Role of the frontal cortex in solving the exploration-exploitation trade-off

Mehdi Khamassi, René Quilodran, Pierre Enel, Peter Dominey, Emmanuel Procyk

ROLE OF THE FRONTAL CORTEX IN SOLVING THE EXPLORATION-EXPLOITATION TRADE-OFF: MODEL-BASED ANALYSIS OF SINGLE-UNIT RECORDINGS

Mehdi Khamassi, René Quilodran, Pierre Enel, Peter F. Dominey, Emmanuel Procyk
(1) Inserm U846, Bron, France ; (2) Stem Cell and Brain Research Institute, Bron, France ; (3) Université de Lyon, Université Lyon I, Lyon, France.
INSERM U846, 18 avenue du Doyen Lépine, 69500 Bron, France
correspondence: mehdi.khamassi@inserm.fr

ABSTRACT
While many electrophysiological recordings and computational modeling work have investigated the role of the frontal cortex in reinforcement learning (learning by trial-and-error to adapt action values), it is not yet clear how the brain flexibly regulates in a task-appropriate way crucial parameters of learning such as the learning rate and the exploration rate. In a previous work, we proposed a computational model based on the meta-learning theoretical framework where the frontal cortex extracts feedback signals 1) to update action values based on a reward prediction error; 2) to estimate the level of exploration based on the current reward average; 3) to select action based on this exploration rate. This model helped us draw a set of experimental predictions. Here we show a model-based analysis of single-unit recordings in the monkey prefrontal cortex so as to test these predictions. We found neural subpopulations activities consistent with these three functions. We also found global properties of the recorded neural ensemble – such as variations in spatial selectivity – which were predicted by our model. Such an approach, gathering computational modeling and neurophysiology, can help understand complex activities of neural ensembles related to decision making.

KEY WORDS
Modeling; Neurophysiology; Decision-Making; Frontal Cortex; Reinforcement Learning; Meta-Learning

1. Introduction

Previous results on neural bases of decision making in the frontal cortex showed crucial mechanisms that could participate both to reinforcement learning processes [1] and to the auto-regulation of exploration-exploitation behaviors [2]. 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 [3, 4, 5]. 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 modeled with the reinforcement learning (RL) framework [1, 6, 7]. However, it is not clear how these mechanisms integrate within these neural structures, and interact to produce coherent decision-making under explore-exploit trade-off.
In a previous work [8], we proposed a computational model where ACC filters dopaminergic input signals – assumed to convey a reward prediction error [9] – and uses this signal both to update action values and to regulate the level of exploration. Inspired by the meta-learning theoretical framework [10], such a regulation consisted in estimating the current reward average and using this information to tune the β parameter called the exploration rate. This model led to a series of experimental predictions that we test here by presenting a model-based analysis of single-unit recordings in monkey prefrontal cortex. We find activities consistent with our predictions, revealing separate neural ensembles in ACC encoding action values and exploration rate, and integration of these information in DLPFC to enable action selection under varying exploration-exploitation trade-off. Global properties of the recorded neural ensemble – such as variations in spatial selectivity – are also consistent with our model predictions.

2. Previous model

Figure 1 summarizes the model previously developed in [8] to reproduce the task used in [7]. In the model, ACC and DLPFC contain a 3*3 grid representing different areas on a touch screen. At each trial, four targets are presented on the screen. Following the RL framework [11], the ACC learns the action value *Q(a)* associated to pressing each possible target *a*. After pressing a target, the action value is compared with the presence/absence of reward so as to compute a Reward Prediction Error (RPE):

$$RPE \leftarrow Q(a) - r \quad (1)$$

where r is the reward.
The value of the performed action is updated according to the following equation:

$$Q(a) = Q(a) + (\alpha . RPE) \quad (2)$$

where α is the learning rate ($0 < \alpha < 1$).

Action values are transmitted to the DLPFC which selects the next action to perform based on the Boltzman softmax rule:

$$P(target_i) = \frac{\exp(\beta . Q(target_i))}{\sum_j \exp(\beta . Q(target_j))} \quad (3)$$

where β regulates the exploration rate ($0 < \beta$). A small β leads to a very similar probability for each action and thus to an exploratory behavior. A high β increases the difference between the highest action value and the others, and thus produces an exploitative behavior.

In parallel, a modulatory variable (MV) is computed so as to dynamically regulate the exploration rate β. Inspired by the meta-learning theoretical framework [9], the idea is to increase exploration when the average reward decreases, and to increase exploitation when it increases:

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

with $\alpha_+ = 2{,}5$ and $\alpha_- = 0{,}25$ to tackle sharp changes between exploration and exploitation phases as observed in monkey behavior in the task used by [7].
MV is used to modulate the exploration rate β within the DLPFC:

$$\beta = \frac{\omega_1}{\left(1 + \exp(\omega_2 . [1 - MV] + \omega_3)\right)} \quad (5)$$

with $\omega_1 = 10$, $\omega_2 = -6$ and $\omega_3 = 1$, which has a sigmoid function that produces a low β when *MV* is high (exploration) and a high β when *MV* is low (exploitation).

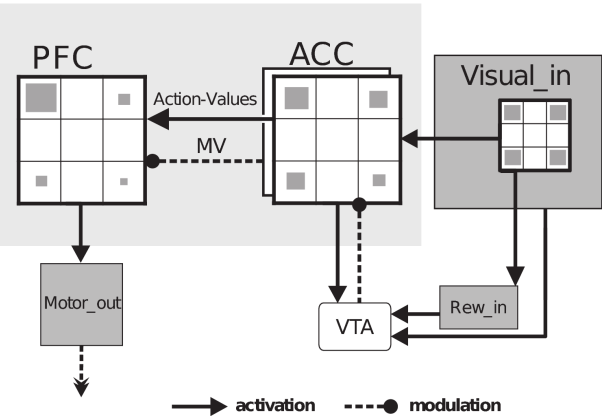

Figure 1. A simplified representation of the computational model developed by [8]. Visual input (4 square targets presented on a touch screen; or a circle in a center indicating a problem-change; or a triangle in the top center indicating a reward) is sent to the Anterior Cingulate Cortex (ACC) and the dopaminergic system (Ventral Tegmental Area; VTA). The VTA sends a Reward Prediction Error signal to ACC which updates action values associated to pressing each possible target. Action values are sent to the Dorsolateral Prefrontal Cortex (DLPFC) which selects an action based on the current exploratory rate β. In parallel, ACC computes a Modulatory Variable (MV) which represents the current reward average. MV is used to modulate the β value in DLPFC so that the latters explores more after errors, and exploits more after correct trials.

This model enabled to draw a set of experimental predictions [8]. Here we test two of these predictions by recording ACC and DLPFC neuronal activities in the task employed by [7]:

1. There should exist *MV* neurons within the ACC – with an increase of activity after each error trial and a decrease of activity after correct trials.
2. The effect of *MV* on the exploration rate β within DLPFC should produce a higher contrast between neuronal activities representing action values during exploitation phases (as symbolized by square surfaces in figure 1), and a lower contrast during exploration phases.

3. Model-based analysis of brain data

We simulated this algorithmic model on the task used in [7] so as to test for correlations between model variables and neuronal activities recorded in ACC and DLPFC in this task.

3.1 Task

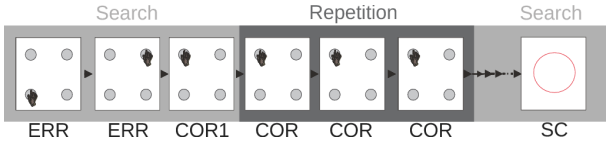

Figure 2. Problem Solving Task employed in [7] where monkey prefrontal cortical neurons analyzed here were recorded. A typical problem starts with a *Search* phase where the animal searches for the rewarding target among four presented on a touch screen. The monkey makes a series of error trials (*ERR*) until it finds the correct target (first correct trial, *COR1*). Then a repetition phase is imposed where the animal needs to repeat the same choice for 3 to 11 trials. Finally, a Signal to Change (*SC*) indicating that new problem starts, meaning that the correct target location will be changed in 90% of the cases.

3.2 Fitting monkey behaviour

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 (so that the model learns based on the same experience as the monkey). 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 monkeys choices. For each behavioral session, we optimize the model by finding the set of parameters that provides the highest likelihood of fitting monkeys choices. We take into account individual spatial biases of each monkey by initializing action values associated to each target based on target preferences measured during the previous session for the same monkey.

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.3 shows simulation results on a

sample of 100 trials for 1 monkey (Monkey M, session MB5_2782).

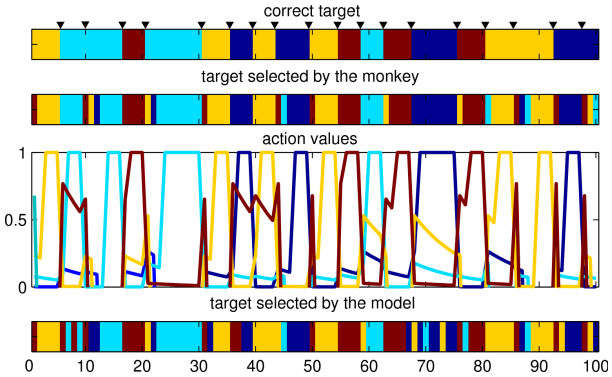

Figure 3. Simulation of the reinforcement learning model on 100 trials. Each color is associated with a different target. The top line denotes the problem sequence experienced by both the monkey and the model. Black triangle indicate the presentation of a Signal to Change (SC). The second line shows the monkeys 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 based on individual spatial preferences, 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.

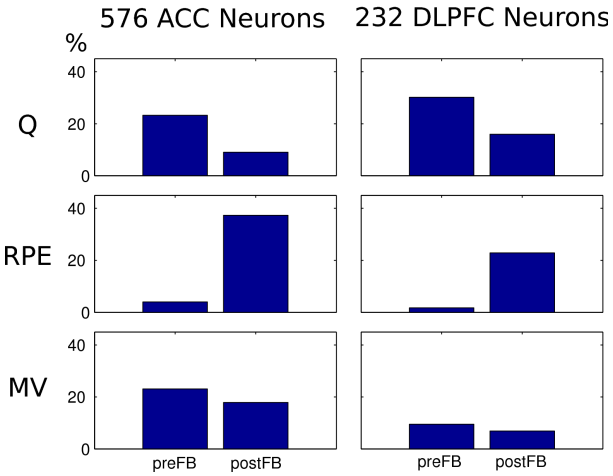

Figure 4. Proportions of ACC and DLPFC neurons with an activity correlated with one of the three model variables (Q, RPE, MV), either during the pre-feedback period (preFB) or during the post-feedback period (postFB).

3.3 Correlations between model variables and neural activities

Once the model is set to fit behavioral data, we can use variables in the model as regressors to test for correlations with single-unit activity recorded in the monkey anterior cingulate cortex (ACC) and dorsolateral prefrontal cortex (DLPFC) in the same task. In this section, we present such model-based analysis. We used a multiple regression analysis to test possible correlations between each neuron's activity measured in its preferred 500ms-period within the trial and the three model variables: action values Q, reward prediction error RPE, modulatory variable MV. As a control, we combined this method with a bootstrap: we randomized 1000 times the trials order for the model's variables and tested if the considered neuron's activity was still correlated. A neuron's activity was considered as significantly correlated with one of the model's variables if the strength of correlation was higher than strengths of 950 random samples.

Figure 4 summarizes the proportions of cells in ACC and DLPFC that are correlated with a model variable. Consistently with previous reports on RPE encoding in the ACC, we found a high proportion of ACC neurons correlated with the Reward Prediction Error (RPE) in the post-feedback period. In addition, we also found neurons correlated with one the four action values both in the ACC and DLPFC. The proportion of action value neurons is higher in the DLPFC than in the ACC, consistently with the former's supposed role in action selection (77% for DLPFC versus 48% for ACC; chi-square test, 1 df, T=27.7228088, p = 7.2370e-8).

In addition, and as predicted by our model, there are neurons in ACC correlated with the Modulatory Variable (MV), and their proportion is higher than in DLPFC (52% in ACC versus 28% in DLPFC; chiquare, 1 df, T = 19.6090119, p = 4.9732e-06).

Figure 5 shows examples of neural activities correlated with some model variables.

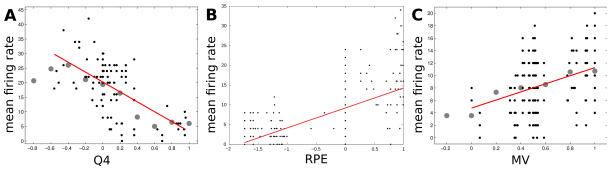

Figure 5. Examples of neurons with activity correlated either with the action value of the fourth target (*Q4*), the reward prediction error (*RPE*), the modulatory variable (*MV*).

3.3 Global properties of recorded neural ensembles

In addition to individual neuron activities which could encode information similar to separate computations carried by our model, we predicted that there should be properties at the neural ensemble level due to the variation in exploration rate. More precisely, the model predicted that there should be an increase in contrast between neural activities representing different target values during the repetition period of this task due to the exploitative mode triggered by the increase in the β value in equations 4-5 of the model.

We observe such phenomenon materialized by an increase of spatial selectivity in DLPFC during the repetition period. Such increase is statistically significant, when looking at all DLPFC cells, but also when considering only action value cells, as the model predicted: among 85 DLPFC action value neurons, the average spatial selectivity of 62 (73%) cells having a significant spatial selectivity (either during SEARCH or REPETITION) significantly increases during REPETITION (the mean index of spatial selectivity variation between SEARCH and REPETITION equals 0,0993, which has a median different from 0 [signrank p = 0.0016] and a mean different from zero [t-test p 9.4003e-04]). Thus, DLPFC's increase in spatial selectivity during the repetition period can be in part due to an increase in the contrast between activities of action value neurons, as postulated by the model.

4. Conclusion

Acumulating evidence suggest that the frontal cortex could contribute to flexible behaviors and to learning based on feedback obtained from the environment [1,3,5]. Recent electrophysiological findings suggest a specialization of the frontal cortex where the Anterior Cingulate Cortex (ACC) monitores performance to modulate decision-making in the Dorsolateral Prefrontal Cortex (DLPFC) [2,6,12]. Several computational models have tackled this specialization, either by considering that ACC monitores conflict between competing actions to increase the gain in the DLPFC [13], or proposing that ACC computes the current error-likelihood [4]. Our model proposes a more general principle to explain ACC function in terms of meta-learning [10]. The ACC could be generally involved in monitoring performance relative to the current environment's properties to tune parameters of reinforcement learning and action selection. Consistently with this proposition, Rushworth and colleagues have recently shown that the ACC in humans is important to track the environment's volatility (variations in the reward rate) and adapt subsequent behavior [14].

In this paper, we used our computational model to analyze single-unit recordings in the monkey ACC and DLPFC. We were able to formally relate activities of subpopulations of neurons to different model variables, thus confirming their possible implications in different computations employed in the model (reinforcement learning based on action values and reward prediction errors; meta-regulation of exploration based on MV). In addition to information measured at the cellular level, we found global properties of neural ensembles that were predicted by the model: the model predicted there should be an increase in spatial selectivity in DLPFC during exploitation phases, due to the increased contrast between action values induced by a high β in equation 3. This validates a possible role for the ACC-DLPFC system in dynamic regulation of the exploration rate. Further investigations will be required to test other predictions formulated with our model in the same task, and to see whether our model makes verified predictions in other protocols.

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 feedbacks (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.

Acknowledgements

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

References

[1] Barraclough, D., Conroy, M., Lee, D., Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience, 7*(4), 2004, 404-410.

[2] Procyk, E., Tanaka, Y., Joseph, J., Anterior cingulate activity during routine and non-routine sequential behaviors in macaques. *Nature Neuroscience, 3*(5), 2000, 502-508.

[3] Aston-Jones, G., Cohen, J., Adaptive gain and the role of the locus coeruleus-norepinephrine system in optimal performance. *Journal of Computational Neurology, 493*(1), 2005, 99-110.

[4] Brown, J., Braver, T., Learned predictions of error likelihood in the anterior cingulate cortex. *Science, 307(5712)*, 2005, 1118-1121.

[5] Dosenbach, N., 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.

[6] Matsumoto, M., Matsumoto, K., Abe, H., Tanaka, K., Medial prefrontal cell activity signaling prediction errors of action values. *Nature Neuroscience, 10*, 2007, 647-656.

[7] Quilodran, R., Rothe, M., Procyk, E., Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron, 57(2)*, 2008, 314-325.

[8] Khamassi, M., Quilodran, R., Enel, P., Procyk, E., Dominey, P., A computational model of integration between reinforcement learning and task monitoring in the prefrontal cortex. Proc. 11th Int. Conf. on Simulation of Adaptive Behavior (SAB2010), Paris, France, in press.

[9] Schultz, W., Dayan, P., Montague, P., A neural substrate of prediction and reward. *Science, 275(5306)*, 1997, 1593-1539.

[10] Doya, K., Metalearning and neuromodulation. *Neural Networks, 15(4-6)*, 2002, 495-506.

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

[12] Seo, H., Lee, D., Behavioral and Neural Changes after Gains and Losses of Conditioned Reinforcers. *Journal of Neuroscience, 29*(11), 2009, 3627–3641.

[13] Botvinick, M., Braver, T., Barch, D., Carter, C., Cohen, J., Conflict monitoring and cognitive control. *Psychological Review, 108*, 2001, 624-652.

[14] Rushworth, M., Behrens, T., Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience, 11*(4), 2008, 389-397.