HAL
open science

Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters

Mehdi Khamassi, Pierre Enel, Peter Ford Dominey, Emmanuel Procyk

To cite this version:

Mehdi Khamassi, Pierre Enel, Peter Ford Dominey, Emmanuel Procyk. Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. Progress in Brain Research. Decision Making Neural and Behavioural Approaches, 2013, 10.1016/B978-0-444-62604-2.00022-8 . hal-01628829

HAL Id: hal-01628829

https://hal.science/hal-01628829

Submitted on 4 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters

Mehdi Khamassi[1,2], Pierre Enel[1], Peter Ford Dominey[1], Emmanuel Procyk[1]

[1] INSERM U846, Stem Cell and Brain Research Institute, 69500 Bron, France; Université de Lyon, Lyon 1, UMR-S 846, 69003 Lyon, France

[2] Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie-Paris 6, 4 place Jussieu, 75252 Paris Cedex 05, France; CNRS UMR 7222, 75005 Paris, France

ABSTRACT: Converging evidence suggest that the medial prefrontal cortex (MPFC) is involved in feedback categorization, performance monitoring and task monitoring, and may contribute to the online regulation of reinforcement learning (RL) parameters that would affect decision-making processes in the lateral prefrontal cortex (LPFC). Previous neurophysiological experiments have shown MPFC activities encoding error likelihood, uncertainty, reward volatility as well as neural responses categorizing different types of feedback, for instance distinguishing between choice errors and execution errors. Rushworth and colleagues have proposed that the involvement of MPFC in tracking the volatility of the task could contribute to the regulation of one of RL parameters called the learning rate. We extend this hypothesis by proposing that MPFC could contribute to the regulation of other RL parameters such as the exploration rate and default action values in case of task shifts. Here we analyze the sensitivity to RL parameters of behavioral performance in two monkey decision-making tasks, one with a deterministic reward schedule and the other with a stochastic one. We show that there exist optimal parameters values specific to each of these tasks, that need to be found for optimal performance and that are usually hand-tuned in computational models. In contrast, automatic online regulation of these parameters using some heuristics can help producing a good, although non-optimal, behavioral performance in each task. We finally describe our computational model of MPFC-LPFC interaction used for online regulation of the exploration rate and its application to a human-robot interaction scenario. There, unexpected uncertainties are introduced by the human introducing cued task changes or by cheating. The model enables the robot to autonomously learn to reset exploration in response to such uncertain cues and events. The combined results provide concrete evidence specifying how prefrontal cortical subregions may cooperate to regulate RL parameters. It also shows how such neurophysiologically inspired mechanisms can control advanced robots in the real-world. Finally, the model's learning mechanisms that were challenged in the last robotic scenario provide testable predictions on the way monkeys may learn the structure of the task during the pre-training phase of the previous laboratory experiments.

INTRODUCTION

The Reinforcement Learning (RL) theory has been widely and successfully used to describe neural mechanisms of decision-making based on action valuation, and on learning of action values based on reward prediction and reward prediction errors (Houk et al., 1995; Sutton and Barto, 1998). Its extensive use in the computational neuroscience literature is grounded on the observation that dopaminergic neurons respond according to a reward prediction error (Schultz et al., 1997), that dopamine strongly innervates the prefrontal cortex and striatum and there modifies synaptic plasticity (Humphries et al., 2010; Reynolds et al., 2001), and that prefrontal cortical and striatal

neurons encode a variety of RL-consistent information (Daw et al., 2006; Khamassi et al., 2008; Samejima et al., 2005; Sul et al., 2010).

However, RL models rely on crucial parameters (e.g. learning rate, exploration rate, temporal discount factor) that need to be dynamically tuned to cope with variations in the environment. In most computational neuroscience work, experimenters explore the parameters space and find a set of parameters which work for a specific task (Chavarriaga et al., 2005; Daw et al., 2005; Frank, 2005; Khamassi et al., 2005). In contrast, animals are able to adjust their behavior to many different situations, show gradual adjustment of their learning characteristics along familiarization with the task (Luksys et al., 2009), and are able to re-explore their environment in response to drastic changes. If one postulates that the brain implements RL-like decision-making mechanisms, one needs to understand how the brain regulates such mechanisms, in other words how it "tunes parameters". Kenji Doya has formalized such principles of regulation of RL parameters in a Meta-Learning theoretical framework, proposing computational solutions to learn which set of parameters is appropriate to control learning during a given task (Doya, 2002). Here we argue that accumulating evidence suggest that the medial prefrontal cortex might play a key role in detecting task changes and variations of the agent's own performance and in consequently adjusting parameters of learning. We illustrate the need for dynamically adjusting RL parameters in two decision-making tasks where we previously recorded monkey MPFC activity (Amiez et al., 2006; Quilodran et al., 2008) by performing simple simulations of a classic RL algorithm that show that different values of the parameters are required to produce optimal performance in different phases of the tasks. Then we present the computational model that we have proposed (Khamassi et al., 2011) to describe how MPFC may interact with LPFC to regulate decision-making based on the history of feedback and thus based on the RL parameters that appear to be required in the present context. We simulate this model in the two monkey decision-making tasks to extract concrete predictions on expected simultaneous MPFC and LPFC neural activities. We finish by illustrating the functioning of the model in a human-robot interaction game to show its performance when coping with real-world uncertainties and to make further predictions on how monkeys may learn the structure of the studied decision-making tasks during the pre-training stage.

THE MPFC AS A REGULATOR OF DECISION-MAKING

Prefrontal cortical mechanisms underlying the regulation of decision-making have been largely studied in terms of "cognitive control" (Badre and Wagner, 2004; Botvinick et al., 2001; Mars et al., 2011; Miller and Cohen, 2001), a high-level of behavioral regulation in new and challenging situations where behavioral routines need to be modified or reorganized, and is hypothesized to involve interactions between subdivisions of the prefrontal cortex (PFC), especially the medial and lateral PFC.

Within the medial frontal cortex, the anterior cingulate cortex (ACC), and in particular area 24c, has an intermediate position between limbic, prefrontal, and premotor systems (Amiez et al., 2005a; Paus et al., 2001). ACC neuronal activity tracks task events and encodes reinforcement-related information (Amiez et al., 2005a; Procyk et al., 2001). Muscimol injections in dorsal ACC induce strong deficits in finding the best behavioral option in a probabilistic learning task and in shifting responses based on reward changes (Amiez et al., 2006; Shima and Tanji, 1998). Dorsal ACC lesions also induce failures in integrating reinforcement history to guide future choices (Kennerley et al., 2006). These data converge toward describing a major role of ACC in integrating reward information over time, which is confirmed by single-unit recordings (Seo and Lee, 2007), and thereby in decision-making based on action-reward associations. This function contrasts with that of the orbitofrontal cortex, which is necessary for stimulus-reward associations (Rudebeck et al., 2008).

In addition, the ACC certainly has a related function in detecting and valuing unexpected but behaviorally relevant events. This notably includes the presence or absence of reward outcomes and

failure in action production, and has been largely studied using event-related potentials in humans and unit recordings in monkeys. The modulation of phasic ACC signals by prediction errors, as defined in the reinforcement learning framework, supports the existence of a key functional relationship with the dopaminergic system (Amiez et al., 2005b; Holroyd and Coles, 2002). In the dopamine system, the same cells encode positive and negative reward prediction error (RPE) by a phasic increase and a decrease in firing, respectively (Bayer and Glimcher, 2005; Morris et al., 2006; Schultz et al., 1997). By contrast, in the ACC, different populations of cells encode positive and negative prediction errors, and both types of error result in an increase in firing (Matsumoto et al., 2007; Quilodran et al., 2008; Sallet et al., 2007). Moreover, ACC neurons are able to discriminate choice errors (choice-related RPE) from execution errors (motor-related RPE, e.g. break of eye fixation; Quilodran et al., 2008). These two error types should be treated differently because they lead to different post-error adaptations. This suggests that while the dopaminergic RPE signal could be directly used for adapting action values, ACC RPE signals also relate to a higher level of abstraction of information, like *feedback categorization*. In line with this, Alexander and Brown recently proposed that ACC signals unexpected non-occurrences of predicted outcomes (Alexander and Brown, 2011). Although their model cannot account for ACC correlates of positive prediction errors – putatively signaling unexpected occurrences of non-predicted outcomes – (Matsumoto et al., 2007; Quilodran et al., 2008) nor for the implication of ACC in action valuation (MacDonald et al., 2000; Kennerley et al., 2006; Rushworth and Behrens, 2008; Seo and Lee, 2008), their model elegantly explains a large amount of reported ACC post-feedback activity and highlights its role in detecting relevant events for behavioral regulation.

A third important aspect of ACC function was revealed by the discovery of changes in neural activity between exploratory and exploitative trials (Procyk et al., 2000; Quilodran et al., 2008), or between volatile and stable rewarding schedules (Behrens et al., 2007). Kolling et al. (2012) have recently found that ACC encodes the average value of the foraging environment. This suggests a more general involvement of ACC in translating results of performance monitoring and task monitoring into a regulatory level.

Koechlin and colleagues have proposed that ACC might regulate the level or rate of cognitive control in LPFC as a function of motivation based on action cost-benefit estimations (Kouneiher et al., 2009). The temporality of activations of the two structures appears consistent with the hypothesis that at times of instructive events performance monitoring (mainly ACC) is followed by adjustment in control and selection (in LFPC). Temporality was studied both by unit recordings in non-human primates (Johnston et al., 2007), and by EEG studies in human (Silton et al., 2010). The former study showed that the effect of task switching appear earlier in ACC than in LFPC (Johnston et al., 2007). The EEG study revealed phasic and early non-selective activations in ACC as opposed to a late LPFC activation correlated with performance. However, Silton and colleagues underlined that when task relevant information is taken into account, late ACC activity appears to be influenced by earlier activation in LPFC (Silton et al., 2010). Data from our laboratory show that after relevant feedback leading to adaptation, advanced activation is seen in ACC before activation of LPFC at the population level for high gamma power of LFP (Rothé et al., 2011).

Rushworth and colleagues have recently highlighted the presence at the level of ACC activity of information relevant to the modulation of one of the reinforcement learning parameters: the learning rate α (Behrens et al., 2007). Their study is grounded on theoretical accounts suggesting that feedback information from the environment does not always have the same uncertainty and will be treated differently dependent on whether the environment is stable or unstable. In unstable and constantly changing ('volatile') environments, rapid behavioral adaptation is required in response to new outcomes, and so a higher learning rate is required. In contrast, the more stable the environment the less reward prediction errors should influence future actions. In the latter situation, more weight should be attributed to previous outcomes and the learning rate should remain small.

These crucial variables of volatility and uncertainty correlate with the BOLD response in the ACC at the time of outcomes (Behrens et al., 2007). Experimental controls in these studies allowed these signals influencing the learning rate to be identified independently from signals representing the prediction error. This suggests that variations in ACC activity reflect the flexible adaptation of parameter α (i.e. the learning rate) based on task requirements, and that previous reports of ACC activity encoding reward prediction errors might be a consequence of such a meta-learning function (Matsumoto et al., 2007; Quilodran et al., 2008). In line with this interpretation, as we mentioned above, the RPE-like activities that we have recorded in the ACC appear to participate to a feedback categorization process with a high-level of abstraction, and thus encode specific events that are relevant for various adaptations in the context of a task (Amiez et al., 2005; Quilodran et al., 2008).

Here we will argue that observed changes between two distinct modes of activity in ACC between exploratory and exploitative trials (Procyk et al., 2000; Quilodran et al., 2008) can be modeled by a mechanism regulating the exploration parameter β. As we will see, this points out to a general role of ACC in dynamically regulating various reinforcement learning parameters based on task events and measures of the agent's own performance.

COMPUTATIONAL PRINCIPLES OF META-LEARNING

Reinforcement Learning (RL) is a research field within computer science that studies how an agent can appropriately adapt its behavioral policy so as to reach a particular goal in a given environment (Sutton and Barto, 1998). Here, we assume this goal to be maximizing the amount of reward obtained by the agent. RL methods rely on Markov Decision Processes. This is a mathematical framework for studying decision-making which supposes that the agent is situated in a probabilistic or deterministic environment, that it has a certain representation of its state (e.g. its location in the environment, the presence of stimuli or rewards, its motivational state), and that future states depend on the performance of particular actions in the current state. Thus the objective of the agent is to learn the value associated to performance of each possible action *a* in each possible state *s* in terms of the amount of reward that they provide. Such state-action value or *quality* is noted *Q(s,a)*. In a popular class of RL algorithms called Temporal-Difference Learning, which has shown strong resemblance with dopaminergic signaling (Schultz et al., 1997), the agent iteratively performs actions and updates action values based on a Reward-Prediction Error:

$$\delta_t = r_t + \gamma.\max_a Q(s_t, a) - Q(s_{t-1}, a_{t-1}) \quad (1)$$

where r_t is the reward obtained at time *t*, $Q(s_{t-1},a_{t-1})$ is the value of action a_{t-1} performed in state s_{t-1} at time *t-1* which lead to the current state s_t, and "γ.max $Q(s_t, a)$" is the quality of the new state s_t, that is, the maximal value that can be expected from performing any action *a*. The latter term is weighted by a parameter *γ* (0 ≤ γ < 1) called the discount factor, which gives the temporal horizon of reward expectations. If *γ* is tuned to a high value, the agent has a behavior oriented towards long-term rewards. If *γ* is tuned to a value close to 0, the agent focuses on immediate rewards (Tanaka et al., 2004; Schweighofer et al., 2007).

The reward prediction error δ_t constitutes a reinforcement signal based on the unpredictability of rewards (*e.g.* unpredicted reward will lead to a positive reward prediction error and thus to a reinforcement; Sutton and Barto, 1998). Action values are then updated with this reward prediction error term:

$$Q(a_{t-1}, s_{t-1}) \leftarrow Q(a_{t-1}, s_{t-1}) + \alpha.\delta_t \quad (2)$$

where *α* is a second parameter called the learning rate (0 ≤ *α* ≤ 1). Tuning *α* will determine whether new reinforcements will drastically change the representation of action values (case where

α is close to 1), or if instead an action should be repeated several times before its value is significantly changed (case where α is close to zero).

Once action values are updated, an action selection process enables a certain exploration-exploitation trade-off: the agent should most of the time select the action with the highest value (*exploitation*) but should also sometimes select other actions (*exploration*) to possibly gather new information, especially when the agent detects that the environment might have changed (Ishii et al., 2002). This can be done by transforming each action value into a probability of performing the associated action *a* in the considered state *s* with a Boltzmann softmax equation:

$$P(a/s) = \frac{\exp(\beta.Q(a,s))}{\sum_i \exp(\beta.Q(a_i,s))} \tag{3}$$

where β is a third parameter called the exploration rate ($0 \leq \beta$). Although it is always the case that the action with the highest value has a higher probability of being performed, exploration is further regulated in the following way: when β is set to a small value, action probabilities are close to each other (*e.g.* flat probability distribution) so that there is a non-null probability of selecting an action whose value is not the greatest (exploration). When β is high, the difference between action probabilities is increased so that the action with the highest value is almost always selected (exploitation).

Clearly, these equations devoted to action value learning and action selection rely on crucial parameters: α, β, γ. Most computational models use fixed parameters, hand-tuned for a given task or problem (Chavarriaga et al., 2005; Daw et al., 2005; Frank et al., 2005; Khamassi et al., 2005). However, animals face a variety of tasks and deal with continuously varying conditions. If animal learning does rely on RL as suggested (*e.g.* Luksys et al., 2009; Samejima et al., 2005), there must exist some brain mechanisms to decide, in each particular situations, which set of parameters is appropriate (*e.g.* when an animal performs stereotypical behavior in its nest, or repetitive food gathering behavior in an habitual place, learning rate and exploration rate should not be the same as those used when the animal discovers a new place). Moreover, within a given task or problem, it is more efficient to dynamically regulate these parameters, so as to optimize performance (*e.g.* it is appropriate to initially explore more in a new 'task' while the rule for obtaining rewards is not yet known, to explore less when the rule has been found and the environment is stable, and to re-explore more when a rule change is detected).

The dynamic regulation of parameters is referred to as *meta-learning* by Kenji Doya (Doya, 2002). Meta-learning is a general principle which enables to solve problems of non-stationary systems in the machine learning literature, but the principle does not assume specific methods for the regulation itself. We invite readers interested in particular solutions to refer to methods such as 'ε-greedy' which chooses the action believed to be best most of the time, but occasionally (with probability ε) substitutes a random action (Sutton and Barto, 1998), upper-confidence bound policies 'UCB' which selects actions based on their associated reward averages and the number of times they were selected so far (Auer et al., 2002), EXP3-S for Exponential-weight algorithm for Exploration and Exploitation which is also based on a Boltzmann softmax function (Cesa-Bianchi et al., 2006), uncertainty-based methods awarding bonuses to actions whose consequences are uncertain (Daw et al., 2006), and reviews of these methods applied to abruptly changing environments (Garivier and Moulines, 2008; Hartland et al., 2006).

Although mathematically different, these methods stand on common principles to regulate action selection. Most are based on estimations of the agent's performance, which we will refer to as *performance monitoring*, and on estimations of the stability of the environment across time or its variance when abrupt environmental changes occur, which we will refer to as *task monitoring*. The former employs measures such as the average reward measured with the history of feedback obtained by the agent, or the number of times a given action has already been performed. The latter

often considers the environment's uncertainty, which in economic terms refers to the risk (the known probability of a given reward source), and the volatility (the variance across time of this risk).

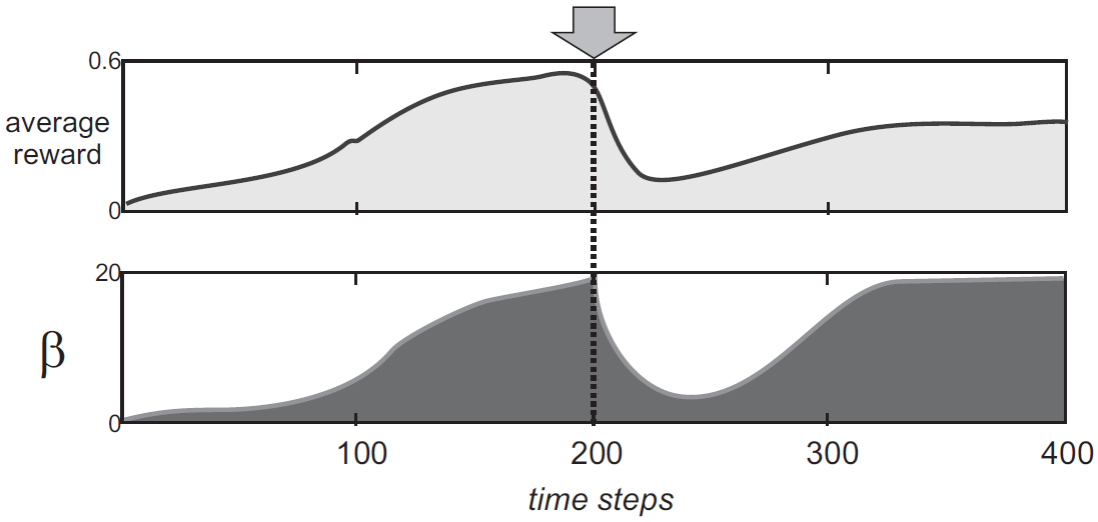

Figure 1. Simulation of a meta-learning algorithm. Adapted from (Schweighofer and Doya, 2003). A change in the task condition from short-term reward to long-term reward at timestep #200 produces a drop in average reward obtained by the agent and thus results in the adaptation of the exploration parameter β.

A simple example of implementation of a meta-learning algorithm was proposed by Schweighofer and Doya (2003) where an agent has to solve a non-stationary Markov decision task also used in human fMRI experiments (Schweighofer et al., 2007; Tanaka et al., 2004). In this task, the agent has two possible actions (pressing one of two buttons). The task is decomposed in two conditions: a short-term condition where one button is associated with a small positive reward and the other button with small negative reward; a long-term condition such that a button with small negative rewards has to be pressed on some steps in order to obtain much larger positive reward in a subsequent step. The authors used a reinforcement learning algorithm where parameters were subject to automatic dynamic regulation. The general principle of the algorithm is to operate such regulation based on variations in the average reward obtained by the agent. **Figure 1** schematizes a sample simulation. The agent learned the short-term condition, starting with a small parameter β (i.e. large exploration level), which progressively increased and produced less exploration as long as the average reward increased. At mid-session, the task condition was changed from short-term condition to long-term condition, resulting in a drop in the average reward obtained by the agent. As a consequence, the parameter β varied allowing more randomness in the agent's actions (due to a small β value), thus allowing the agent to quickly discover the new appropriate contingencies of the task. After some time, the agent learns the new task condition and converges to a more exploitative behaviour (large β value) so as to reduce errors due to exploratory behavior while the environment is now known and stable.

This type of computational process appears suitably robust to account for animal behavioral adaptation. The meta-learning framework has been formalized with neural mechanisms in mind. Doya proposed that the level of different neuromodulators in the prefrontal cortex and striatum might operate the tuning of specific parameters for learning and action selection (Doya, 2008). We will argue below that the meta-learning framework indeed offers valuable tools to study neural mechanisms of decision-making and learning, especially within the medial and lateral prefrontal cortex. This framework offers formal descriptions of the functional biases observed in each structure and also provides explanatory principles for their interaction and role in the regulation of behavior. In the next paragraph, we describe the computational model of the MPFC-LPFC system that we have proposed. Then we simulate it on two particular decision-making tasks on which we previously

recorded MPFC activity. We show that dynamically regulating RL parameters during these tasks based on some heuristics can produce a higher performance than keeping these parameters fixed during the whole task.

METHODS: COMPUTATIONAL MODEL

In (Khamassi et al., 2011) we have proposed a neurocomputational model for the interactions between MPFC and LPFC involved in behavioural regulation during probabilistic and deterministic reinforcement learning tasks performed by monkeys (**Figure 2**). The model largely relies on reinforcement learning principles allowing an agent to adapt its behavioral policy by trial-and-error so as to maximize reward (Sutton and Barto, 1998). Based on the greater anatomical projections of the dopaminergic system to MPFC than to LPFC (Fluxe et al., 1974) and based on previous neurophysiological recordings, we made the assumption that action values are learned and stored in the MPFC through dopaminergic input (Amiez et al., 2005; Holroyd and Cole, 2002; Kennerley et al., 2006; Matsumoto et al., 2007; Rushworth et al., 2007) – although this does not exclude that these values are learned and stabilized in conjunction with the striatum (Samejima et al., 2005) through cortico-basal loops (Alexander et al., 1990). These values are transmitted to the LPFC which selects the action to perform with a certain exploration-exploitation trade-off determined by the current setting of the β parameter (**Equation 3**).

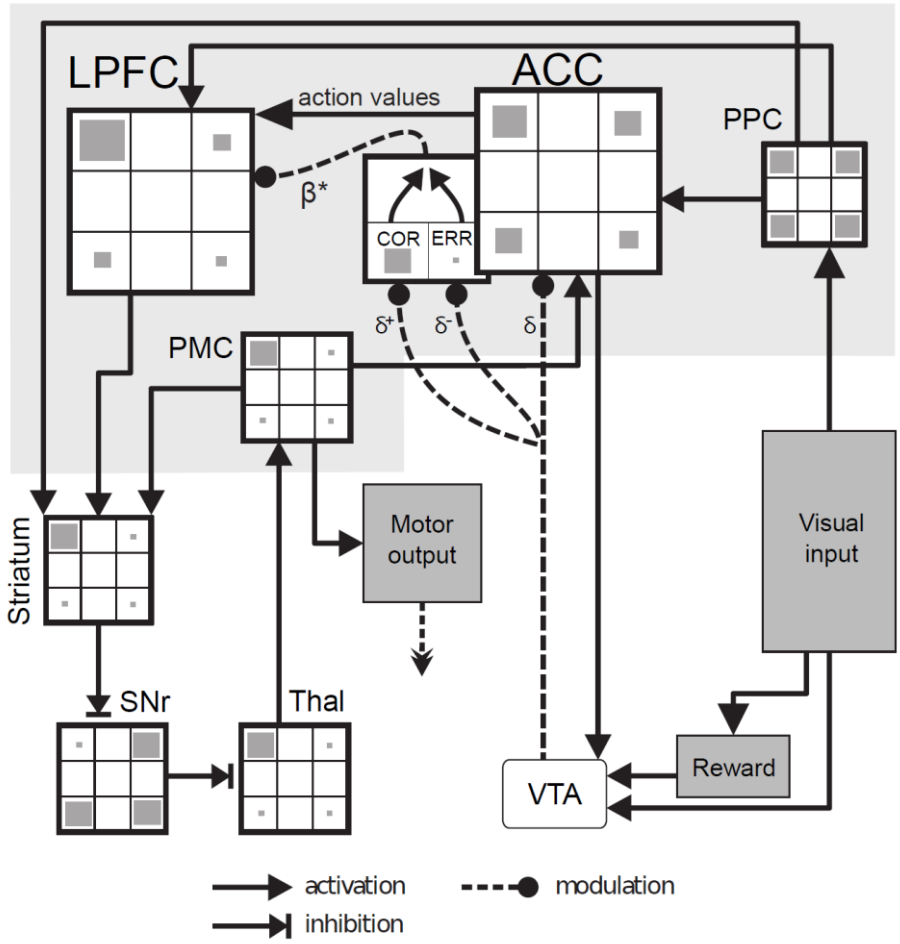

Figure 2. Computational model. Visual input (e.g. targets seen on a screen or objects on a table) is sent to the Posterior Parietal Cortex (PPC). The Anterior Cingulate Cortex (ACC) stores and updates the action value associated with choosing each possible object. When a reward is received, a reinforcement learning signal is computed in the Ventral Tegmental Area (VTA) and is used both to update action values and to compute an outcome history in ACC (COR: correct neuron; ERR: error neuron) used to modulate the desired exploration

level β^*. Action values are sent to the Lateral Prefrontal Cortex (LPFC) which performs action selection. A winner-take-all ensures a single action to be executed at each moment. This is performed in the cortico-basal ganglia loop consisting of Striatum, Substantia Nigra Reticulata (SNr) and Thalamus (Thal) until the Premotor Cortex (PMC). Finally, the output of the PMC is used to command the robot and as an efferent copy of the chosen action sent to ACC.

In addition, the model keeps track of the agent's performance and the variability of the environment to adjust behavioral parameters. Thus the MPFC component monitors positive and negative feedback (Holroyd and Coles, 2002; Brown and Braver, 2005; Sallet et al., 2007; Quilodran et al., 2008) and encodes the outcome history (Seo and Lee, 2007). Thus, in addition to the projection of dopaminergic neurons to MPFC action values, dopamine signals also influence a set of MPFC feedback categorization neurons (**Figure 2**): error (ERR) neurons respond only when there is a negative δ signal; correct (COR) neurons respond only when there is a positive δ signal. COR and ERR signals are then used to update a variable encoding the outcome history (β^*):

$$COR(t) = \delta(t),\ if\delta(t) \geq 0 \qquad (4)$$
$$ERR(t) = -\delta(t),\ if\delta(t) < 0$$

$$\beta^*(t) \leftarrow \beta^*(t) + \eta_+.\mathrm{COR}(t) + \eta_-.\mathrm{ERR}(t)$$

where η+ and η- are updating rates, and $0 < \beta^* < 1$. Such a mechanism was inspired by the concept of vigilance employed by Dehaene and Changeux (1998) to modulate the activity of workspace neurons whose role is to determine the degree of effort in decision-making. As for the vigilance which is increased after errors, and decreased after correct trials, the asymmetrical learning rates (η+ and η-) enables sharper changes in response to either positive or negative feedback depending on the task. In the present model, these parameters have been tuned to capture global behavioral properties and changes in reaction times of monkeys' behavior during a problem-solving task (Khamassi et al., 2011): small progressive changes after errors; sharp changes once the correct answer is found to promote exploitation.

The adjustment of behavioral parameters based on such outcome history follows meta-learning principles (Doya, 2002; Ishii et al., 2002) and is here restricted to the tuning of the β parameter which regulates the exploration rate of the agent. Following previous machine learning models, the exploration rate β is adjusted based on variations of the average reward (Auer et al., 2002; Schweighofer and Doya, 2003) and on the occurrence of uncertain events (Daw et al., 2006; Yu and Dayan, 2005). In short, a decrease of the outcome history – denoting a drop of performance – results in a decrease of β (more exploration); an increase in the outcome history – denoting an improvement in performance – results in an increase of β (more exploitation). The resulting parameter modulates action selection within the lateral prefrontal cortex, consistent with its involvement in the exploration-exploitation trade-off (Cohen et al., 2007; Daw et al., 2006; Frank et al., 2009). In addition, the repetitive occurrence of particular uncertain events that turn out to be systematically followed by a drop of performance (*e.g.*, abrupt cued and initially unknown changes in the task condition) can be learned as requiring a reset of β to its initial low value β_0 (*i.e.*, the model restarts to explore each time it detects such events). In order to learn that particular cues or objects require a reset of exploration, the model associates so-called "meta-values" to each cue and object involved in the task. These meta-values are initialized to zero. Each time the presentation of a cue/object is followed by a decrease in the reward average, the corresponding meta-value is decreased according to the following equation:

$$M(o_i, t) \leftarrow M(o_i, t) + \omega.\theta(t) \qquad (5)$$

where $M(o_i,t)$ is the meta-value associated to cue/object o_i at time t, ω is an update rate and $\theta(t)$ is the estimated reward average at time t.

When the meta-value associated with any object is below a certain threshold *T* (empirically fixed to require approximately 10 presentations before learning in the robotic simulations presented in the third result section), subsequent presentations of this object to the model automatically trigger a reset of the exploration level β(t) to its initial value β_0; The rest of the time, the exploration level is determined by the current outcome history $\beta^*(t)$:

$$\beta(t) = \begin{cases} \beta_0, if \exists i, [M(o_i,t) < T] \wedge [o_i \equiv presented] \\ f(\beta^*(t)), otherwise \end{cases} \quad (6)$$

where *T* is the chosen threshold and *f(.)* is a sigmoid function transforming the outcome history (between 0 and 1) into an appropriate exploration level (between 0 and 10).

This part of the model provides a normative way of regulating the exploration level without specifying the precise underlying physiological mechanism. Interestingly, although the precise molecular and cellular mechanisms in the prefrontal cortex underlying shifts between exploration and exploitation are not yet known, there is however accumulating evidence that differential levels of activation of dopamine receptors D1 and D2 in the prefrontal cortex may produce distinct states of activity: a first state entertaining multiple network representations nearly simultaneously and thus permitting "an exploration of the input space"; a second state where the influence of weak inputs on PFC networks is shut off so as to stabilize one or a limited set of representations, which would then have complete control of PFC output, and would thus promote exploitation (Durstewitz and Seamans, 2008). Other models have been proposed to regulate the exploration-exploitation trade-off in action selection via a neuromodulation of extrinsic and inhibitory synaptic weights between competing neurons in the prefrontal cortex (Krichmar, 2008). A strong common point between these two types of models is to produce an alternation between a state with a high entropy in the action probability distribution (exploration) and a state with a low entropy in the action probability distribution (exploitation), which principle is here abstracted through the use of Boltzmann's softmax function (**Equation 3**).

RESULTS (I): DETERMINISTIC TASK

In (Khamassi et al., 2011), this model was first simulated on a deterministic problem solving task performed by monkeys (Quilodran et al., 2008) so as to reproduce monkey behavioral performance. In this task, 4 targets are presented on a touch screen at each trial. Monkeys have to find by trial-and-error which target is associated to reward (search phase). Once the correct target is found (first rewarded trial of the problem), monkeys have to repeat this choice during 3, 7 or 11 trials (repetition phase). Such variability of the duration of the repetition phase was imposed to prevent monkeys from expecting the end of this phase and thus from behaving differently. After the end of the last repetition trial, a Signal to Change (SC) is presented on the screen which indicates the beginning of a new problem: the rewarding target is changed and the animal has to perform a new search. Animals have been well pre-trained on this task and analysis of the behavior of 4 monkeys (Khamassi et al., 2011) shows that they choose the previously rewarded target after less than 20% of SC presentation, and rather re-explore other possible targets in more than 80% of the times.

We previously found that our computational model can well reproduce global properties of monkey behavior in this task (number of errors, average duration of each phase...). Here we want to show that using some metal-learning principles – *i.e.* employing different exploration parameters β_S and β_R for the search and repetition phases – can produce a better performance on this task than employing a single constant exploration parameter for the two phases. To do so, we made

simulations of a simple Q-learning model (using **Equations 1, 2 and 3** described above) on a sample sequence of 286 problems (corresponding to 1724 trials) performed by a monkey and explored the ability of combinations of parameters α, β_S and β_R (with γ=0) to either maximize the likelihood that the model makes the rewarded choice at each trial (*reward maximization*) or maximize the likelihood that the model reproduces monkey's choice at each trial (*fit maximization*). We tested different parameter sets in the following way:

- α: from 0.1 to 1.0 with 0.1 steps,
- β_S: 0, then from exp(-2.3) to exp(5.7) with exp(0.1) steps (i.e. $0 < \beta_S < 299$),
- β_R: 0, then from exp(-2.3) to exp(5.7) with exp(0.1) steps (i.e. $0 < \beta_R < 299$).

Figure 3 shows the performance for both reward maximization (left) and fit maximization (right) obtained by the model as a function of combinations of the two exploration parameters (β_S and β_R). The figure shows that the best performance is obtained with different exploration levels between search and repetition: $0 \le \beta_S \le 10^0$ and $10^0 \le \beta_R \le 10^1$. In other words, a low exploration parameter β_S is required during search (*i.e.* more exploration), and a higher exploration level is required during repetition ($\beta_R \ge \beta_S$, *i.e.* more exploitation). In contrast, a model which uses the same exploration level during the two phases ($\beta_S = \beta_R$) would be situated on the diagonal of the plotted matrix and would thus not be in the region were reward is maximized. Interestingly, since the monkey had been well pre-trained and its behavior was stereotypical and nearly optimal, the combination of exploration parameters that maximize the fit is very close to the combination of parameters that maximize reward, with a slightly smaller required β_S to accurately fit monkey's behavior (**Figure 3**).

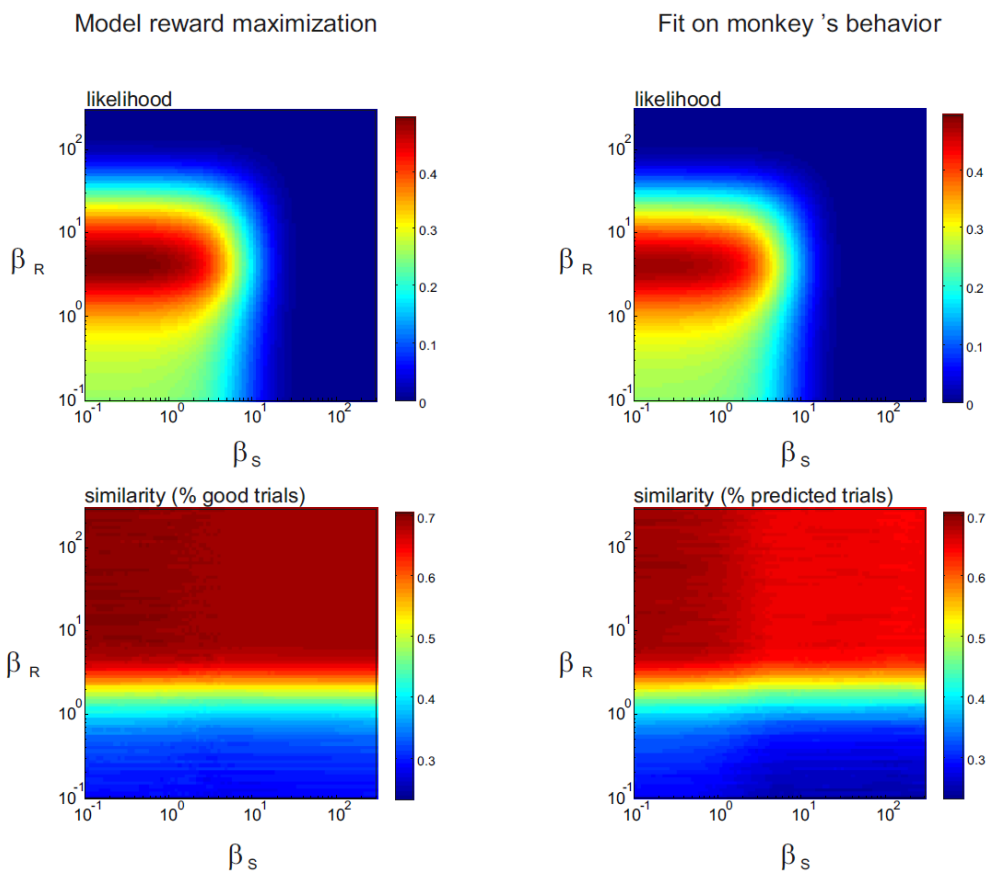

Figure 3. Effect of different combinations of parameters on the model's performance during the deterministic task of (Quilodran et al., 2008). (Left) performance (likelihood) of the model in maximizing

reward during the sampled problems of the task. (Right) performance (likelihood) of the model in fitting monkey's choices during the sampled problems of the task. Bottom charts show the % of correct trials corresponding to the likelihood (Top charts) obtained with each combination of parameters.

These results illustrate that enabling a dynamic regulation of the exploration parameter β and using some heuristics (*e.g.* using a small β during the search phase, after perceiving the Signal to Change, to promote exploration; increasing β after the first rewarded trial to promote exploitation during the repetition phase) can be relevant to solve such deterministic decision-making task. In addition, our neurocomputational model having been built so as to respect anatomical constraints and to reproduce global properties of monkey behavior in this task (Khamassi et al., 2011), we can generate a list of experimental predictions that have to be tested by future simultaneous neurophysiological recordings of the medial and lateral prefrontal cortex during this task:

1. We should find feedback categorization neurons (Matsumoto et al., 2007; Quilodran et al., 2008) and neurons encoding the outcome history (Seo and Lee, 2007) mainly in the medial prefrontal cortex (MPFC) which is hypothesized to be involved in performance monitoring.
2. The desired exploration level extracted from the current performance estimation should modulate the decision process – putatively through a mechanism similar to the softmax function (**Equation 3**; Krichmar, 2008) – in the lateral prefrontal cortex (LPFC). Thus exploration-based modulation should effect only on LPFC action probability neurons and not on MPFC action value neurons. In the model, we made the choice to keep original action values (that is, not altered by the exploration-based modulation) in the MPFC 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 (Procyk and Goldman-Rakic, 2006) – in LPFC than in MPFC due to the decision-making process based on the softmax function (which increases the contrast between action values when β is high).
4. There should be an increase of spatial selectivity in LPFC but not in MPFC during the repetition period. Such increase of spatial selectivity in LPFC neurons in the model is due to the higher β parameter used in the softmax function during the repetition phase than during the search phase so as to produce correct robust performance during repetition.

RESULTS (II): PROBABILISTIC TASK

We then tried to generalize the above mentioned model by testing it on a more probabilistically rewarded decision-making task developed by (Amiez et al., 2006). In this task monkeys were also facing a touch screen and had to find which one of two targets had the best rewarding rate. However, in this case, the reward distribution was stochastic rather than deterministic. 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 (**Figure 4A**). Thus, although each "problem" in this task also comprised a search phase and a repetition phase, a single rewarded trial was not sufficient to find out the best target. Monkeys had to sample several outcomes for each target before being able to estimate each target's value. As a consequence, there was no sharp change between search and repetition 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. At the end of the repetition period a new problem started, like in the deterministic version of the task. However, if after 50 trials the monkey had not entered the repetition phase, the problem was considered as failed, it was aborted and a new problem started. The exact same behavioral protocol and behavioral measures were used to evaluate the model's performance in the task.

In addition to analyzing the influence of the parameters α and β on the performance of the model, we also enabled the model to reset its Q-values at the beginning of each problem, in response to the presentation of the Signal to Change, and looked at the influence of different initial Q-values (namely 'Q_{init}' parameter) on the exploration process. Since the transition from the search phase to the repetition phase is not as clear as for the deterministic task, instead of using two separate exploration parameters (i.e. β_S and β_R), we compared a version of the model with a single fixed β and a model using the dynamic regulation of β based on measurement of the outcome history β^* (Khamassi et al., 2011; **Equations 4-6**). Finally, the performance was measured both in terms of the number of trials required by the model to find the best target and the optimal target ratio, that is the number of successful (non-aborted) problems.

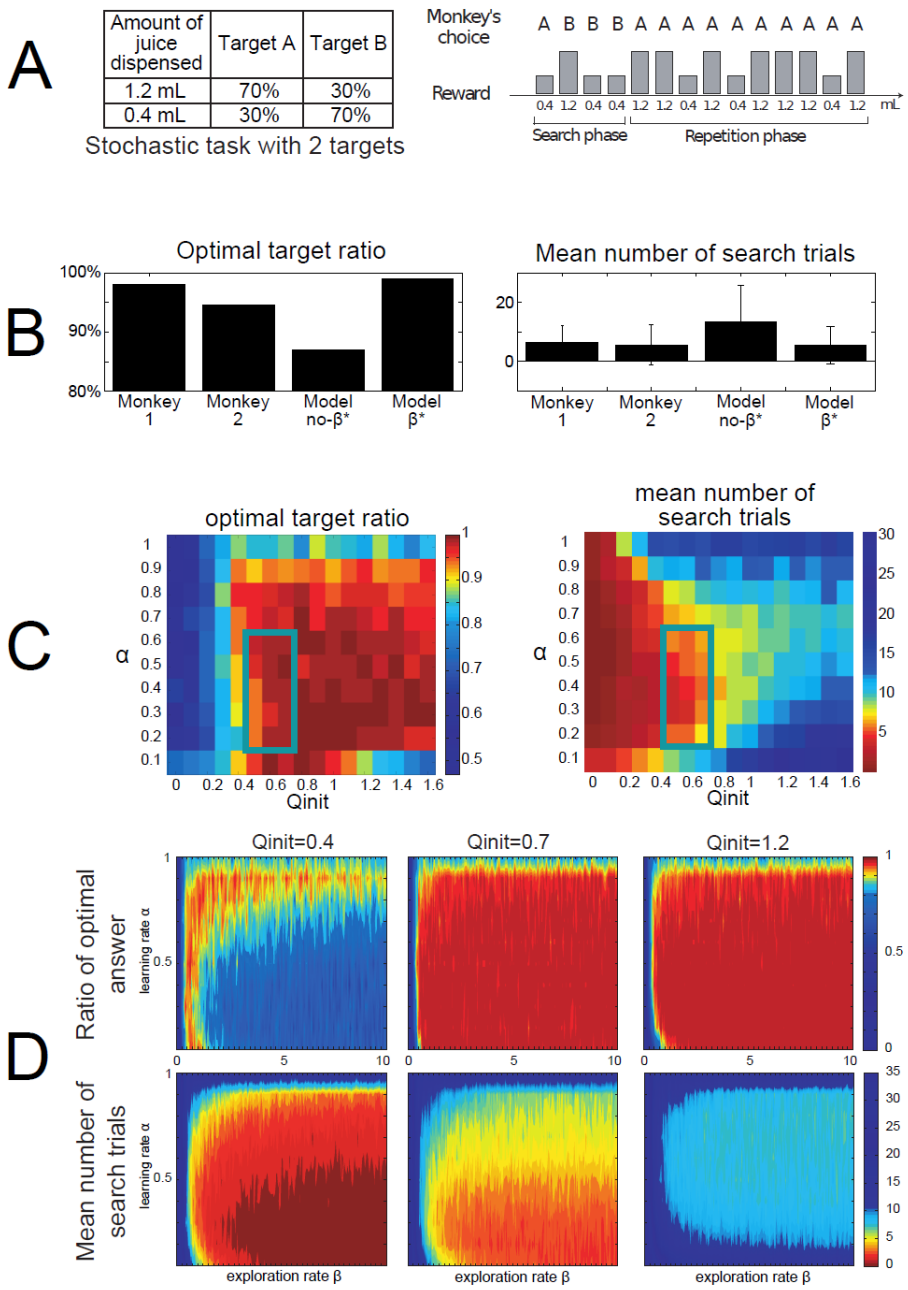

Figure 4. Simulation of the model on the probabilistic task of (Amiez et al., 2006). (A-Left) Probability of getting a large or small reward when choosing target A or B. (A-Right) Typical problem decomposed in search and repetition phases. (B) Compared performance of monkeys and models with and without the meta-learning

mechanism to dynamically regulate the exploration parameter β. The optimal target ratio is the percentage of successfully completed problems. (C) Regions of the parameters space that produce optimal performances on this task. (D) The performance also depends on the initial Q-values to which targets are reset at the beginning of each new problem and which also influence the level of exploration.

A naive test on the stochastic task with the optimal parameters used with the deterministic task and a fixed exploration level – that is without the β^*-based mechanism for dynamic exploration regulation ($\alpha = 0.9$, $\beta = 5.2$, $Q_{init} = 0.4$) – 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 "Model no- β^*" on **Figure 4B**). The adaptation of the parameters with an exploration rate β regulated based on the outcome history (Khamassi et al., 2011) was more successful (see "Model β^*" on **Figure 4B**). Roughly, the optimal α is between 0.4 and 0.6, and the optimal Q_{init} between 0.6 and 0.8 (**Figure 4C**). 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 (Amiez et al., 2006). Interestingly, optimization of the model in the stochastic task led to a lower learning rate ($\alpha = 0.5$) than optimization of the model in the deterministic task ($\alpha = 0.9$; Khamassi et al., 2011). This illustrates the necessity in probabilistic reward schedules to slowly integrate outcome information and to repeat several times rewarded actions before being confident of one's own behavior (Behrens et al., 2007).

In addition, the optimization including the exploration level 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 small ($\beta > 5$) and α was between 0.2 and 0.9 (**Figure 4D**). In the stochastic task, the regulation of β based on the outcome history elicits values close to 10, the highest values possible for β in these simulations, hence corresponding to the values where β is optimal for this stochastic task. This was in part due to the nature of this task in which only 2 targets were available, decreasing the search space. So the best strategy was clearly exploitative.

Further analyses showed that the two indicators of performance had opposite tendencies with respect to the initial Q-values. As shown in **Figure 4D**, 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. Thus there appears to be a trade-off between minimizing the length of search phase and maximizing the chance to complete the problem. An average initial Q-value can balance these two effects so as to have a relatively good performance with the two indicators. Further analyses revealed that the initial Q-value is highly correlated to the search period length (correlation coefficient is 0.99 with p-value < 10e-14).

These results show the importance of the reset of Q-values when a new problem started in the stochastic task. The initial Q-values should not be 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 chances to elicit good performance than the rest of the parameter space. Interestingly, electrophysiological data from the medial prefrontal cortex (MPFC) recorded during this stochastic problem solving task 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; Amiez et al., 2006). 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 MPFC participates in the extraction of information from the environment to

regulate RL parameters, but also that MPFC sets the action values used as reference to initiate exploratory behavior.

RESULTS (III): HUMAN-ROBOT INTERACTION GAME

Finally, in (Khamassi et al., 2011) we tested our neurocomputational model on a robotic platform to test its ability to cope with real-world uncertainties (**Figure 5-left**). Instead of having to choose between 4 targets on a touch screen, the iCub humanoid robot had to perform a simple human-robot interaction game so as to find, among a set of 4 cubes on a table, which cube had a circle on its hidden face (corresponding to the reward). The monkeys visual system was simplified so as to a priori recognize four different shapes: the alignment of the cubes corresponding to a GO signal (**Figure 5A-B**); the circle corresponding to the reward (**Figure 5C**); a wooden board which was initially set as a neutral object (*i.e.* null initial meta-value; **Figure 5D**); and human's hands on the cubes also initially set as neutral (**Figure 5F**). Since we focused on the dynamic regulation of decision-making without taking into account motor control aspects, the robot simply had to point out the chosen cube and the human then grasped and flipped the cube so as to show the robot its hidden face.

The first interesting result is that the neurocomputational model enabled the robot to cope with the intrinsic perceptual uncertainty generated by such type of human-robot interaction: if it failed to perceive the outcome of a trial due to the human's movements or due to an inability to recognize a shape, the robot would simply adapt its choice with reinforcement learning processes until finding the correct cube; if the robot had not found the circle after having chosen each possible cube, it would simply continue to explore until finding it; if the robot had mistakenly recognized a circle under the wrong cube, it would simply choose it again, recognize that it's an error, and then continue to explore other cubes (Khamassi et al., 2011).

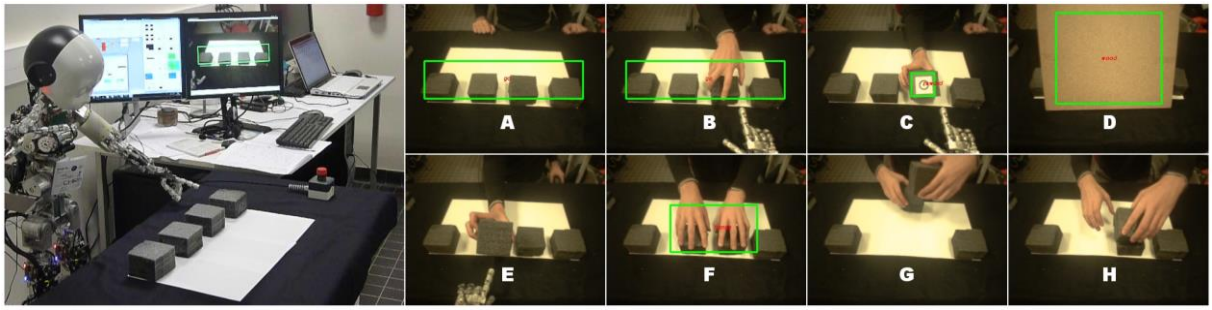

Figure 5. Human-robot interaction scenario used to test the ability of the model to cope with real-world uncertainties. (Left) The model is tested on the iCub humanoid robot that has to learn to choose the rewarding cube among a set on a table. (Right) Illustration of the visual perceptions of the robot during different task events. The alignment of the cubes indicates a GO signal. The circle under the correct cube is the rewarding signal. The wooden board and the human's hands on the cubes are initially set as neutral signals to which the model will progressively learn to associate a reset of exploration.

The second experiment that we did was to use the initially neutral objects as 'Signals to Change' (SC) similar to the previous monkey tasks: each time they are presented, the rewarding cube's location is changed. More precisely, the wooden board is used to hide the cubes while the human shuffles the cubes; the human's hands on the cubes were used to represent some sort of "cheating" behavior by the human. While in the previous experiments the model and the monkeys knew a priori that a particular signal SC (*i.e.* a circle shown on the touch screen) was associated with a change in the task condition, and thus a shift in the rewarded target, here we wanted the model to autonomously learn that some cues are always followed by errors and thus should be associated to

an environmental change that requires a new exploration. This learning process was intended to propose a hypothetical mechanism by which monkeys could learn the structure of the previous tasks during their pre-training phases. To do so, null meta-values were initially associated to each perceivable shape, and each time the presentation of one shape was followed by a drop in the average reward, the model would decrease the corresponding meta-value (**Equation 5**). If this is consistently repeated for a given shape, its meta-value will decrease below a certain threshold which would subsequently trigger a new exploration phase each time the shape is perceived again (**Equation 6**; Khamassi et al., 2011).

With this principle, the robot learned that presentation of the board was always followed by a drop in the average reward. Thus the board acquired a negative meta-value and the robot systematically shifted its behavior and restarted to explore each time the board appeared again. Interestingly, such learning process led to an improvement of the performance of the robot. During the second part of each game, the robot made fewer errors on average during search phases, and required fewer trials to find the correct cube. Concretely, before the exploration reset was learned, in 65 problems initiated by a board presentation, the robot took on average 3.5 trials to find the correct cube. After the exploration reset was learned for the wooden board, in 36 problems initiated by a board presentation, the robot took on average 2.2 trials to find the correct cube. The difference is statistically significant (Kruskal-Wallis test, $p < 0.001$).

Such meta-learning mechanism constitutes a prediction on the way monkeys may learn to react to the Signal to Change (SC) during the pre-training phases of the previous problem solving-tasks. Future recordings and analyses of monkeys' behavior during pre-training should reveal whether they indeed learn to correctly repeat the rewarded choice before learning to re-explore each time the SC is presented, or whether it is the opposite.

CONCLUSIONS

Accumulating evidence suggest that the frontal cortex could contribute to flexible goal-directed behaviors and to learning based on feedback obtained from the environment (Mars et al., 2011; Miller and Cohen, 2001). Recent electrophysiological findings suggest a specialization of the frontal cortex where the medial prefrontal cortex (MPFC) monitors performance to modulate decision-making in the lateral prefrontal cortex (LPFC) (Matsumoto et al., 2007; Procyk et al., 2000; Seo and Lee, 2009). Several computational models have tackled this specialization, either by considering that MPFC monitors conflict between competing actions to increase the gain in the LPFC (Botvinick et al., 2001), proposing that MPFC computes the current error-likelihood (Brown and Braver, 2005), or proposing that MPFC detect salient unpredicted events relevant for behavioral adaptation (Alexander and Brown, 2011). We extended these lines of argument by proposing a computational model describing MPFC function in terms of meta-learning (Doya, 2002). The MPFC could be generally involved in monitoring performance relative to the current environment's properties so as to tune parameters of reinforcement learning and action selection. Consistently with this proposition, Rushworth and colleagues have recently shown that the MPFC in humans is important to track the environment's volatility (variations in the reward rate) and adapt subsequent behavior (Behrens et al., 2007).

The model synthesizes a wide range of anatomical and physiological data concerning the MPFC-LPFC system (Khamassi et al., 2011). In addition, certain aspects of the neural activity produced by the model during performance of the tasks resembles previously reported MPFC neural patterns that where not a priori built into the model (Procyk et al., 2000; Quilodran et al., 2008). Specifically, like neurons in the MPFC, in the model MPFC feedback categorization neurons responded more to the first correct trial and not to subsequent correct trials, a consequence of the high learning rate suitable

for the deterministic task. This provides a functional explanation for these observations. Moreover, detailed analyses of the model's activity properties during simulations provide testable predictions on the proportion of neurons in MPFC and LPFC that should carry information related to different variables in the model, or that should vary their spatial selectivity between search and repetition phases. In the future we will test hypotheses emerging from this model on simultaneously recorded MPFC and LPFC activities during such decision-making tasks.

The work presented here also illustrated the robustness of biological hypotheses implemented in this model by demonstrating that it could allow a robot to solve similar tasks in the real-world. Comparison of simulated versus physical interaction of the robot with the environment showed that real-world performance produced unexpected uncertainties that the robot had to accommodate (e.g. obstructing vision of an object with its arm and thus failing to perceive it, or perceiving a feature in the scene which looked like a known object but was not). The neuro-inspired model provided learning abilities that could be suboptimal in a given task but which enabled the robot to adapt to such kind of uncertainties in each of the experiments. Besides, the model enabled the robot to show efficient behavioral adaptation during human-robot interaction and to successfully adapt to unexpected uncertainty introduced by the human (e.g. cheating). The robot could also learn that new objects introduced by the human could be associated with changes in the task condition. This was achieved by learning meta-values associated with different objects. These meta-values could either be reinforced or depreciated depending on variations in the average reward that followed presentation of these objects. The object which was used to hide cubes on the table while the human changed the position of the reward was learned to have a negative meta-value and triggered a new behavioral exploration by the robot after learning. Such meta-learning processes may explain the way monkeys learn the significance of the Signal to Change during the pre-training phase of the two studied laboratory experiments. In future work, we will analyze such pre-training behavioral data and test whether the model can explain the evolution of monkey behavioral performance along such process.

Such kind of pluridisciplinary approach can provide tools both for a better understanding of neural mechanisms of behavioral adaptation and for the design of artificial systems that can autonomously extract regularities from the environment and interpret various types of feedback (rewards, feedback from humans) to appropriately adapt their choices.

REFERENCES

Alexander GE, Crutcher MD, DeLong MR (1990) Basal ganglia-thalamocortical circuits: parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. Prog Brain Res 85:119-46.

Alexander WH, Brown JW (2011) Medial prefrontal cortex as an action-outcome predictor. Nat Neurosci 14:1338-1344.

Amiez C, Joseph JP, Procyk E (2005a) Primate anterior cingulate cortex and adaptation of behaviour. In: From monkey brain to human brain (Dehaene S, Duhamel JR, Hauser MD, Rizzolatti G, eds): MIT Press.

Amiez C, Joseph JP, Procyk E (2005b) Anterior cingulate error-related activity is modulated by predicted reward. Eur J Neurosci 21:3447-3452.

Amiez C, Joseph JP, Procyk E (2006) Reward encoding in the monkey anterior cingulate cortex. Cereb Cortex 16:1040-1055.

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit. Machine Learning 47:235-256.

Badre D, Wagner AD (2004) Selection, integration, and conflict monitoring; assessing the nature and generality of prefrontal cognitive control mechanisms. Neuron 41:473-487.

Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47:129-141.

Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. Nat Neurosci 10:1214-1221.

Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD (2001) Conflict monitoring and cognitive control. Psychol Rev 108:624-652.

Brown JW, Braver TS (2005) Learned predictions of error likelihood in the anterior cingulate cortex. Science 307:1118-1121.

Cesa-Bianchi N, Gabor L, Stoltz G (2006) Regret minimization under partial monitoring. Math Oper Res 31.

Chavarriaga R, Strösslin T, Sheynikhovich D, Gerstner W (2005) A computational model of parallel navigation systems in rodents. Neuroinformatics 3: 223-42.

Cohen JD, McClure SM, Yu AJ (2007) Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philos Trans R Soc Lond B Biol Sci 362:933-42.

Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8:1704-1711.

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876-879.

Doya K (2002) Metalearning and neuromodulation. Neural Netw 15:495-506.

Doya K (2008) Modulators of decision making. Nature Neurosci 11:410-16.

Durstewitz D, Seamans JK (2008) The dual-state theory of prefrontal cortex dopamine function with relevance to catechol-o-methyltransferase genotypes and schizophrenia. Biol Psychiatry 64:739-749.

Fluxe K, Hokfelt T, Johansson O, Jonsson G, Lidbrink P, Ljungdahl A (1974) The origin of the dopamine nerve terminals in limbic and frontal cortex. Evidence for mesocortico dopamine neurons. Brain Research 82:349–55.

Frank MJ (2005) Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. J Cogn Neurosci 17(1):51-72.

Frank MJ, Doll BB, Oas-Terpstra J, Moreno F (2009) Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat Neurosci 12:1062-8.

Garivier A, Moulines E (2008) On upper-confidence bound policies for nonstationary bandit problems. Arxiv preprint arXiv:0805.3415.

Hartland C, Gelly S, Baskiotis N, Teytaud O, M. S (2006) Multi-armed bandit, dynamic environments and meta-bandits. In: NIPS-2006 workshop, Online trading between exploration and exploitation. Whistler, Canada.

Holroyd CB, Coles MG (2002) The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol Rev 109:679-709.

Houk JC, Adams J, Barto AG (1995) A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Models of information processing in the basal ganglia, pp 249-270. Cambridge, MA: MIT Press.

Humphries MD, Prescott TJ (2010) The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward. Prog Neurobiol 90:385-417.

Ishii S, Yoshida W, Yoshimoto J (2002) Control of exploitation-exploration meta-parameter in reinforcement learning. Neural Netw 15:665-687.

Johnston K, Levin HM, Koval MJ, Everling S (2007) Top-down control-signal dynamics in anterior cingulate and prefrontal cortex neurons following task switching. Neuron 53:453-462.

Kennerley SW, Walton ME, Behrens TE, Buckley MJ, Rushworth MF (2006) Optimal decision making and the anterior cingulate cortex. Nat Neurosci 9:940-947.

Khamassi M, Lachèze L, Girard B, Berthoz A, Guillot A (2005) Actor-critic models of reinforcement learning in the basal ganglia: From natural to artificial rats. Adapt Behav 13(2):131-48.

Khamassi M, Mulder AB, Tabuchi E, Douchamps V, Wiener SI (2008) Anticipatory reward signals in ventral striatal neurons of behaving rats. Eur J Neurosci 28:1849-1866.

Khamassi M, Lallée S, Enel P, Procyk E, Dominey PF (2011) Robot cognitive control with a neurophysiologically inspired reinforcement learning model. Frontiers in Neurorobotics 5:1.

Kolling N, Behrens TE, Mars RB, Rushworth MF (2012) Neural mechanisms of foraging. Science 336(6077):95-98.

Kouneiher F, Charron S, Koechlin E (2009) Motivation and cognitive control in the human prefrontal cortex. Nat Neurosci 12:939-945.

Krichmar JL (2008) The neuromodulatory system – a framework for survival and adaptive behavior in a challenging world. Adapt Behav 16:385-399.

Luksys G, Gerstner W, Sandi C (2009) Stress, genotype and norepinephrine in the prediction of mouse behavior using reinforcement learning. Nat Neurosci 12:1180-1186.

MacDonald AW, 3rd, Cohen JD, Stenger VA, Carter CS (2000) Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. Science. 288: 1835-1838.

Mars RB, Sallet J, Rushwort MFS, Yeung N (2011) Neural basis of motivational and cognitive control. Cambridge, MA: MIT Press.

Matsumoto M, Matsumoto K, Abe H, Tanaka K (2007) Medial prefrontal cell activity signaling prediction errors of action values. Nat Neurosci 10:647-656.

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167-202.

Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H (2006) Midbrain dopamine neurons encode decisions for future action. Nat Neurosci 9:1057-1063.

Paus T (2001) Primate anterior cingulate cortex: where motor control, drive and cognition interface. Nat Rev Neurosci 2:417-424.

Procyk E, Tanaka YL, Joseph JP (2000) Anterior cingulate activity during routine and non-routine sequential behaviors in macaques. Nat Neurosci 3:502-508.

Procyk E, Joseph JP (2001) Characterization of serial order encoding in the monkey anterior cingulate sulcus. Eur J Neurosci 14:1041-1046.

Procyk E, Goldman-Rakic PS (2006) Modulation of dorsolateral prefrontal delay activity during self-organized behavior. J Neurosci 26:11313-11323.

Quilodran R, Rothé M, Procyk E (2008) Behavioral shifts and action valuation in the anterior cingulate cortex. Neuron 57(2):314–325.

Reynolds JN, Hyland BI, Wickens JR (2001) A cellular mechanism of reward-related learning. Nature 413:67-70.

Rothé M, Quilodran R, Sallet J, Procyk E (2011) Coordination of High Gamma Activity in Anterior Cingulate and Lateral Prefrontal Cortical Areas during Adaptation. J Neurosci 31:11110-11117.

Rudebeck PH, Behrens TE, Kennerley SW, Baxter MG, Buckley MJ, Walton ME, Rushworth MF (2008) Frontal cortex subregions play distinct roles in choices between actions and stimuli. J Neurosci 28:13775-13785.

Rushworth MF, Behrens TE (2008) Choice, uncertainty and value in prefrontal and cingulate cortex. Nat Neurosci 11:389-397.

Sallet J, Quilodran R, Rothé M, Vezoli J, Joseph JP, Procyk E (2007) Expectations, gains, and losses in the anterior cingulate cortex. Cogn Affect Behav Neurosci 7:327-336.

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

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593-1599.

Schweighofer N, Doya K (2003) Meta-learning in reinforcement learning. Neural Netw 16:5-9.

Schweighofer N, Tanaka SC, Doya K (2007) Serotonin and the evaluation of future rewards: theory, experiments, and possible neural mechanisms. Ann N Y Acad Sci 1104:289-300.

Seo H, Lee D (2007) Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. J Neurosci 27:8366-8377.

Seo H, Lee D (2008) Cortical mechanisms for reinforcement learning in competitive games. Philos Trans R Soc Lond B Biol Sci 363:3845-3857.

Seo H, Lee D (2009) Behavioral and neural changes after gains and losses of conditioned reinforcers. Journal of Neuroscience, 29(11):3627–3641.

Shima K, Tanji J (1998) Role for cingulate motor area cells in voluntary movement selection based on reward. Science 282:1335-1338.

Silton RL, Heller W, Towers DN, Engels AS, Spielberg JM, Edgar JC, Sass SM, Stewart JL, Sutton BP, Banich MT, Miller GA (2010) The time course of activity in dorsolateral prefrontal cortex and anterior cingulate cortex during top-down attentional control. Neuroimage 50:1292-1302.

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, MA: MIT Press.

Sul JH, Kim H, Huh N, Lee D, Jung MW (2010) Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. Neuron 66:449-460.

Tanaka SC, Doya K, Okada G, Ueda K, Okamoto Y, Yamawaki S (2004) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. Nat Neurosci 7:887-893.

Yu AJ, Dayan P (2005) Uncertainty, neuromodulation, and attention. Neuron 46(4):681-92.