Meta-learning, cognitive control, and physiological interactions between medial and lateral prefrontal cortex

Mehdi Khamassi, Charles R E Wilson, Marie Rothé, René Quilodran, Peter F Dominey, Emmanuel Procyk

Meta-learning, cognitive control, and physiological interactions between medial and lateral prefrontal cortex

Authors:

Mehdi Khamassi[1,2], Charles R.E. Wilson[1], Marie Rothé[1], René Quilodran[3], Peter F. Dominey[1], Emmanuel Procyk[1]

authors addresses:

[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, CNRS UMR 7222, 4 place Jussieu, F-75252, Paris Cedex 05, France

[3] Escuela de Medicina, Departamento de Pre-clínicas, Universidad de Valparaíso, Hontaneda 2653, Valparaíso, Chile

Acknowledgement. MK and PFD are funded by Agence Nationale de la Recherche Amorces , CREW by Neurodis Foundation, MR by Fondation pour la Recherche Médicale, RQ by Facultad de Medicina Universidad de Valparaíso , EP by Agence National de la Recherche, Centre National de la Recherche Scientifique (CNRS), and Fondation de France.

1- INTRODUCTION

Reacting to errors and adapting choices to achieve long-term goals are fundamental abilities used in reasoning and problem solving. These abilities require the proper operation of executive functions which allow decision-making and the organization of behavior in new and challenging situations. Several theoretical models propose that this involves a superior cognitive control of action in particular when routines need to be modified or reorganized [20, 64, 74]. There is evidence for a range of sub-component processes, including selection, active maintenance and use of information for planning (working memory), inhibition, and performance monitoring [51]. In problematic situations, automatic responding becomes inefficient or suboptimal, and so cognitive control has to be triggered in order to promote the selection of appropriate actions given the circumstances. It is clear that the proper functioning of these processes is not dependent on the integrity of one particular brain structure but on a specific distributed network. Converging evidence suggests that subdivisions of the prefrontal cortex house an important part of this network, but the mechanisms used to implement these processes remain unclear.

In the past 15 years, the Reinforcement Learning (RL) theory has been 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 [29, 79]. Its extensive

use in the computational neuroscience literature is grounded on the observation that dopaminergic neurons respond according to reward prediction error [70], that dopamine strongly innervates the prefrontal cortex and striatum and there modifies synaptic plasticity [30, 63], and that prefrontal cortical and striatal neurons encode a variety of RL-consistent information [19, 36, 69, 78].

However, RL models rely on crucial meta-parameters (e.g. learning rate, exploration rate, temporal discount factor) that need to be dynamically tuned to cope with variations in the environment. 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 meta-parameters". Regulation of decision-making has been largely studied in terms of "cognitive control", and is hypothesized to involve interactions between subdivisions of the prefrontal cortex (PFC), especially the medial and lateral PFC. We argue here that neural data concerning such interactions can be interpreted with the Meta-Learning theoretical framework proposed by Kenji Doya to synthesize computational principles for regulating RL meta-parameters [21].

2- THEORETICAL BASES OF META-LEARNING

Reinforcement Learning (RL) is a research field within computer science that studies how an agent can appropriately adapt its behavioural policy so as to reach a particular goal in a given environment [79]. 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 stochastic 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: *Q(s,a)*. In a popular class of RL algorithms called Temporal-Difference Learning, which has shown strong resemblance with dopaminergic signaling [70], 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 $\gamma.\max_a Q(s_t, a)$ 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 meta-parameter γ ($0 \leq \gamma < 1$) called the discount factor, which gives the temporal horizon of reward expectations. If γ is tuned to a high value, the

agent has a behaviour oriented towards long-term rewards. If *γ* is tuned to a value close to 0, the agent focuses on immediate rewards.

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 [79]). 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 meta-parameter called the learning rate ($0 \leq \alpha \leq 1$). Tuning α will determine whether new reinforcement will drastically change the representation of action values, or if instead an action should be repeated several times before its value is significantly changed (see part 6 for further explanation).

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 [32]. 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:

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

where β is a third meta-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 so that there is a high probability of selecting an action whose action value is not the greatest (exploration). When β is high, the difference between action probabilities is increased so that the action with the highest action value is almost always selected (exploitation).

Clearly, these equations devoted to action value learning and action selection rely on crucial meta-parameters: α, β, γ. Most computational models use fixed meta-parameters, hand-tuned for a given task or problem. However, animals face a variety of tasks and deal with continuously varying conditions. If animal learning does rely on RL as suggested e.g. [45, 69] there must exist some brain mechanisms to decide, in each particular situations, which set of meta-parameters is appropriate (*e.g.* when an animal performs stereotypical behaviour in its nest, or repetitive food gathering behaviour 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 meta-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 meta-parameters has been called *meta-learning* by Kenji Doya. Meta-learning is a general principle which allows us 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 [79], upper-confidence bound policies 'UCB' which selects actions based on their associated reward averages and the number of times they were selected so far [6], EXP3-S for Exponential-weight algorithm for Exploration and Exploitation which is also based on a Boltzmann softmax function [14], uncertainty-based methods awarding bonuses to actions whose consequences are uncertain [19], and reviews of these methods applied to abruptly changing environments [23, 27].

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 economical terms refers to the risk (the known probability of a given reward source), and the volatility (the variance), across time of this risk.

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 [71, 80]. In this task, the agent has two possible actions (pressing one of two buttons). The task was 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 had 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 meta-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 shows a sample simulation. The agent learned the short-term condition, starting with a small meta-parameter beta (i.e. large exploration rate), 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, meta-parameters varied allowing more randomness in the agent's actions

(due to a small beta value), and leading the agent to focus on immediate reward (due to a small gamma value) which is more appropriate when the environment is unstable. After some time, the agent learns the new task condition and converges to a more exploitative behaviour (large beta value) combined with a more long-term oriented behavioural policy (large gamma value) appropriate for this new task condition.

This type of computational process appears suitably robust to account for animal behavioural 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 meta-parameters for learning and action selection [21].

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 behaviour. In order to describe how meta-learning can improve the functional descriptions of prefrontal areas, we will first present a short neurobiological overview.

3- ANATOMY, PHYSIOLOGY, AND FUNCTION OF PFC AREAS

The PFC is a large area of cortex, and there have been several attempts to subdivide it on both anatomical and functional lines. It seems clear now that the PFC has both an overall functional role that is not localised within its subdivisions, but also significant differences in function between those regions [84]. The prefrontal cortex's anatomical heterogeneity, observed in its local cytoarchitectonic organization and in the connectivity pattern of areas, reveals a functional heterogeneity (See chapter 1). PFC areas are highly interconnected but each seems contributing to specific functions of the prefrontal cortex as a whole [41]. One standard functional high order grouping of PFC areas defines lateral (LPFC), orbital, and medial subdivisions. The PFC is the target of multiple neuromodulatory afferents (including strong dopaminergic inputs), and it appears that impaired functioning of these systems results in numerous psychiatric and neurological disorders.

Several theories are proposed regarding the function of lateral prefrontal cortex (LPFC) [22, 26, 51, 56]. Most theories are based on the fact that LPFC neural activity participates in bridging cues and responses separated in time and space by actively representing task relevant information, i.e. information relative to targets, responses, and goals. Debates on functional dissociations within LPFC are intense, but most admit that active representation of information is a key feature of LPFC function. Active maintenance and the ability to link information across time delays are at the core of the role of LPFC in the control and sequential organization of behaviour. Although still under investigation, it has been proposed that the maintenance and control of information involve several mechanisms,

somewhat dependant on dopaminergic input, and related to recurrent excitation within LPFC, and between LPFC and distant areas (see reviews by [17, 51, 52]). The coding properties of LPFC tonic activity are modified between routine and non-routine or exploratory behaviours [57], suggesting a neurophysiological correlate of the cognitive control and modulation predicted by theories.

Crucial information required for action planning during adaptive behaviours is also encoded within LPFC activity. LPFC neurons encode information about the animal's responses as well as states of the environment [25, 83]. LPFC neurons represent the sequence of steps and state transitions that lead from the present to the desired goal [7, 54] which is reminiscent of goal-oriented action planning, also referred to as model-based reinforcement learning [18]. The quality and quantity of expected or obtained reward exert an influence on prefrontal delay activities [1, 42, 82]. Several lines of evidence suggest that the LPFC does not simply sum task-relevant information, but rather integrates reward-related information into knowledge about spatial location [39, 42]. Simultaneous information coding related to spatial location and reward takes place in this region as well as in the caudate nucleus [39]. Although spatial selectivity relates well to the role of LPFC in action selection, these hypotheses fail to provide a functional explanation for observed variations in spatial selectivity in LPFC. Spatial selectivity variations were observed depending on task phases and independent of the actual selection [58]. As we will see later, consistent with previous computational models describing the effect of average reward on variation in exploration rate within the LPFC [49], meta-learning principles enable good predictions of variations in spatial selectivity in LPFC between exploration and exploitation phases [38].

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 [3, 55]. ACC neuronal activity tracks task events and encode reinforcement-related information [3, 59, 75]. Muscimol injections in dorsal ACC induce strong deficits in finding the best behavioural option in a probabilistic learning task and in shifting responses based on reward changes [4, 75]. Dorsal ACC lesions induce failures in integrating reinforcement history to guide future choices [34]. These data converge toward describing a major role of ACC in integrating reward information over time, which is confirmed by single-unit recordings [72], 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 [65].

In addition, the ACC certainly has a related function in detecting and valuing unexpected but behaviourally 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 [2, 28]. 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 [9, 53, 70]. 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 [48, 62, 68]. Moreover, ACC neurons are able to discriminate choice errors (choice-related RPE) from execution errors (motor-related RPE, e.g. break of eye fixation), [62]. 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*.

A third important aspect of ACC function was revealed by the discovery of changes in neural activity between exploratory and exploitative trials [60, 62], or between volatile and stable rewarding schedules [10]. This suggests a more general involvement of ACC in translating results of performance monitoring and task monitoring into a regulatory level.

From this short review, clear functional dissociations appear between ACC and LPFC. However, we shall see later that a fine description of dissociations and interactions is required for a good functional description of these two regions.

4- DISSOCIATIONS AND INTERACTIONS BETWEEN ACC AND LPFC

Dissociations. Studies on ACC - LPFC co-activations in various cognitive tasks significantly helped to dissociate their specific roles or describe their interactions [35, 46, 50, 76]. An influential proposal is that ACC and LPFC are respectively involved in detection/monitoring of response conflict and in implementing cognitive control to cope with it [35, 46]. The dissociation is supported by evidence for correlations between sustained LPFC activation and the level of cognitive control on the one hand, and rapid changes in ACC activation during task practice on the other [50].

Overall, ACC appears to be important when a task requires behavioural adaptation. In an fMRI study involving task shifts, the ACC was active especially after cues that were informative regarding behavioural adaptation while LPFC was activated even after non informative cues [44]. Other fMRI studies pointed to a general role of ACC in assigning motivational priorities to task sets at any time, as opposed for a role for LPFC of dealing with interference arising from recently used task sets [31]. This last view is highly consistent with the theory according to which ACC has a major role in decision making by relating actions to their outcomes [67].

Comparative electrophysiological studies show a certain level of redundancy of coding and similar response patterns in ACC and LPFC, but also stress the complementary properties of activity from the two structures. For instance, differential activations related to reward encoding have been shown in ACC and LPFC [43]. In this study, ACC neurons encoded both reward and the behavioural response, while LPFC neurons mostly coded for the response. Matsumoto et al. have reported that ACC neurons were more likely to encode Response-Outcome associations, while LPFC neurons encoded Stimulus–Response associations [47]. Seo and Lee have shown, using a dynamic binary choice task, that more LPFC than ACC unit activity correlates with the difference between the reward values of two alternative choices.

That is, LPFC seems to indicate the best option to a greater degree, whereas there is more evidence in ACC for encoding reinforcement history [73]. Importantly, this study showed that both structures share some aspects of reinforcement-related computation. Overall these data converge toward a bias for ACC to encode performance monitoring signals, whereas LPFC neurons show a bias toward properties reflecting action selection. Note, however, that when one considers the overall properties of encoding by single units the dissociation is not absolute.

Interactions. While ACC and LPFC have been mainly highlighted in the literature in terms of their respective function, their contribution to cognitive control might be fully realized in their interaction. The typology and function of these interactions are still unclear and are the topic of ongoing investigations.

The study of the dynamic of conflict resolution appears to show correlated increases of ACC and LPFC activations in the face of conflict [8, 35]. In the context of the cognitive control loop scheme this has been interpreted as a sequential and directed involvement of ACC and LPFC in the response to and resolution of conflict (see below). However the occurrence of ACC-LPFC interaction only in situations involving conflict resolution is debated [31]. Koechlin and colleagues have instead 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 [40].

By means of electrophysiological recordings in the monkey, Tsujimoto and colleagues have shown synchronous local field potentials (LFP) between areas 9 and 32, homologous to subparts of LPFC and ACC regions in humans, during a variety of cognitive tasks [81]. Similar results have been found in EEG in humans with the Eriksen-Flanker task [13] where oscillatory activity in the theta band (4-8 Hz) in the medial prefrontal cortex was enhanced after errors, associated to a transient synchronization with the LPFC, and followed by a behavioural adjustment. Gehring and Knight showed that in patients with a lesion in LPFC, the well studied medial frontal error-related negative potential, putatively produced from ACC, was still present but did not discriminate between errors and correct trials anymore [24]. At the behavioural level the same patients showed difficulties in adapting responses following errors. These data question the direction of influence between ACC and LFPC, although the effect could also be explained by an increased detection of response conflicts under abnormal cognitive control [16].

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 [33], and by EEG studies in human [76]. The former study showed that the effect of task switching appear earlier in ACC than in LFPC [33]. 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. 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 both for unit activity and high gamma power of LFP (see figure 2).

5- PFC AND THE REGULATION OF COGNITIVE CONTROL

The functions of prefrontal areas have been widely studied within a framework that strongly echoes meta-learning principles: the cognitive control loop theory [11, 15]. The cognitive control loop describes the modulation of the control level in order to adapt to immediate needs imposed by the environment. It also enables a shift from routine behaviours in a known context requiring little attention and concentration, to more flexible behaviours involving rapid and active control. Two main phases are necessary for the regulation and implementation of cognitive control. The first consists of the systematic detection and evaluation of the relevance of performed actions. This information is used, in a second phase, to regulate cognitive control and to potentiate appropriate action selection to reach a particular goal. Norman and Shallice had formalized such a system with two components: an entity for automatic action selection and an attentional supervisory/control system [74]. The more familiar the environment and the more stably rewarded the actions are, the more the system tends towards automatic functioning. In contrast, complex situations impose an active recovery of control to deal with new contexts and select appropriate actions. Botvinick and colleagues followed the same perspective by proposing conflict detection as a central mechanism to regulate cognitive control[11].

The neural substrates proposed to support the mechanisms described by these theories largely involve interactions between prefrontal areas [11, 51]. Particularly, the medial prefrontal cortex, including the ACC for its role in performance and task monitoring, and the LPFC for its role in action planning and in the implementation of cognitive control.

Several computational models have been developed describing functioning of the cognitive control loop, and explicitly referring to ACC and LPFC. The respective roles attributed to these two structures are always based around the canonical view that ACC processes errors and monitors performance to regulate control in LPFC where action selection is implemented. The « global workspace » model described by Dehaene and colleagues in 1998, explains how control is resumed in situations where routines get interrupted, where errors are made or where an environmental change is detected. This model is composed of separate specialized modules regulated by a global workspace. The postulate being that, considering functions attributed to ACC and LPFC, these two structures are perfectly appropriate to accomplish regulation [20].

Cohen and colleagues describe an auto-regulated system responding to the demands of control by adjusting the exploration-exploitation trade-off [15]. In this model, ACC detects action consequences and attributes them a value. This information is then used to modulate the cognitive control rate in LPFC via the locus coeruleus. Their model explicitly mentions

noradrenergic innervation of the LPFC as a possible intermediate substrate to translate ACC modulation. The resulting increased cognitive control will facilitate behavioural adaptation in an appropriate way in associative areas. In a later version of the model, ACC is dedicated to conflict monitoring while the orbitofrontal cortex (OFC) monitors performance by estimating the reward average [49]. ACC and OFC would exert a systematic regulation of LC, which in turn modulates the level of exploration in the LPFC. Although the model fails to take into account ACC's involvement in reinforcement learning mechanisms such as reward prediction error signalling and action value updating, it has the merit of implicitly echoing the meta-learning framework by proposing a regulation of exploration based on reward average. Moreover, the model proposes a neural implementation of exploration regulation by varying the contrast between neural activities associated with competing actions, similar to the effect of the Boltzmann softmax function presented earlier (Eq. 3). Our recent work using the meta-learning framework helped reconcile and integrate ACC mechanisms related to reinforcement learning and mechanisms related to performance monitoring. Moreover, as described in the next paragraph, it predicted formal variations of influence on LPFC mediated by ACC functions which could be verified by simultaneous recordings of the two structures [38].

6- NEURAL CORRELATES OF META-PARAMETERS REGULATION

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 meta-parameters: the learning rate α [10]. Their study is grounded on theoretical accounts suggesting that feedback information from the environment does not 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 behavioural 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 [10]. 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 meta-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 [48, 62]. This hypothesis can also explain differences in the time window over which previous reward information is encoded in ACC and related structures as measured in different protocols involving different volatilities: a low learning rate produces a slow integration of reward information and thus preserves previous reward information over a

large time window. In contrast, a high learning rate quickly erases information about previous rewards. Consistently with this in Sugrue et al. reward contingencies remained stable for hundreds of trials, which allowed outcomes from more than 30 trials ago to still have some influence over the values of choice options [77]. In Kennerley et al. (2006), the monkeys experienced a more volatile environment that switched approximately every 25 trials[34]. As a consequence, a much shorter reward integration period was reported in this study. In an adaptation of the matching pennies game Seo and Lee showed that monkey's choice in a given trial was potentially influenced by the choice outcomes in multiple previous trials as expressed by a slow updating (low learning rate $\alpha = 0.24$) of action value functions, and ACC unit activity reflected the persistence of reward information across trials [72].

Quilodran et al. (2008) used a very volatile environment (problem solving task, PST) where the action reward contingency could be obtained from one single outcome and where the task rule shifted after fewer than 10 trials on average[62]. This task enabled us to clearly dissociate exploratory and exploitative trials. Animals had to find which target presented in a set of 4 is rewarded. In each block (problem) the animal can explore targets until discovering the rewarded one, and then exploit (repeat) its choice for at least 4 trials. The target was then changed to initiate a new problem. This produced a complete reset of monkeys' action values at each new problem, independent of the previous problem [37]. Consistent with the theoretical relationship between volatility and learning rate, we found that monkey behaviour in the PST fit the best with a reinforcement learning model using a very high learning rate ($\alpha = 0.9$). In this task the learning rate is not expected to change over time, which implies a high but stable volatility. However, the exploratory rate should be varied to optimally regulate decision stochasticity. Previous recordings of either ACC or LFPC neurons in this task revealed strong firing rate variations between exploratory (uncertain) trials and repetitive trials [58, 62]. Recent investigations done in our laboratory using the PST showed neural correlates of regulation of meta-parameter β (i.e. exploration rate) using recordings from both ACC and LPFC. We developed a computational model providing a formal description of ACC-LPFC interactions so as to be able to draw experimental predictions [37]. The model integrates ACC's role in adapting action values based on dopaminergic reward-prediction errors and reward history [28, 66], its function in performance monitoring through feedback categorization mechanisms [62], and its role in regulating LPFC's function [5]. Finally, we integrated Cohen and Aston-Jones' proposal that the exploration rate is regulated within the LPFC based on performance monitoring [15, 49]. To do so, our LPFC part filters action values sent by the ACC with equation (3) (see paragraph 2) where β is regulated by feedback history measured in ACC (Fig. 3A). Simulation of the model led to a set of experimental predictions that were verified by preliminary analyses of recordings from ACC and LPFC in the PST: (1) an overall decrease of activity during repetition trials was observed only in the LPFC; (2) target selectivity was globally higher in LPFC than in ACC; (3) an increase of target selectivity was observed during repetition trials, consistent with the hypothesized exploitative mode of the system (Fig. 3B). Analysis of single-unit activity in this protocol also revealed correlates of

information related to different variables in the model and confirmed the hypothesized function of ACC-LPFC interactions in this task.

7- CONCLUSION

The cognitive control theory has previously stressed the importance of performance monitoring and task monitoring in the ACC to regulate the level of control within the LPFC. It appears that the meta-learning framework can complete this picture by providing testable computational principles that could formally underlie the regulation of such control. This framework explains the finding of a diversity of performance monitoring processes previously associated with ACC function, such as estimations of error-likelihood [12]. It also supports the previously highlighted fundamental role of this structure in relating actions to outcomes [67]. Recent investigations have explicitly referred to the involvement of the ACC in the regulation of reinforcement learning meta-parameters [10]. This and our studies suggest that ACC might contribute to adapting the learning rate based on estimations of the environment's volatility, and the exploration rate based on feedback history and reward average.

However, the current picture drawn from the dissociation between ACC and LPFC function is not yet complete. Recent findings including our own analyses, suggest that function is somewhat distributed over ACC and LPFC and that they both contain information related to action valuation, action selection, and their regulation. Also, while most theoretical approaches focussed on ACC's influence over the LPFC, the opposite has to be considered. As mentioned above, Gehring and Knight showed that in patients with LPFC lesion, the medial frontal error-related negative potential, associated to ACC, was still present but did not discriminate between errors and correct trials anymore [24]. Moreover, anatomical data collected in our laboratory suggest that anatomical connections in both directions exist and have different patterns suggesting different functional effects [61] (see also Chapter 1). Thus information flow from LPFC to ACC appears important and has to be taken into account to better understand ACC-LPFC interactions. Further investigations will be required to understand how ACC and LPFC share information, interact, and still show dissociable contributions to specific functions. The combinations of neurophysiological, interruptive, and computational approaches will be essential to answer such complex questions.

REFERENCES

1. Amemori K, Sawaguchi T (2006) Contrasting effects of reward expectation on sensory and motor memories in primate prefrontal neurons. Cereb Cortex 16:1002-1015.
2. Amiez C, Joseph JP, Procyk E (2005) Anterior cingulate error-related activity is modulated by predicted reward. Eur J Neurosci 21:3447-3452.
3. Amiez C, Joseph JP, Procyk E (2005) 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.
4. Amiez C, Joseph JP, Procyk E (2006) Reward encoding in the monkey anterior cingulate cortex. Cereb Cortex 16:1040-1055.
5. Aston-Jones G, Cohen JD (2005) An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annu Rev Neurosci 28:403-450.
6. Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit. Machine Learning 47:235-256.
7. Averbeck BB, Sohn JW, Lee D (2006) Activity in prefrontal cortex during dynamic selection of action sequences. Nat Neurosci 9:276-282.
8. Badre D, Wagner AD (2004) Selection, integration, and conflict monitoring; assessing the nature and generality of prefrontal cognitive control mechanisms. Neuron 41:473-487.
9. Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47:129-141.
10. Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. Nat Neurosci 10:1214-1221.
11. Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD (2001) Conflict monitoring and cognitive control. Psychol Rev 108:624-652.
12. Brown JW, Braver TS (2005) Learned predictions of error likelihood in the anterior cingulate cortex. Science 307:1118-1121.
13. Cavanagh JF, Cohen MX, Allen JJ (2009) Prelude to and resolution of an error: EEG phase synchrony reveals cognitive control dynamics during action monitoring. J Neurosci 29:98-105.
14. Cesa-Bianchi N, Gabor L, Stoltz G (2006) Regret minimization under partial monitoring. MathOper Res 31.
15. Cohen JD, Aston-Jones G, Gilzenrat MS (2004) A systems-level perspective on attention and cognitive control. In: Cognitive Neuroscience of attention (Posner MI, ed), pp 71-90. New York: Guilford.
16. Cohen JD, Botvinick M, Carter CS (2000) Anterior cingulate and prefrontal cortex: who's in control? Nat Neurosci 3:421-423.
17. Constantinidis C, Procyk E (2004) The primate working memory networks. Cogn Affect Behav Neurosci 4:444-465.
18. Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8:1704-1711.
19. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876-879.
20. Dehaene S, Kerszberg M, Changeux JP (1998) A neuronal model of a global workspace in effortful cognitive tasks. Proc Natl Acad Sci U S A 95:14529-14534.
21. Doya K (2002) Metalearning and neuromodulation. Neural Netw 15:495-506.
22. Fuster JM (1997) The prefrontal cortex. Anatomy, physiology and neuropsychology of the frontal lobe, 3rd Edition: Lippincott-Raven.
23. Garivier A, Moulines E (2008) On upper-confidence bound policies for nonstationary bandit problems. Arxiv preprint arXiv:0805.3415.
24. Gehring WJ, Knight RT (2000) Prefrontal-cingulate interactions in action monitoring. Nat Neurosci 3:516-520.

25. Genovesio A, Brasted PJ, Wise SP (2006) Representation of future and previous spatial goals by separate neural populations in prefrontal cortex. J Neurosci 26:7305-7316.
26. Goldman-Rakic PS (1987) Circuity of primate prefrontal cortex and regulation of behavior by representational memory. In: Higher functions of the brain (Plum F, ed), pp 373-414. Bethesda: American physiological society.
27. 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.
28. 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.
29. 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.
30. 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.
31. Hyafil A, Summerfield C, Koechlin E (2009) Two mechanisms for task switching in the prefrontal cortex. J Neurosci 29:5135-5142.
32. Ishii S, Yoshida W, Yoshimoto J (2002) Control of exploitation-exploration meta-parameter in reinforcement learning. Neural Netw 15:665-687.
33. 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.
34. Kennerley SW, Walton ME, Behrens TE, Buckley MJ, Rushworth MF (2006) Optimal decision making and the anterior cingulate cortex. Nat Neurosci 9:940-947.
35. Kerns JG, Cohen JD, MacDonald AW, 3rd, Cho RY, Stenger VA, Carter CS (2004) Anterior cingulate conflict monitoring and adjustments in control. Science 303:1023-1026.
36. 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.
37. Khamassi M, Quilodran R, Enel P, Procyk E, Dominey PF (2010) A computational model of integration between reinforcement learning and task monitoring in the prefrontal cortex. In: Simulation of Adaptive Behaviour. Paris: Springer.
38. Khamassi M, Quilodran R, Procyk E, Dominey PF (2009) Anterior cingulate cortex integrates reinforcement learning and task monitoring. In: Society For Neuroscience 39th annual meeting. Chicago, IL.
39. Kobayashi S, Kawagoe R, Takikawa Y, Koizumi M, Sakagami M, Hikosaka O (2007) Functional differences between macaque prefrontal cortex and caudate nucleus during eye movements with and without reward. Exp Brain Res 176:341-355.
40. Kouneiher F, Charron S, Koechlin E (2009) Motivation and cognitive control in the human prefrontal cortex. Nat Neurosci 12:939-945.
41. Lee D, Rushworth MF, Walton ME, Watanabe M, Sakagami M (2007) Functional specialization of the primate frontal cortex during decision making. J Neurosci 27:8170-8173.
42. Leon MI, Shadlen MN (1999) Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. Neuron 24:415-425.
43. Luk CH, Wallis JD (2009) Dynamic encoding of responses and outcomes by neurons in medial prefrontal cortex. J Neurosci 29:7526-7539.
44. Luks TL, Simpson GV, Feiwell RJ, Miller WL (2002) Evidence for anterior cingulate cortex involvement in monitoring preparatory attentional set. Neuroimage 17:792-802.
45. 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.
46. 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.

47. Matsumoto K, Suzuki W, Tanaka K (2003) Neuronal correlates of goal-based motor selection in the prefrontal cortex. Science 301:229-232.
48. Matsumoto M, Matsumoto K, Abe H, Tanaka K (2007) Medial prefrontal cell activity signaling prediction errors of action values. Nat Neurosci 10:647-656.
49. McClure SM, Gilzenrat MS, Cohen JD (2006) An exploration–exploitation model based on norepinephrine and dopamine activity. In: Advances in neural information processing systems (Weiss Y, Sholkopf B, Platt J, eds), pp 867–874: MIT Press, Cambridge, MA.
50. Milham MP, Banich MT, Claus ED, Cohen NJ (2003) Practice-related effects demonstrate complementary roles of anterior cingulate and prefrontal cortices in attentional control. Neuroimage 18:483-493.
51. Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167-202.
52. Montague PR, Hyman SE, Cohen JD (2004) Computational roles for dopamine in behavioural control. Nature 431:760-767.
53. Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H (2006) Midbrain dopamine neurons encode decisions for future action. Nat Neurosci 9:1057-1063.
54. Mushiake H, Saito N, Sakamoto K, Itoyama Y, Tanji J (2006) Activity in the lateral prefrontal cortex reflects multiple steps of future events in action plans. Neuron 50:631-641.
55. Paus T (2001) Primate anterior cingulate cortex: where motor control, drive and cognition interface. Nat Rev Neurosci 2:417-424.
56. Petrides M (1998) Specialized systems for the processing of mnemonic information within the primate frontal cortex. In: The prefrontal cortex. Executive and cognitive functions (Roberts AC, Robbins TW, Weiskrantz L, eds), pp 103-116. Oxford: Oxford University press.
57. Procyk E, Gao WJ, Goldman-Rakic PS (2001) prefrontal unit activity during delayed response and self-initiated performance. In: society for neuroscience, p 533.532. San-Diego.
58. Procyk E, Goldman-Rakic PS (2006) Modulation of dorsolateral prefrontal delay activity during self-organized behavior. J Neurosci 26:11313-11323.
59. Procyk E, Joseph JP (2001) Characterization of serial order encoding in the monkey anterior cingulate sulcus. Eur J Neurosci 14:1041-1046.
60. Procyk E, Tanaka YL, Joseph JP (2000) Anterior cingulate activity during routine and non-routine sequential behaviors in macaques. Nat Neurosci 3:502-508.
61. Quilodran R (2009) Réseaux corticaux préfrontaux et adaptation du comportement: physiologie et anatomie quantitative chez le singe. In: PhD thesis Lyon: Université Claude Bernard Lyon I.
62. Quilodran R, Rothé M, Procyk E (2008) Behavioral shifts and action valuation in the anterior cingulate cortex. Neuron 57(2):314–325.
63. Reynolds JN, Hyland BI, Wickens JR (2001) A cellular mechanism of reward-related learning. Nature 413:67-70.
64. Robbins TW (1998) Dissociating executive functions of the prefrontal cortex. In: The prefrontal cortex. Executive and cognitive functions (Roberts AC, Robbins TW, Weiskrantz L, eds), pp 117-130. New York: Oxford University Press.
65. 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.
66. Rushworth MF, Behrens TE, Rudebeck PH, Walton ME (2007) Contrasting roles for cingulate and orbitofrontal cortex in decisions and social behaviour. Trends Cogn Sci.
67. Rushworth MF, Walton ME, Kennerley SW, Bannerman DM (2004) Action sets and decisions in the medial frontal cortex. Trends Cogn Sci 8:410-417.
68. 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.
69. Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. Science 310:1337-1340.

70. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593-1599.
71. Schweighofer N, Doya K (2003) Meta-learning in reinforcement learning. Neural Netw 16:5-9.
72. 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.
73. Seo H, Lee D (2008) Cortical mechanisms for reinforcement learning in competitive games. Philos Trans R Soc Lond B Biol Sci 363:3845-3857.
74. Shallice T (1988) From neuropsychology to mental structure: Cambridge Univ Press.
75. Shima K, Tanji J (1998) Role for cingulate motor area cells in voluntary movement selection based on reward. Science 282:1335-1338.
76. 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.
77. Sugrue LP, Corrado GS, Newsome WT (2004) Matching behavior and the representation of value in the parietal cortex. Science 304:1782-1787.
78. 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.
79. Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, MA
London, England: MIT Press.
80. 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.
81. Tsujimoto T, Shimazu H, Isomura Y, Sasaki K (2010) Theta oscillations in primate prefrontal and anterior cingulate cortices in forewarned reaction time tasks. J Neurophysiol 103:827-843.
82. Watanabe M (1996) Reward expectancy in primate prefrontal neurons. Nature 382:629-632.
83. Watanabe M, Sakagami M (2007) Integration of cognitive and motivational context information in the primate prefrontal cortex. Cereb Cortex 17 Suppl 1:i101-109.
84. Wilson CR, Gaffan D, Browning PG, Baxter MG (2010) Functional localization within the prefrontal cortex: missing the forest for the trees? Trends Neurosci.

FIGURE CAPTIONS

Figure 1. Simulation of a meta-learning algorithm. Adapted from [71]. A change in the task condition from short-term reward to long-term reward at timestep #200 produces an adaptation of meta-parameters' values.

Figure 2. Latencies of neural responses after feedback in ACC and LPFC. A. From unit activity recorded in the PST task[62]: neurons selective to incorrect feedbacks (INC, after an incorrect choice) discharge at comparable latencies in ACC and LPFC (black curves). However, neurons responding to salient feedbacks (first correct CO1 and INC feedbacks after incorrect choice and after the first reward delivery; grey) have a shorter latency in ACC than in LPFC. **B.** Latencies of significant high gamma power increase in LFP after incorrect feedbacks in ACC (black) and LPFC (grey).

Figure 3. A. Theoretical scheme of the hypothesized respective roles of ACC and LPFC in action value learning and exploration regulation (β*) in the PST task. The interaction of these structures with the striatum through cortico-basal ganglia-thalamo-cortical anatomical loops is not represented here. **B.** Global physiological tendencies measured in ACC and LPFC consistent with the theoretical scheme.

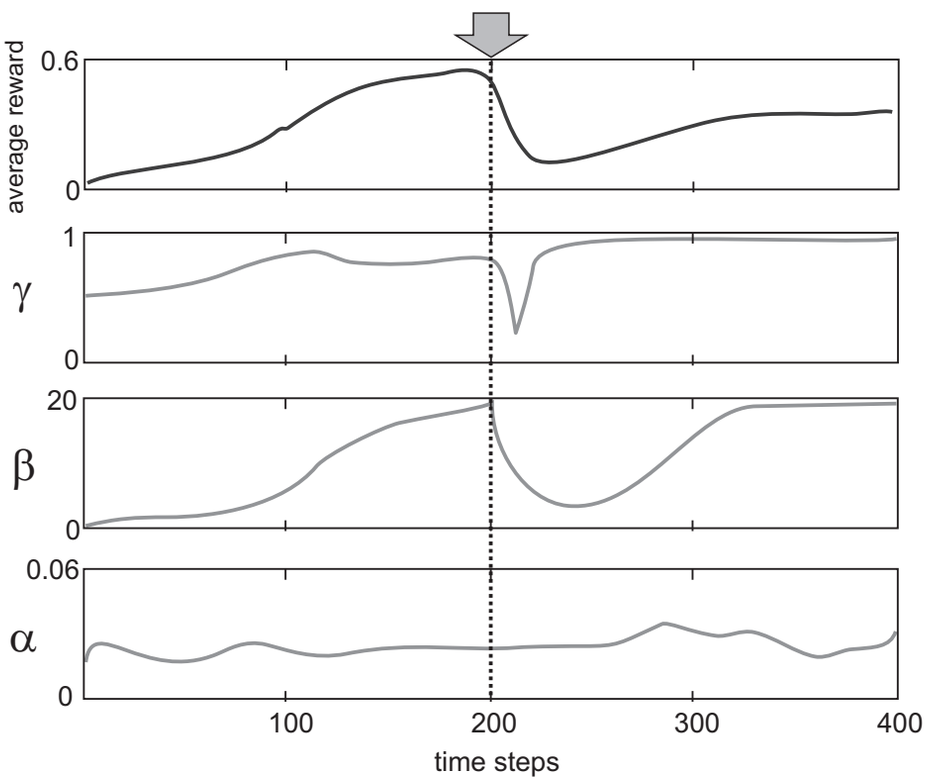


Figure 1

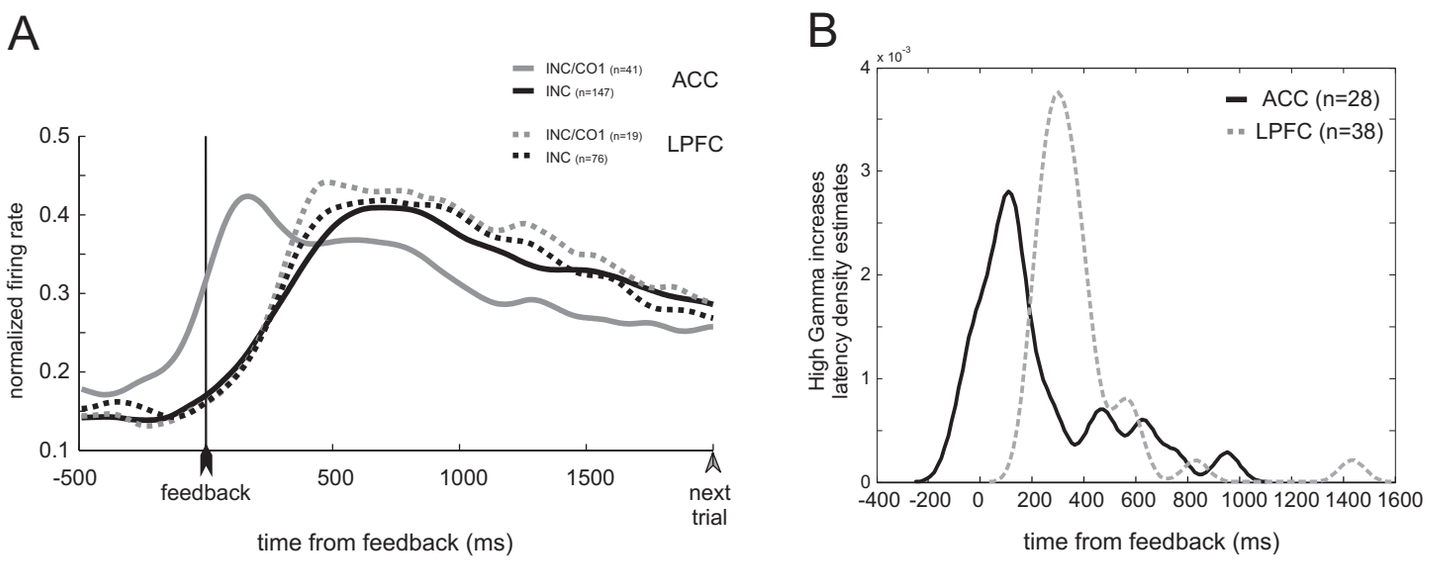
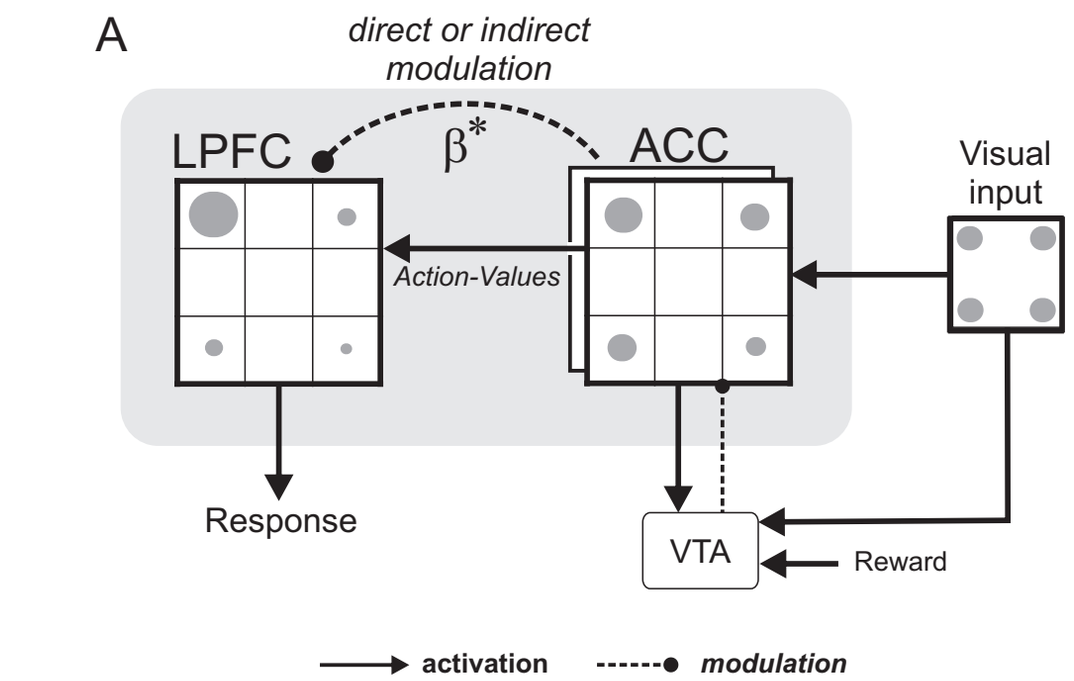

A
INC/CO1 (n=41)
INC (n=147)
ACC
INC/CO1 (n=19)
INC (n=76)
LPFC
normalized firing rate
0.5
0.4
0.3
0.2
0.1
-500
feedback
500
1000
1500
next trial
time from feedback (ms)
B
x 10-3
4
3
2
1
0
ACC (n=28)
LPFC (n=38)
High Gamma increases latency density estimates
-400 -200 0 200 400 600 800 1000 1200 1400 1600
time from feedback (ms)

Figure 2


A
direct or indirect
modulation
LPFC
β*
ACC
Visual
input
Action-Values
Response
VTA
Reward
activation
modulation

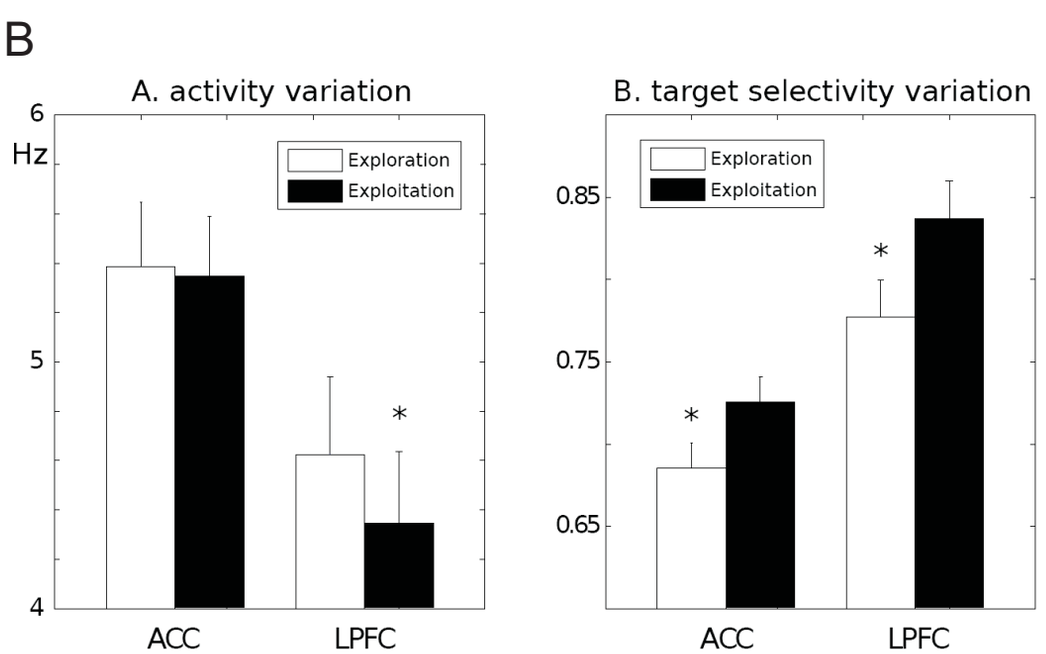

B
A. activity variation
Hz
6
5
4
Exploration
Exploitation
*
ACC
LPFC
B. target selectivity variation
0.85
0.75
0.65
Exploration
Exploitation
*
*
ACC
LPFC

Figure 3