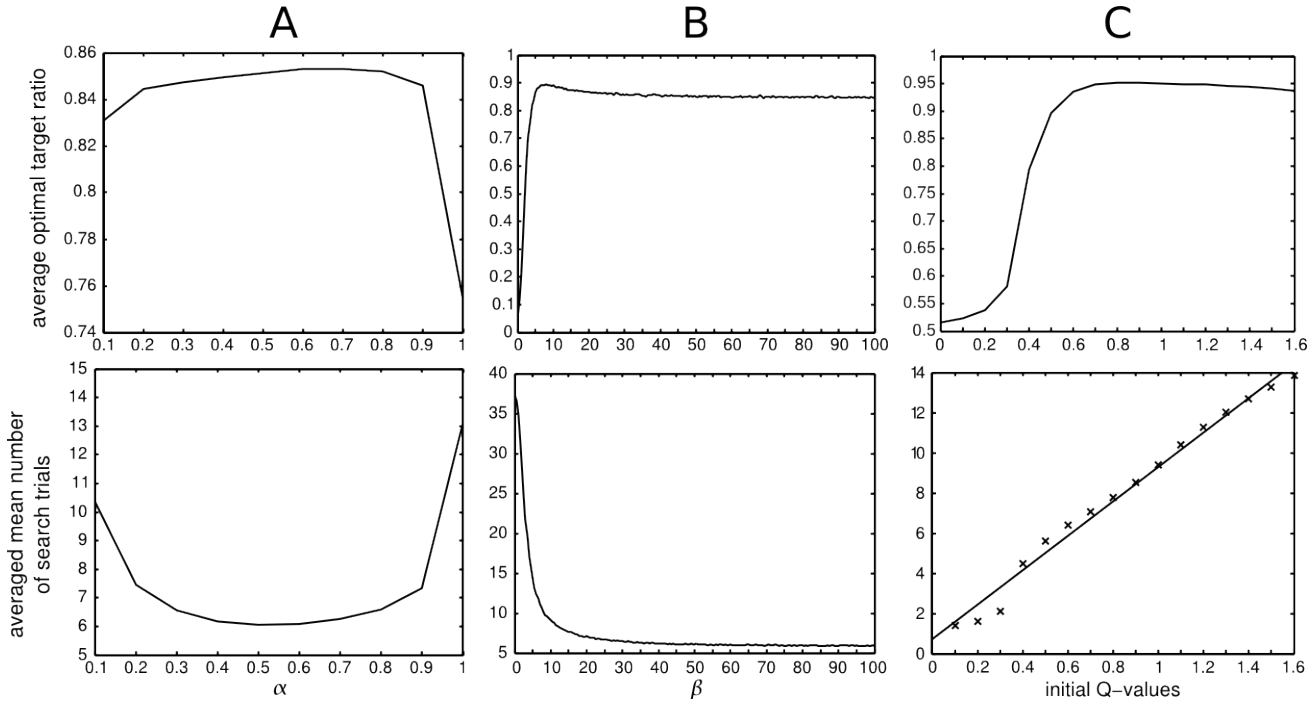

Figure 3. Averaged mean number of search trials and optimal-target ratio as function of the three parameters alpha, beta and initial Q-value. **(A)** Averaged indicators of performance as function of alpha. The model's performances were particularly low for alpha = 1. The model did not learn for the feedback it received. **(B)** Averaged indicators of performance as function of beta. Apart from the smallest value of beta (beta < 15), the performances were very stable and satisfactory. **(C)** Averaged indicators of performance as function of the initial Q-value. Initial Q-values had opposite effects on the indicators of performance. The optimal-target ratio reaches satisfactory values for an initial Q-value of 0.6. The averaged mean search period is correlated with the initial Q-values with correlation coefficient 0.99 and p-value < 10^{-14}.

indicators. Further analyzes revealed that the initial Q-value is highly correlated to the search period length (correlation coefficient is 0.99 with p-value < 10-14 ; fig. 3 C).

4. Discussion

The main findings of this study were that (1) the modulatory variable (MV) successfully adjusted the exploration rate to adapt it to the stochastic PST and (2) the value assigned to the Q-values at each problem start had a strong influence on the model performance.

Interestingly, and as predicted with the model, the optimal learning rate α for this task was lower than the optimal α for the deterministic PST. In the deterministic task, α was rather high (0.9 on average) reflecting the deterministic reward distribution. Whereas, in accordance with our hypothesis, optimal α was lower (between 0.4 and 0.6) with Amiez PST because of the stochasticity of the task.

As for the optimization of β, it is remarkable that the more exploitative the better the performances (low β induced a too lengthy search period because the model was too exploratory). Unlike our initial hypothesis, this is in part due to the nature of the PST in which only 2 targets are available, decreasing the search space, so the optimal strategy is clearly exploitative. In accordance with this finding, β was adjusted with MV to its highest possible value (around 10). The optimized model with a fixed exploration rate beta reached a nearly optimal behavior. In contrast, the model with a dynamic exploration rate revealed good performance, although not optimal, but nevertheless closer to monkeys' performance in this task. This suggests that such brain inspired adaptive mechanisms are not optimal but might have been selected through evolution because they can produce a good performance in different conditions.

Optimization showed the importance of the reset of Q-values when a new problem started. The initial Q-value should be no smaller than the smallest possible reward (0.4), otherwise the model persists in selecting the target it chose at the first trial of a problem. Hence, with low initial Q-values the strategy was clearly not exploratory and the optimal target was chosen only half of the time. However we observed high search phase lengths when the Q-values were reset to high values, especially when higher than the highest possible reward (1.2). Because the action values were high, they required more trials to converge especially when the learning rate was low. We can consider that initial Q-values between the lowest and highest reward possible have more chance to elicit good performance than the rest of the parameter space. Interestingly, electrophysiological data from the ACC recorded during the stochastic PST showed that neurons in this region encode the 'task value', i.e. the expected value of the most rewarded option (0.96 = 0.7*1.2 + 0.3 *0.4) [15]. The expected value indeed falls between the range of values to which the model should be reset for optimal performance. These data reinforce the idea that ACC extracts information from the environment to regulate the RL parameters, but also that ACC sets the action values used as reference to initiate exploratory

behavior.

5. Conclusion

The present work focused on the adaptation of a reinforcement learning model to a probabilistically rewarded task. The exploration of three parameters unveiled the key role of the initial action values on the performance of the model, but also the possible mechanisms by which the meta-parameters of reinforcement learning could be regulated. Further work could investigate the task value as a possible cue for action value resetting. Other perspectives include the implementation of this RL framework in a more realistic neuronal model. Indeed, modeling connectivity and interaction between ACC and DLPFC could yield interesting insight on the mechanisms of RL in the prefrontal cortex.

References

[1] M. Khamassi, R. Quilodran, P. Enel, E. Procyk, and P.F. Dominey, A computational model of integration between reinforcement learning and task monitoring in the prefrontal cortex, *Simulation of Adaptive Behavior 2010, in press*.

[2] K. Doya, Metalearning and neuromodulation., *Neural networks : the official journal of the International Neural Network Society*, vol. 15, 2002, pp. 495-506.

[3] R.S. Sutton and A.G. Barto, *Reinforcement learning: an introduction* (Cambridge, MIT press, 1998).

[4] P.R. Montague, S.E. Hyman, and J.D. Cohen, Computational roles for dopamine in behavioural control., *Nature*, vol. 431, 2004, pp. 760-7.

[5] W. Schultz, Multiple reward signals in the brain, *Nature Reviews Neuroscience*, vol. 1, 2000, p. 199–207.

[6] G. Aston-Jones and J.D. Cohen, An integrative theory of locus coeruleus-norepinephrine function: Adaptive Gain and Optimal Performance, *Annual Review of Neuroscience*, vol. 28, 2005, pp. 403-450.

[7] M. Matsumoto, K. Matsumoto, H. Abe, and K. Tanaka, Medial prefrontal cell activity signaling prediction errors of action values., *Nature neuroscience*, vol. 10, 2007, pp. 647-56.

[8] E. Procyk, Y.L. Tanaka, and J.P. Joseph, Anterior cingulate activity during routine and non-routine sequential behaviors in macaques., *Nature neuroscience*, vol. 3, 2000, pp. 502-8.

[9] H. Seo and D. Lee, Cortical mechanisms for reinforcement learning in competitive games., *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 363, 2008, pp. 3845-57.

[10] R. Quilodran, M. Rothé, and E. Procyk, Behavioral shifts and action valuation in the anterior cingulate cortex., *Neuron*, vol. 57, 2008, pp. 314-25.

[11] T.E. Behrens, M.W. Woolrich, M.E. Walton, and M.F. Rushworth, Learning the value of information in an uncertain world., *Nature neuroscience*, vol. 10, 2007, pp. 1214-21.

[12] M. Khamassi, L. Lachèze, B. Girard, A. Berthoz, and A. Guillot, Actor-Critic Models of Reinforcement Learning in the Basal Ganglia: From Natural to Artificial Rats, *Adaptive Behavior*, vol. 13, 2005, pp. 131-148.

[13] P. Dayan and N.D. Daw, Decision theory, reinforcement learning, and the brain., *Cognitive, affective & behavioral neuroscience*, vol. 8, 2008, pp. 429-53.

[14] K. Doya, Modulators of decision making., *Nature neuroscience*, vol. 11, 2008, pp. 410-6.

[15] C. Amiez, J.P. Joseph, and E. Procyk, Reward encoding in the monkey anterior cingulate cortex., *Cerebral cortex*, vol. 16, 2006, pp. 1040-55.

[16] S. Dehaene, M. Kerszberg, and J. Changeux, A neuronal model of a global workspace in effortful cognitive tasks, *Proceedings of the National Academy of Sciences of the United States of America*, 1998.

[17] P.F. Dominey, M. Arbib, and J. Joseph, A model of corticostriatal plasticity for learning oculomotor associations and sequences, *Journal of Cognitive Neuroscience*, 1995.

[18] B. Girard, N. Tabareau, Q.C. Pham, a. Berthoz, and J. Slotine, Where neuroscience and dynamic system theory meet autonomous robotics: a contracting basal ganglia model for action selection., *Neural networks : the official journal of the International Neural Network Society*, vol. 21, 2008, pp. 628-41.

[19] K.N. Gurney, T.J. Prescott, and P. Redgrave, A computational model of action selection in the basal ganglia: I. A new functional anatomy, *Biological Cybernetics*, vol. 84, 2001, pp. 401-410.

[20] G. Chevalier, and M. Deniau, Disinhibition as a basic process of striatal function, *Trends in Neurosciences*, vol. 13, 1990, pp. 277–280.