frontiers in
NEUROROBOTICS

ORIGINAL RESEARCH ARTICLE
published: 12 July 2011
doi: 10.3389/fnbot.2011.00001

Robot cognitive control with a neurophysiologically inspired reinforcement learning model

Mehdi Khamassi[1,2,3,4]*, Stéphane Lallée[1,2], Pierre Enel[1,2], Emmanuel Procyk[1,2] and Peter F. Dominey[1,2]

[1] Stem Cell and Brain Research Institute, INSERM U846, Bron, France
[2] UMR-S 846, Université de Lyon 1, Lyon, France
[3] Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie – Paris 6, Paris, France
[4] CNRS UMR 7222, Paris, France

Edited by:
Jeffrey L. Krichmar, University of California Irvine, USA

Reviewed by:
Jeffrey L. Krichmar, University of California Irvine, USA
Jason G. Fleischer, The Neuroscience Institute, USA

***Correspondence:**
Mehdi Khamassi, UPMC - ISIR UMR 7222, Boîte courrier 173, 4 place Jussieu, 75005 Paris, France.
e-mail: mehdi.khamassi@isir.upmc.fr

A major challenge in modern robotics is to liberate robots from controlled industrial settings, and allow them to interact with humans and changing environments in the real-world. The current research attempts to determine if a neurophysiologically motivated model of cortical function in the primate can help to address this challenge. Primates are endowed with cognitive systems that allow them to maximize the feedback from their environment by learning the values of actions in diverse situations and by adjusting their behavioral parameters (i.e., cognitive control) to accommodate unexpected events. In such contexts uncertainty can arise from at least two distinct sources – expected uncertainty resulting from noise during sensory-motor interaction in a known context, and unexpected uncertainty resulting from the changing probabilistic structure of the environment. However, it is not clear how neurophysiological mechanisms of reinforcement learning and cognitive control integrate in the brain to produce efficient behavior. Based on primate neuroanatomy and neurophysiology, we propose a novel computational model for the interaction between lateral prefrontal and anterior cingulate cortex reconciling previous models dedicated to these two functions. We deployed the model in two robots and demonstrate that, based on adaptive regulation of a meta-parameter β that controls the exploration rate, the model can robustly deal with the two kinds of uncertainties in the real-world. In addition the model could reproduce monkey behavioral performance and neurophysiological data in two problem-solving tasks. A last experiment extends this to human–robot interaction with the iCub humanoid, and novel sources of uncertainty corresponding to "cheating" by the human. The combined results provide concrete evidence for the ability of neurophysiologically inspired cognitive systems to control advanced robots in the real-world.

Keywords: iCub, humanoid robot, reinforcement learning, meta-learning, bio-inspiration, prefrontal cortex

INTRODUCTION

In controlled environments (e.g., industrial applications), robots can achieve performance superior in speed and precision to humans. When faced with limited uncertainty that can be characterized *a priori*, we can provide robots with computational techniques such as finite state machines that can address such expected uncertainty. But in the real-world, robots face unexpected uncertainty – such as new constraints or new objects in a task – and need to be robust to variability in the world.

Exploiting knowledge of primate neuroscience can help in the design of cognitive systems enabling robots to adapt to varying task conditions and to have satisfying, if not optimal, performance, in a variety of different situations (Pfeifer et al., 2007; Arbib et al., 2008; Meyer and Guillot, 2008).

We have previously characterized the functional neurophysiology of the prefrontal cortex as playing a central role in the organization of complex cognitive behavior (Amiez et al., 2006; Procyk and Goldman-Rakic, 2006; Quilodran et al., 2008). The goal of the current research is to test the hypothesis that indeed, a model based on this architecture can be used to control complex robots that rely on potentially noisy perceptual–motor systems.

Recent advances in the neurophysiological mechanisms of decision-making have highlighted the role of the prefrontal cortex, particularly the anterior cingulate cortex (ACC) and dorsolateral prefrontal cortex (LPFC), in flexible behavioral adaptation by learning action values based on rewards obtained from the environment, and adjusting behavioral parameters to varying uncertainties in the current task or context (Miller and Cohen, 2001; Koechlin and Summerfield, 2007; Rushworth and Behrens, 2008; see Khamassi et al., in press for a review). Both the ACC and LPFC appear to play crucial roles in these processes. They both receive inputs from dopamine neurons which are known to encode a reward prediction error coherent with reinforcement learning (RL) principles (Schultz et al., 1997). The LPFC is involved in action selection and planning. The ACC is known to monitor feedback as well as the task and is considered to modulate or "energize" the LPFC based on the motivational state (Kouneiher et al., 2009).

However, there is a contradiction between current models of the ACC–LPFC system, which are either dedicated to reward-based RL functions (Holroyd and Coles, 2002; Matsumoto et al., 2007) or are focused on the regulation of behavioral parameters by means of conflict monitoring and cognitive control (Botvinick et al., 2001; Cohen et al., 2004). Here we propose a novel computational model reconciling these two types of processes, and show that it can reproduce monkey behavior in dealing with uncertainty in a variety of behavioral tasks. The system relies on RL principles allowing an agent to adapt its behavioral policy by trial-and-error so as to maximize reward (Sutton and Barto, 1998). Based on previous neurophysiological data, we make the assumption that action values are learned and stored in the ACC through dopaminergic input (Holroyd and Coles, 2002; Amiez et al., 2005; Matsumoto et al., 2007; Rushworth et al., 2007). These values are transmitted to the LPFC which selects the action to perform. In addition, the model keeps track of the agent's performance and the variability of the environment to adjust behavioral parameters. Thus the ACC component monitors 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). The adjustment of behavioral parameters based on such outcome history follows meta-learning principles (Doya, 2002) and is here restricted to the tuning of the β meta-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 (Yu and Dayan, 2005; Daw et al., 2006). The resulting meta-parameter modulates action selection within the LPFC, consistent with its involvement in the exploration–exploitation trade-off (Daw et al., 2006; McClure et al., 2006; Cohen et al., 2007; Frank et al., 2009).

The model was tested on two robot platforms to: (1) show its ability to robustly perform and adapt under different conditions of uncertainty in the real-world during various neurophysiologically tested problem-solving (PS) tasks combining reward-based learning and alternation between exploration and exploitation periods (Amiez et al., 2006; Quilodran et al., 2008); (2) reproduce monkey behavioral performance by comparing the robot's behavior with previously published and new monkey behavioral data; (3) reproduce global properties of previously shown neurophysiological activities during these tasks.

The PS tasks used here involve a set of problems where the robot should select one of a set of targets on a touch screen. Each problem is decomposed into search (exploration) trials where the robot identifies the rewarded target, and exploitation trials where the robot then repeats its choice of the "best" target. We will see that the robot solved the task with performance similar to that of monkeys. It properly adapted to perceptual uncertainties and alternated between exploration and exploitation.

We then generalized the model to a human–robot interaction scenario where unexpected uncertainties are introduced by the human introducing cued task changes or by cheating. By correctly performing and autonomously learning to reset exploration in response to such uncertain cues and events, we demonstrate that neurophysiologically inspired cognitive systems can control advanced robotic systems in the real-world. In addition, the model's learning mechanisms that were challenged in the last scenario provide testable predictions on the way monkeys may learn the structure of the task during the pre-training phase of Experiments 1 and 2.

MATERIALS AND METHODS

GLOBAL ROBOTICS SETUP

In each experiment presented in this paper, we consider a humanoid agent – a physical robot or a simulation – which interacts with the environment through visual perception and motor commands. The agent perceives objects or geometrical features (i.e., cubes on a table or targets on a screen) via a camera-based vision system described below. The agent is required to choose one of the objects with the objective of obtaining a reward. The reward is a specific visual signal (i.e., a triangle presented on a screen) supposed to represent the juice reward obtained by monkeys during these experiments. For simplicity, perception of the reward signal is hardcoded to trigger an internal scalar reward signal in the computational model controlling the robot. Thus all external inputs are provided to the robot through vision. Experiments 1 and 2 are inspired by our previous monkey neurophysiology experiments (Amiez et al., 2006; Quilodran et al., 2008). They involve interaction with a touch-sensitive screen (IIyama Vision Master Pro 500) where different square targets appear. The agent should search for and find the target with the highest reward value by touching it on the screen (**Figure 1**). Experiment 3 extends monkey experiments to a simple scenario of human–robot interaction that involves a set of cubes on a table. A human is sitting near the table, in front of the robot, and shuffles the cubes. The robot has to find the cube with a circle on its hidden face, corresponding to the reward.

GLOBAL STRUCTURE OF THE EXPERIMENTS

The three experiments have the same temporal structure. Here we describe the details of this structure, and then provide the specifics for each experiment.

All experiments are composed of a set of problems where the agent should search by trial-and-error in order to find the most rewarding object among a proposed ensemble. Each problem is decomposed into search (exploration) trials where the agent explores different alternatives until finding the best object, and repetition (exploitation) trials where the agent is required to repeat choice of the best object several times (**Figure 2**). After the repetition, a problem-changing cue (PCC) signal is shown to the agent to indicate that a new problem will start. In 90% of the new problems the identity of the best object is changed. In Experiments 1 and 2, the PCC signal is known *a priori*. Experiment 3 tests the flexibility of the system, as the PCC is learned by the agent. Experiment 1 is deterministic (only one object is rewarded while the others are not). Experiment 2 is probabilistic (each object has a certain probability of association with reward) and thus tests the ability of the system to accommodate such probabilistic conditions.

EXPERIMENT 1

The first experiment is inspired by our previous neurophysiological research described in (Quilodran et al., 2008). Four square targets are presented on the touch screen (see **Figure 2**). At each

FIGURE 1 | Lynxmotion SES robotic arm in front of a touch screen used for Experiment 1. The screen is perceived by a webcam. The arm has a gripper with a sponge surrounded by aluminum connected to the ground. This produces a static current when contacting the screen and enables the screen to detect when and where the robot touches it. This setup allows us to test the robot in the same experimental conditions as the non-human primate subjects in our previous studies (Amiez et al., 2006; Quilodran et al., 2008).

problem, a single target is associated with reward with a probability of one (deterministic). At each trial, the four targets appear on the screen and remain visible during a 5-s delay. The robotic arm should touch one of the targets before the end of the delay. Once a touch is detected on the screen, the targets disappear and the choice is evaluated. If the correct target is chosen, a triangle appears on the screen, symbolizing the juice reward monkeys obtain. For incorrect choices, the screen remains black for another 5-s delay and then a new search trial starts. Once the correct target is chosen through a process of trial-and-error search, a repetition phase follows, lasting until the robot performs three correct responses, no matter how many errors it made. At the end of the repetition phase, a circle appears on the screen, indicating the end of the current problem, and the start of a new one. Similarly to monkey experiments, in about 90% cases, the correct target is different between two consecutive problems, requiring a behavioral shift and a new exploration phase.

EXPERIMENT 2

Experiment 1 tests whether the model can be used under deterministic conditions, but leaves open the question as to whether it can successfully perform under a probabilistic reward distribution. Experiment 2 allows us to test the functioning of the model in such probabilistic conditions, directly inspired by our neurophysiological research described in (Amiez et al., 2006). In contrast with Experiment 1, the agent can choose only between two targets. In each problem, one target has a high probability (0.7) of producing a large reward and a low probability (0.3) of producing a small one. The other target has the opposite distribution (**Table 1**). Problems in this task are also decomposed in search and repetition trials. However, in contrast to Experiment 1, there is no sharp change between search and repetition phases. Instead, trials are *a posteriori* categorized as repetition trials, as follows. Each problem continues until the agent makes five consecutive choices of the best target, followed by selection of the same target for the next five trials or five of the next six trials. However, if after 50 trials the monkey has not entered the repetition phase, the current problem is aborted and considered unsuccessful. Similarly to Experiment 1, the end of each problem is cued by a PCC indicating a 90% probability of change in reward distribution among targets.

EXPERIMENT 3

The third experiment constitutes an extension of Experiment 1 to a simple human–robot interaction scenario. The experiment is

Table 1 | Reward probabilities used in Experiment 2.

Amount of "juice" dispensed as reward	Target A	Target B
1.2 mL	0.7	0.3
0.4 mL	0.3	0.7

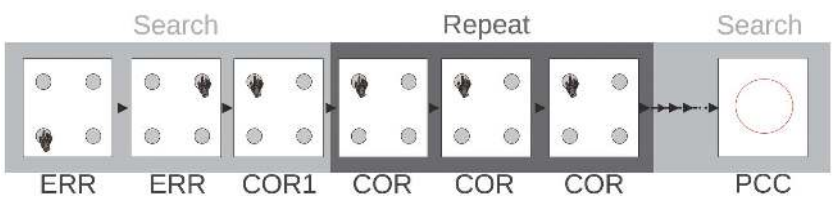


FIGURE 2 | Task used in Experiment 1. Four targets appear on the screen. Only one is associated with reward. The robot searches for the correct target. When the correct target is found, three repetitions of the correct choice are required before a problem-changing cue (PCC) appears. ERR, error; COR1, first correct trial; COR, subsequent correct trials.

performed with the iCub, a humanoid robot developed as part of the RobotCub project (Tsagarakis et al., 2007). The task performed by the iCub robot is illustrated in **Figure 3** and its temporal structure is described in **Figure 4**. In this task, four cubes are lying on a table. One of the cubes has a circle on its hidden face, indicating a reward. The human can periodically hide the cubes with a wooden board (**Figure 4D**) and change the position of the rewarding cube. This mimics the PCC used in the previous experiments. The difference here is that the model has to autonomously learn that presentation of the wooden board is always followed by a change in condition, and should thus be associated with a shift in target choice and a new exploration phase.

FIGURE 3 | iCub robot performing Experiment 3. The robot chooses among four cubes on a table. The left screen tracks simulated activity in the neural-network model. The right screen shows the perception of the robot.

MONKEY BEHAVIORAL VALIDATION

To validate the ability of the neurocomputational model to control the robot, we compared the robot's behavioral performance with monkey data previously published as well as original monkey behavioral data. Average behavioral performances of Monkeys 1 and 2 performing Experiment 2 were taken from (Amiez et al., 2006). Trial-by-trial data of monkey M performing Experiment 1 were taken from (Quilodran et al., 2008). In addition, we analyzed unpublished data performed by three other monkeys (G, R, S) on Experiment 1 in our laboratory.

NEURAL-NETWORK MODEL DESCRIPTION

Action selection is performed with a neural-network model[1] whose architecture is inspired by anatomical connections in the prefrontal cortex and basal ganglia in monkeys (**Figure 5**). The model was programmed using the neural simulation language (NSL) software (Weitzenfeld et al., 2002). Each module in our model contains a $3 * 3$ array of leaky integrator neurons whose activity topographically encodes different locations in the visual space (i.e., nine different locations on the touch screen for Experiments 1 and 2, or on the table for Experiment 3). At each time step, a neuron's membrane potential *mp* depended on its previous history and input *s*:

$$\tau \frac{\partial \mathrm{mp}}{\partial t} = -\mathrm{mp} + s \tag{1}$$

where τ is a time constant. The average firing rate output of the neuron is then generated based on a non-linear (sigmoid) function of the membrane potential. We used $\partial t = 100$ ms, which means that we simulated 10 iterations of the model per second of real

[1]The source code of the model and a tutorial document can be downloaded at: http://chronos.isir.upmc.fr/~khamassi/projects/ACC-LPFC_2011/

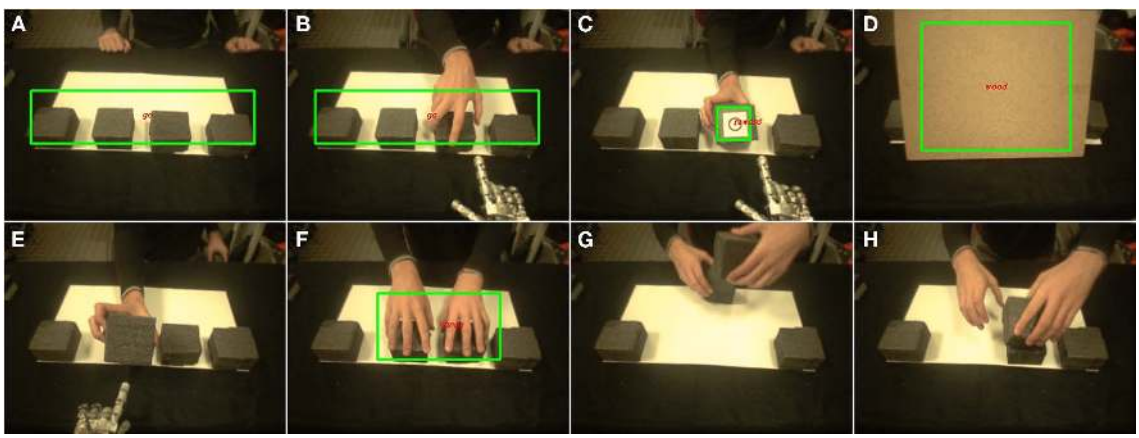


FIGURE 4 | Scenes perceived through the eyes of the iCub robot during Experiment 3. Labeled green rectangles indicate visual features recognized by the robot. The robot chose (by pointing to) one of four cubes on a table **(A,B)**. The human revealed the hidden side of the indicated cube. One of the cubes had a circle on its hidden face, indicating a reward **(C)**. At the end of a problem, the human could hide the cubes with a wooden board **(D)**, and changed the position of the rewarded cube. In early stages, this was followed by an error **(E)**. Once the robot learned the appropriate meta-value of the board, the human could cheat by unexpectedly changing the reward location **(F–H)**.

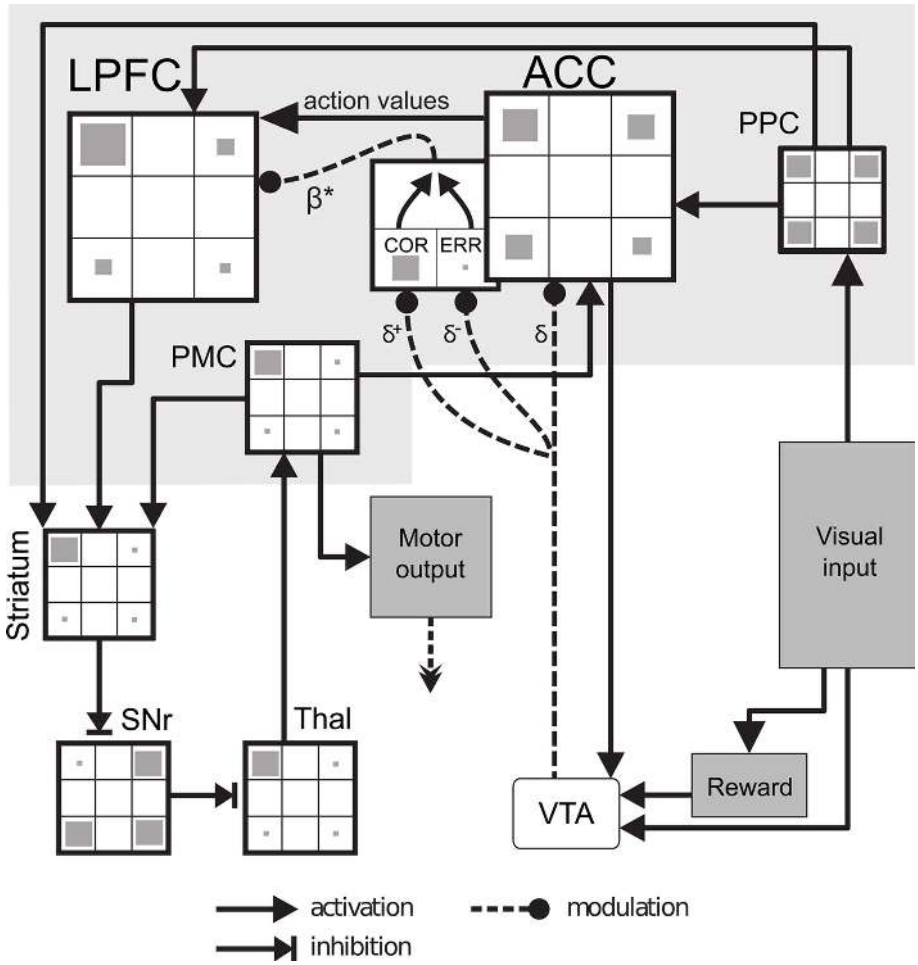


FIGURE 5 | Neural-network model. Visual input (targets seen on the screen or cubes on the table) is sent to the posterior parietal cortex (PPC). The anterior cingulate cortex (ACC) stores and updates the action value associated with choosing each possible object. When a reward is received, a reinforcement learning signal is computed in the ventral tegmental area (VTA) and is used both to update action values and to compute an outcome history (COR, correct neuron; ERR, error neuron) used to modulate the exploration level β^* in ACC. 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.

time. A parameter table is provided in the appendix, summarizing the number of neurons and parameters in each module of the model. Here we describe the role of each of these modules.

VISUAL PROCESSING

Visual information perceived by the camera is processed by a commercial object recognition software (SpikeNet; Delorme et al., 1999). Prior to each experiment, SpikeNet was trained to recognize a maximum of four different geometrical shapes (square, triangle, circle in Experiments 1 and 2; cube, wooden board, hands, circle in Experiment 3). During the task, perception of a particular shape at a particular location activates the corresponding neuron in the $4 * 3 * 3$ input matrix in the visual system of the model.

A time persistence in the visual system enables the perception of an object to progressively vanish instead of instantaneously disappear. This is necessary for robotic tests of the model during which spurious discontinuities in the perception of an object should not influence the model's behavior.

CORTICAL MODULES

In order to decide which target to touch or cube to choose, the model relies on the estimation of action values based on a Temporal-Difference learning algorithm (Sutton and Barto, 1998). In our model, this takes place in ACC, based on three principal neurophysiological findings: First – anatomical projections of the dopaminergic system that have been demonstrated to have greater strength to ACC than to LPFC (Fluxe et al., 1974). Second – the observed ACC responses to reward prediction errors (Holroyd and Coles, 2002; Amiez et al., 2005; Matsumoto et al., 2007). Third – the observed role of ACC in action value encoding (Kennerley et al., 2006; Lee et al., 2007; Rushworth et al., 2007). For Experiments 1 and 2, these action values are initialized at the

beginning of each new problem, after presentation of the PCC signal. This is based on the observation that, after extensive pre-training, monkeys show a choice shift after more than 80% of the PCC presentation (mean for Monkey G: 95%; M: 97%; R: 61%; S: 77%). In Experiment 3, the model autonomously learns to reinitialize action values (Experiment 3 Results, below).

Anterior cingulate cortex action value neurons project to LPFC, and to dopamine neurons in the ventral tegmental area (VTA) module to compute an action-dependent reward prediction error:

$$\delta = r - Q(a_i) \tag{2}$$

where a_i, $i \in \{1..4\}$ is the performed action, and r is the reward set to 1 when the corresponding cue is perceived.

In the neuroscience literature of decision-making, subjects' behavior can be well captured by RL models by computing a reward prediction error once every trial, at the feedback time, even in the case where no reward is obtained (Daw et al., 2006; Behrens et al., 2007; Seo and Lee, 2007). Here, we wanted to avoid such *ad hoc* informing of the model when the absence of reward should be considered as a feedback. Thus, dopamine neurons of the model produce a reward prediction error signal in response to any salient event (appearance or disappearance of a visual cue). In addition to being more parsimonious with respect to robotic implementation of the model, this is consistent with more general theories of dopamine neurons arguing that dopamine neurons respond to any task-relevant stimulus to prevent sensory habituation (Horvitz, 2000; Redgrave and Gurney, 2006). This reinforcement signal is sent to ACC and affects synaptic plasticity of an action value neuron only when it co-occurs with a motor efference copy sent by the premotor cortex (PMC):

The reinforcement signal δ is sent to ACC which updates synaptic weights associated to the corresponding action value neuron:

$$Q(a_i) \leftarrow Q(a_i) + \alpha \cdot \delta \cdot \mathrm{trace}(a_i) \tag{3}$$

where *trace* is the efferent copy sent by the PMC to reinforce only the performed action, and α is a learning rate.

While ACC is considered important for learning action values, decision on the action to make based on these values is known to involve the LPFC (Lee et al., 2007). Thus in the model, action values are sent to LPFC which makes a decision on the action to trigger (**Figure 5**). This decision relies on a Boltzmann softmax function, which controls the greediness versus the degree of exploration of the system:

$$P(a_i) = \frac{\exp(\beta \cdot Q(a_i))}{\sum_j \exp(\beta \cdot Q(a_j))} \tag{4}$$

where β regulates the exploration rate ($0 < \beta$). A small β leads to almost equal probabilities for each action and thus to an exploratory behavior. A high β increases the difference between the highest action probability and the others, and thus produces an exploitative behavior. As shown in **Figure 5**, such action selection results in more contrast between action neurons' activities in LPFC than in ACC during repetition phases where β is high, thus promoting exploitation.

As we wanted to adhere to the mathematical formulation employed for model-based analysis of the prefrontal cortical data recorded during decision-making (Daw et al., 2006; Behrens et al., 2007; Seo and Lee, 2007), the activity of leaky integrator neurons in our LPFC modules is algorithmically filtered at each time step by Eq. 4. We invite the reader to refer to (McClure et al., 2006; Krichmar, 2008) for a neural implementation of this precise mechanism of decision-making under exploration–exploitation trade-off.

BASAL GANGLIA LOOP

In order to prevent the robot from executing two actions at the same time when activity in LPFC related to non-selected action remains non-null, we finally implemented a winner-take-all mechanism in the basal ganglia. It has been proposed that the basal ganglia are involved in clean action selection so as to permit a winner-takes-all mechanism (Humphries et al., 2006; Girard et al., 2008). Here we simplified our previous basal ganglia loop models (Dominey et al., 1995; Khamassi et al., 2006) to a simple relay of inhibition which permits the neurophysiologically grounded disinhibition of a single selected action in the Thalamus at a given moment (**Figure 5**).

COGNITIVE CONTROL MECHANISMS

In addition to RL mechanisms, we provide the system with cognitive control mechanisms which will enable it to flexibly adjust behavioral parameters during learning. Here this is restricted to the dynamical regulation of the exploration rate β used in Eq. 4 based on the outcome history, following meta-learning principles (Schweighofer and Doya, 2003).

A substantial number of studies have shown ACC neural responses to errors (Holroyd and Coles, 2002) as well as positive feedback, a process interpreted as feedback categorization (Quilodran et al., 2008). In addition, neurons have been found in the ACC with an activity reflecting the outcome history (Seo and Lee, 2007). Thus, in our model, in addition to the projection of dopaminergic neurons to ACC action values, dopamine signals also influence a set of ACC feedback categorization neurons (**Figure 5**): 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*):

$$\begin{aligned} &\mathrm{COR}(t) = \delta(t), \text{ if } \delta(t) \geq 0 \\ &\mathrm{ERR}(t) = -\delta(t) \text{ if } \delta(t) < 0 \\ &\beta^*(t) \leftarrow \beta^*(t) + \alpha_+ \cdot \mathrm{COR}(t) + \alpha_- \cdot \mathrm{ERR}(t) \end{aligned} \tag{5}$$

where $\alpha_+ = -2.5$ and $\alpha_- = 0.25$ are updating rates with β* ($0 < \beta^* < 1$). 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 (α_+ and α_-) enables sharper changes in response to either positive or negative feedback depending on the task.

β* is then transferred to LPFC where it regulates the exploration rate β. In short, β* is algorithmically filtered by a sigmoid function

which reverses its sign, and constraints it to a range between 0 and 10:

$$\beta = \frac{\omega_1}{(1 + \exp(\omega_2 \cdot [1 - \beta^*] + \omega_3))} \quad (6)$$

where $\omega_1 = 10$, $\omega_2 = -6$ and $\omega_3 = 1$. This equation represents a sigmoid function that produces a low β when β* is high (exploration) and a high β when β* is low (exploitation).

Finally, the ACC module also learns meta-values associated with different perceived objects which represent how each of these objects is associated with variations of average reward. This will enable the robot to learn that, during Experiment 3, presentation of the wooden board is always followed by a drop in the average reward, and thus should be associated with a negative meta-value. This part of the model represents the learning process that takes place in monkeys during pre-training phases preceding Experiments 1 and 2. During such pre-training, monkeys progressively learn that different problems are separated by a PCC signal.

In the model, a reward average is computed and meta-values of objects that have been seen during the trial are updated based on variations in the reward average as computed at the end of the current trial:

$$M(o_i, t) \leftarrow M(o_i, t) + \eta \cdot \theta(t) \quad (7)$$

where η is a learning rate and $\theta(t)$ is the estimated reward average.

When the meta-value associated with any object is below a certain threshold (empirically fixed to require approximately 10 presentations before learning; see parameter table in Appendix), presentation of this object to the robot automatically triggers a reset of action values and β* variable – action values are reset to random values while β* is increased so that it produces a low β (corresponding to exploration). As a consequence, the robot will display exploratory behavior after such reset.

MOTOR COMMANDS

Motor output from the model's PMC module is sent to the robotic devices via port communication with YARP (Metta et al., 2006).

RESULTS

EXPERIMENT 1

We first performed a first series of 11 sessions with the Lynxmotion SES 5DOF robotic arm (http://www.lynxmotion.com) on the problem-solving task described above. This corresponded to a total of 112 problems and 717 trials. **Figure 6** shows a sample performance of the model on two consecutive problems – corresponding to 14 trials. Each trial lasted a few seconds and resulted in the selection of one of the four targets – corresponding to different colors on the third chart of **Figure 6**. At the beginning of a trial, the perception of the onset of the four targets on the screen produced an increase of activity of ACC and LPFC neurons (first two charts on **Figure 6**). The neuron with the highest activity activated a selection of the corresponding target by the robot. At the end of the trial, the offset of the targets with or without reward (depending on the correctness of the robot's choice) resulted in a drop of ACC and LPFC activity and return of the robot's arm to its initial position (end of target choice on the third chart of **Figure 6**). During the first problem, the robot selected three successive targets (indicated by the green, blue and brown blocks in **Figure 6**) corresponding to error trials until the correct target was chosen (the target illustrated as orange in **Figure 6**) and a reward was obtained (ACC COR neuron **Figure 6**). The errors lead to a progressive increase of activity of β* along the search phase – producing more exploratory behavior – and a drop of β* after the first reward – promoting exploitation during repetition (Fifth chart of **Figure 6**). Such activity may explain our finding that many ACC neurons respond more during the search phase than during the repetition phase (Procyk et al., 2000; Quilodran et al., 2008).

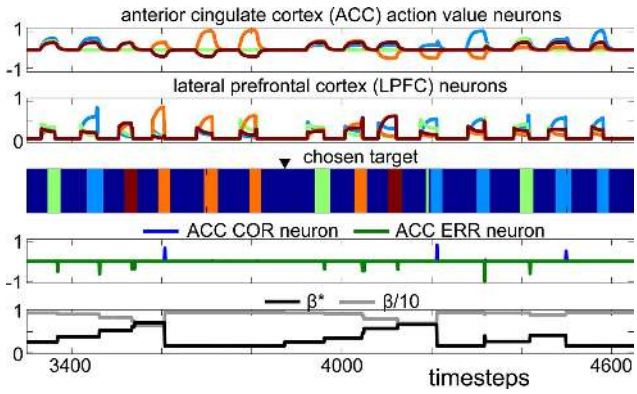


FIGURE 6 | Simulation of the model on two consecutive problems during Experiment 1. Each color represents a different target chosen by the robot. The black triangle above the "chosen target" chart indicates the presentation of the Problem-Changing Cue before the start of a new problem. The *x*-axis represents time. The first chart shows the activity of ACC action value neurons. The second chart shows LPFC action neurons. The fourth chart shows ACC feedback categorization neurons, indicating errors (ERR) and correct (COR) trials, and induced by dopaminergic reward prediction error signals. The last chart shows the evolution of the exploration rate β in the model. This simulation illustrates the correct execution of the task by the robot and shows the incremental variation of the exploration rate in response to positive and negative feedback.

In the model, we made the hypothesis that feedback categorization responses in the ACC would emerge from reward prediction error signals (Eq. 5; Holroyd and Coles, 2002). Interestingly, the high learning rate α suitable for the task produced a positive reward prediction error (and thus a COR response of ACC feedback categorization neurons) only at the first correct trial, and not at subsequent correct trials during repetition where the reward prediction error in the model was null (**Figure 6**). This may explain why, in monkeys, ACC neurons responding to positive feedback in the same task mainly responded during the first correct trial and less to subsequent correct trials (Quilodran et al., 2008). Indeed, these neurons have been interpreted as responding to dopamine reward prediction error signals. Validating this interpretation, the explanation emerging from the model for the precise pattern of response of these neurons is that subsequent correct trials during repetition were correctly expected and thus did not produce a reward prediction error.

In terms of behavior, the robot quickly adapted to feedback obtained at each trial and rarely repeated choice errors. The second

half of the session shown on **Figure 6** illustrates a case where the robot adapted to uncertainty emerging from perceptual ambiguities. Around time step 3900, a new problem started, cued by the PCC, and the model thus resets its exploration rate and action values. The robot searched for the new correct target (the target illustrated as blue in **Figure 6**), and once found, repeated the correct choice. However, due to visual ambiguity that could occasionally take place during such physical interaction with the environment the robot interpreted the trial as incorrect. Specifically, in this case, while touching the correct target the robot's arm hid the targets on the screen and the system thus perceived targets as vanishing long before reward occurrence. As a consequence, the model generated a negative reinforcement signal which reduced the action value associated with the correct target (time step 4300 on **Figure 6**). This lead to the choice of a different target on the next trial, and finally a return to the correct choice, to properly finish the repetition phase. This demonstrates that perceptual noise inherent in robotic systems can be accommodated by such type of neurophysiologically inspired model.

We next compared the robotic results with real monkey data collected in the same task and tests of the same model in simulation, to assess robustness in real-world conditions and variations in performance due to embodiment. Monkey behavioral data were collected in four monkeys for a total of 7397 problems and 46188 trials. **Figure 7A** shows the average errors during search versus repetition phases. Similar to monkeys, the robot produced approximately 60% errors during the search phase, which is close to optimality (considering that in 90% of new problems, the correct target was different from the previous problem, there were $2/3 = 66.67\%$ chances of choosing a wrong target). During the repetition phase, the robot made approximately 85% correct responses, which was similar to monkeys. In contrast, simulation of the same model made no error during repetition, as task-related perception in the simulation was always perfect.

Performance of the robot was also similar to monkeys when considering the average duration of search and repetition trials (**Figure 7B**). The search phase for the robot lasted 2.5 trials on average which was not different from that of monkeys (Kruskal–Wallis test, $p > 0.31$). The repetition phase lasted less than four trials, again not different from monkeys (Kruskal–Wallis test, $p > 0.78$). The robot's behavior thus did not differ from that of the monkeys. In contrast, the simulation always took exactly three trials during repetition, which was the smallest possible duration and was statistically different from monkey performance during repetition (Kruskal–Wallis test, $p = 1.6\text{e-}12$).

Thus, in addition to respecting known anatomy and reproducing neurophysiological properties observed in the monkey prefrontal cortex during the same task, the model could reproduce global behavioral properties of monkeys when driving a robot[2].

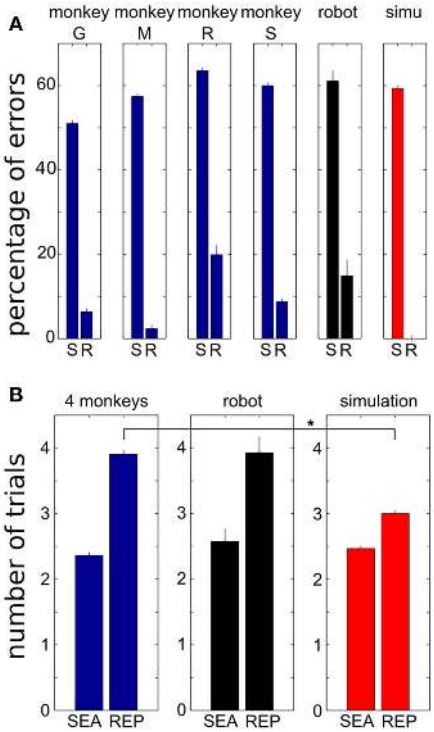


FIGURE 7 | Comparison of simulation results, robotics results and behavioral data obtained in monkeys performing Experiment 1. The percentage of errors **(A)** and the duration of problems **(B)** performed by the robot were not different from that of monkeys. In contrast, simulation of the model provided perfect performance. S/SEA, search; R/REP, repetition. Errorbars: SEM. "*" Indicates significant difference via Kruskal–Wallis test, $p < 0.05$.

EXPERIMENT 2

In order to test the ability of our neuro-inspired model to generalize over variations in task conditions, we next tested it in simulation on a stochastic version of the problem-solving task used in monkeys (Amiez et al., 2006). The reward distribution was stochastically distributed over two possible targets, and so obtaining the largest reward value was possible even when choosing the wrong target (see **Table 1**). Thus a single correct trial was not sufficient to know which target had the highest value. As a consequence, we predicted that the same model with a smaller learning rate α (used in Eq. 3) would better explain monkeys' behavior, as a reduced learning rate would require several successful trials before convergence.

Consistent with our prediction, a naive test on the stochastic task with the parameters used with Experiment 1 and a fixed exploration rate β – that is, without the β^*-mechanism for exploration regulation ($\alpha = 0.9$, $\beta = 5.2$) – elicited a mean number of search trials of 13.3 ± 12.3 with only 87% successful problems – problems during which the most rewarded target was found and correctly repeated ("Model no-β^*" on **Figures 8A,B**). This represented poor performance compared to monkeys. In the original experiment, the two monkeys found the best target in 98% and 94.5% of the problems. The search phase lasted on average 6.4 ± 5.6 and 5.6 ± 6.9 trials respectively (Amiez et al., 2006).

We then explored different values of the learning rate combined with a flexible adaptation of the exploration rate β regulated by the modulatory variable β^*. This provided results closer to

[2]A video of the SES robotic arm performing the PS task can be downloaded at: http://chronos.isir.upmc.fr/~khamassi/projects/ACC-LPFC_2011/

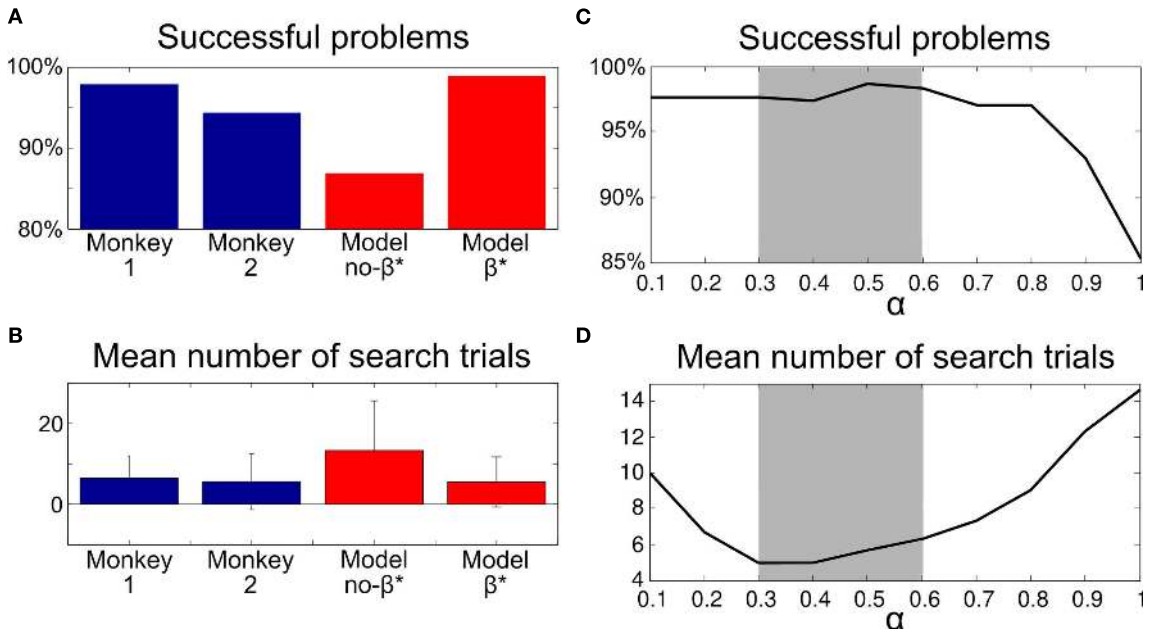


FIGURE 8 | Comparison of simulated results and monkey performance in Experiment 2. (A) Percentage of problems where the agent did eventually find the correct target and passed the criterion for a correct repetition phase. **(B)** Performance during search trials. **(C)** Percentage of Successful problems for different values of α tested with the model with β*. **(D)** Duration of the search phase for the same tests. Without the dynamic regulation of the exploration level computed with β*, the model produced worst performance than monkeys. Errorbars: SD.

monkey performance. Roughly, monkeys' performances could be best approximated with α between 0.3 and 0.6 (**Figures 8C,D**). This produced a mean number of search trials of 5.5 and 99% successful problems ("Model β*" on **Figures 8A,B**).

Interestingly, monkey performance could be best approximated with a mean α around 0.5 during Experiment 2, while a higher mean α (0.9 on average) better explained monkey behavior during Experiment 1. This is consistent with theoretical propositions for efficiently regulating the learning rate α based on the volatility of the task (Rushworth and Behrens, 2008). Indeed, in Experiment 1 the correct target changed every seven trials on average (as illustrated in **Figure 7**) which was more volatile than Experiment 2 where changes of reward distribution occurred less frequently: every 16 trials (∼six search trials as illustrated in **Figure 8**, and 10 repetition trials imposed by the task structure).

Concerning the optimization of β, it is remarkable that the more exploitative the better the performances (low β induced a too lengthy search phase because the model was too exploratory). Unlike our initial hypothesis, this was in part due to the nature of Experiment 2 in which only two targets were available, decreasing the search space, so the best strategy was clearly exploitative. In accordance with this finding, β was systematically adjusted with β* to its highest possible value allowed here (around 10). The optimized model with a fixed exploration rate β reached a nearly optimal behavior – in the sense of reward maximization. In contrast, the model with a dynamic exploration rate achieved good performance (although not as good) but nevertheless closer to monkeys' performance in this task. This suggests that such brain inspired adaptive mechanisms are not optimal but might have been selected through evolution because they can produce satisfactory performance in a variety of different conditions.

EXPERIMENT 3

The last experiment was implemented for two purposes:

- In the previous experiments the model knew *a priori* that a particular signal called PCC 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.
- We also wanted to test our neuro-inspired model on a humanoid robot performing a simple human–robot interaction scenario where the human can introduce unexpected uncertainty or cheat, showing the potential applications of the model to more complex situations.

During the course of eight experiment sessions, the robot performed a total of 151 problems and 901 trials. **Figure 9** shows a sequence of 14 problems performed by the model on the iCub robot during Experiment 3. Similar to Experiment 1, the robot searched for the correct cube and repeated its choice once that cube had been determined.

Also similarly to Experiment 1, we used a "PCC" which was here a wooden board used to hide the cubes while the human changed the position of the rewarded one (**Figure 4D**). An important difference with Experiment 1 was that the model did not *a priori* know what this signal meant and made errors following its presentation during the first part of a session. Since the wooden board was always associated to an error, the robot learned by itself to shift its behavior and restarted to explore when it was later presented. This was achieved by learning meta-values associated to different perceived objects: each time the perception of a given object was

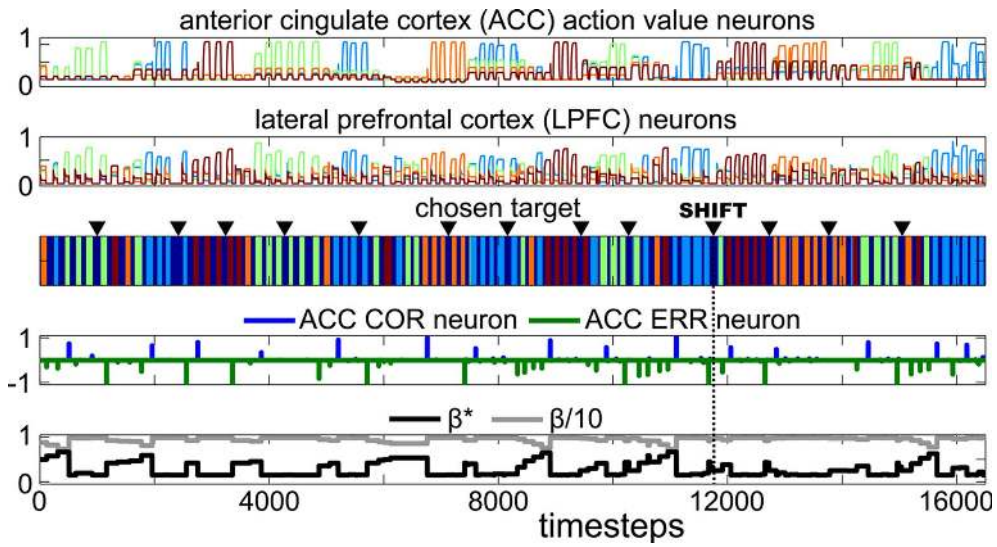


FIGURE 9 | Example session comprising 14 consecutive problems performed by the iCub robot during Experiment 3. The different parts of the figure follow the same legend as **Figure 6**. This session illlustrates that after many presentations of the wooden board (triangles) followed by errors, the robot learned to associate it with a condition change. This happened around time step 12000 (labeled as "SHIFT") where the apparition of the wooden board successfully triggered a reset of action values and of the exploration rate.

followed by a variation (positive or negative) of the average reward obtained by the robot, the meta-value of this object was slightly modified (Eq. 7). 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. When the meta-value of a given object became significantly low, the robot systematically shifted its behavior and restarted to explore each time the object appeared again.

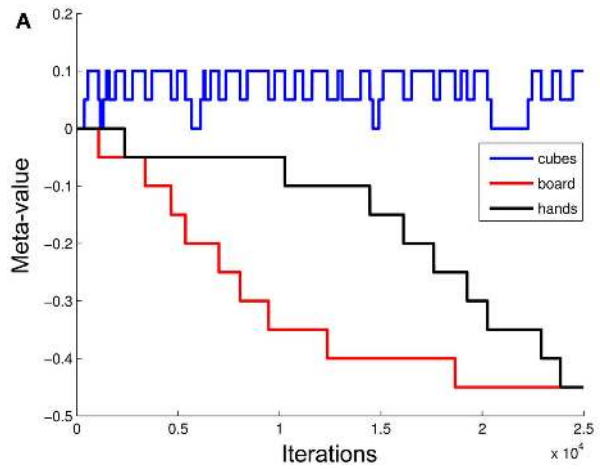
Figure 10A shows the evolution of the meta-values associated with the board, the cubes and perception of the experimenter's hands grasping the cubes. We can see that the board's meta-value incrementally decreased – each time it was presented and followed by an error. In the example session shown on **Figure 9**, the meta-value of the board became sufficiently low to enable a behavioral shift at the beginning of the 11th problem after about 12000 time steps. At that moment, the human hid the cubes with the board, changed the position of the rewarding cube, and the robot directly chose a new cube (exploration).

When looking at all eight experiments performed by the robot, among 55 presentations of the board that occurred in the first 10000 iterations of a session, the robot shifted only five times (9.1% of the time). Among 37 presentations of the board that occurred after the 10000 first iterations, the robot shifted 29 times (78.4%). Thus the iCub robot learned to shift in response to the board.

Such a learned behavioral shift produced an improvement in the robot's performance on the task. During the second part of each session, the robot made fewer errors on average during search phases, and required fewer trials to find the correct cube. Before this shifting was learned, in 65 problems initiated by a board presentation, the robot took on average 3.5 trials to find the correct cube. After shifting learned, 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, $p < 0.001$).

Figure 10 also shows that the meta-value associated with the cubes themselves fluctuated – because perception of the cubes was sometimes followed by correct choices, sometimes by errors – but remained within a certain boundary. As a consequence, the robot did not unlearn the task. If the cubes' meta-value had also significantly declined, the robot would have reset action values at each presentation of the cubes (i.e., at each trial), and would not have been able to find the correct target. Thus, such meta-learning mechanism may be a good model of how animals learn the structure of the task during the pre-training phase of Experiments 1 and 2: (A) Learning that some cues are sometimes followed by rewards, sometimes by errors, and are thus subject to RL; (B) Learning that some other cues such as the PCC are always followed by errors and shall be associated with a task change which requires a reset of action values and exploration each time they are presented.

We finally addressed an additional degree of complexity. During the second half of each experiment, once the robot had learned to shift its choice in response to the wooden board, the human introduced new unexpected uncertainty by occasionally "cheating" in the middle of a problem. The human put his hands on the cubes, grasped them and changed their position without hiding the cubes with the board (as illustrated on **Figures 4F–H**). The robot saw such an event by recognizing the hands on the cubes. This was *a priori* provided to the robot as a possible visual feature, but was not *a priori* associated with any meaning. In a first stage, this event was systematically followed by an error from the robot which selected the cube location associated to the highest value (exploitation), though the human had "cheated" by moving the rewarded cube to a different location. A first degree of flexibility was enabled by the model's RL mechanisms. This permitted the robot to decrease the value of the cube location following this



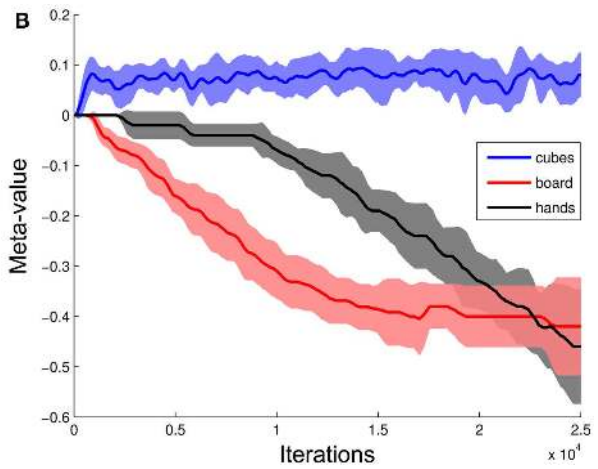


FIGURE 10 | Evolution of objects' meta-values *M*(*o*, *t*) associated to the different perceived objects (cubes, board, and hands) and computed with Eq. 7 (A) during a sample session; (B) averaged over the eight experiments performed by the iCub robot (*N* = 8). After about 10000 iterations of the model, the low meta-value of the board allowed a high probability of triggering a new exploration phase at subsequent presentation of that object to the robot (i.e., SHIFT event on **Figure 9**). Such probability was reached after about 20000 iterations for the hands perceived when the human cheated. Errorbars: SEM.

error, and thus to avoid persistence in failure: among 37 times where the human cheated followed by an error from the robot, in 34 cases (91.9%) the robot shifted at the next trial. In addition and similar to the board, the meta-value of the perceived hands incrementally decreased, finally producing a high probability of triggering a new exploration phase each time it occurred (**Figure 10B**). Thus the robot progressively learned to shift its behavior in response to the human's hands configuration during cheating: Among 16 such events occurring after the 20000 first iterations of a session, the robot shifted 10 times (62.5%) while it shifted in only 3.0% (1/33) of the cases during the first 20000 iterations of each session[3].

DISCUSSION

This work showed the application of a neuro-inspired computational model on a series of robotic experiments inspired by monkey neurophysiological tasks. The last experiment extended such tasks to a simple human–robot interaction scenario.

This demonstrates that a neuro-inspired model could adapt to diverse conditions in a real-world environment by virtue of:

- Reinforcement learning (RL) principles, enabling the capability to learn by trial-and-error, and to dynamically adjust values associated to behavioral options;
- Meta-learning mechanisms, here enabling the dynamic and autonomous regulation of one of the RL meta-parameters called the exploration rate β.

The model synthesizes a wide range of anatomical and physiological data concerning the Anterior Cingulate-Prefrontal Cortical system. In addition, certain aspects of the neural activity produced by the model during performance of the tasks resembles previously reported ACC neural patterns that where not *a priori* built into the model (Procyk et al., 2000; Quilodran et al., 2008). Specifically, like neurons in the ACC, in the model ACC 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 task. This provides a functional explanation for these observations. Detailed analysis of the model's activity properties during simulations without robotic implementation provided testable predictions on the proportion of neurons in ACC 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 (Khamassi et al., 2010). In the future we will test hypotheses emerging from this model on simultaneously recorded ACC and LPFC activities during PS 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 in Experiment 1 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.

By incorporating a model based on neuroscience hypotheses in a robot, we had to make concrete hypotheses on the interaction between brain structures dedicated to different cognitive processes. Robotic constraints prevented us from providing *ad hoc* information often used during perfectly controlled simulations, such as the information that the absence of reward at the end

[3]A video of the iCub robot performing the cube game can be downloaded at: http://chronos.isir.upmc.fr/~khamassi/projects/ACC-LPFC_2011/

of a trial should be considered as a feedback signal for the RL model (Daw et al., 2006; Behrens et al., 2007; Seo and Lee, 2007). Instead, dopamine neurons of our model produced a reward prediction error signal in response to any salient event (appearance or disappearance of a visual cue) and could affect synaptic plasticity of an action value neuron within ACC only when it co-occurred with an efferent copy sent by the PMC. Interestingly, dopamine neurons were previously reported to respond also to salient neutral stimuli (Horvitz, 2000), which was interpreted as a role of dopamine neurons in blocking sensory habituation and sustaining appetitive behavior to learn task-relevant action-outcome contingencies (Redgrave et al., 2008). Moreover, in the case of dopaminergic signaling to the striatum, it has been reported that a motor efference copy is sent to the striatum in conjunction with the phasic response of dopaminergic neurons, which was interpreted as enabling a specific reinforcement of relevant action-outcome contingencies (reviewed in Redgrave et al., 2008). Thus, an interesting neurophysiological experiment that could permit to validate or refute choices implemented in our model would consist in recording dopaminergic neurons during our PS task and see whether: (1) they respond to neutral salient events; (2) their response to trial outcomes is contingent with traced inputs from PMC to ACC.

Importantly, our work demonstrated that the model could also be applied to human–robot interaction. The model enabled the robot to solve the task imposed by the human 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 PCC during the pre-training phase of Experiments 1 and 2. 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.

Future work can also include a refinement of the β^*-based regulation of exploration within the LPFC so as to take into account noradrenergic neuromodulation within a network of interconnected cortical neurons. Indeed, here we wanted to evaluate mathematical principles of meta-learning for the regulation of exploratory decisions. As a consequence, we simply algorithmically transferred the outcome history computed in ACC into the β variable used in the softmax equation for action selection in LPFC (Eq. 4). This does not preclude a neural implementation of such an interaction. It has previously been shown that noradrenergic neurons in the locus coeruleus (LC) shift between two modes of response between exploration and exploitation phases, and that noradrenaline changes the signal-to-noise ratio within the prefrontal cortex (Aston-Jones and Cohen, 2005). Given that ACC projects to LC and drives phasic responses of LC noradrenergic neurons (Berridge and Waterhouse, 2003; Aston-Jones and Cohen, 2005), our model is consistent with such a configuration. A possible improvement of our model would be to replace the algorithmic implementation of the softmax function in our LPFC module by a modulation of extrinsic and inhibitory synaptic weights between competing neurons based on the level of noradrenergic innervation, as proposed by (Krichmar, 2008).

On the robotic side, future work could involve autonomous learning of the relevant objects of each experiment (i.e., those that are regularly presented) and adaptive regulation of the learning rate α when shifting between deterministic and stochastic reward conditions (Experiment 1 and 2 respectively). The latter could be achieved by extracting measures of the dynamics of the different task conditions, such as the reward volatility which is expected to vary between deterministic and stochastic conditions (Rushworth and Behrens, 2008; see (Khamassi et al., in press) for a review of this issue on PS tasks). We also plan to extend the model to social rewards provided by the human to the robots by means of language (Dominey et al., 2009; Lallée et al., 2010).

Such pluridisciplinary approaches provide tools both for a better understanding of neural mechanisms of decision-making and for the design of artificial systems that can autonomously extract regularities from the environment and interpret various types of feedback (rewards, feedbacks from humans, etc...) based on these regularities to appropriately adapt their own behaviors.

ACKNOWLEDGMENTS

This research is supported by the European Commission FP7 ICT Projects CHRIS, Organic, and EFAA, and French ANR Projects Amorces and Comprendre.

REFERENCES

Amiez, C., Joseph, J.-P., and Procyk, E. (2005). Anterior cingulate error-related activity is modulated by predicted reward. *Eur. J. Neurosci.* 21, 3447–3452.

Amiez, C., Joseph, J.-P., and Procyk, E. (2006). Reward encoding in the monkey anterior cingulate cortex. *Cereb. Cortex* 16, 1040–1055.

Arbib, M., Metta, G., and van der Smagt, P. (2008). "Neurorobotics: from vision to action," Chapter 62, in *Handbook of Robotics* eds B. Siciliano and O. Khatib (Berlin: Springer-Verlag), 1453–1480.

Aston-Jones, G., and Cohen, J. D. (2005). Adaptive gain and the role of the locus coeruleus-norepinephrine system in optimal performance. *J. Comp. Neurosci.* 493, 99–110.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit. *Mach. Learn.* 47, 235–256.

Behrens, T. E., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nat. Neurosci.* 10, 1214–1221.

Berridge, C. W., and Waterhouse, B. D. (2003). The locus coeruleus-noradrenergic system: modulation of behavioral state and state-dependent cognitive processes. *Brain Res. Rev.* 42, 33–84.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., and Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychol. Rev.* 108, 624–652.

Brown, J. W., and Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulated cortex. *Science* 307, 1118–1121.

Cohen, J. D., Aston-Jones, G., and Gilzenut, S. (2004). "A systems-level perspective on attention and cognitive control," in *Cognitive Neuroscience of Attention* ed. M. Posner (New York: Guilford Publications), 71–90.

Cohen, J. D., McClure, S. M., and Yu, A. J. (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, 933–942.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., and Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature* 441, 876–879.

Dehaene, S., and Changeux, J. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl. Acad. Sci. U.S.A.* 95, 14529–14534.

Delorme, A., Gautrais, J., Van Rullen, R., and Thorpe, S. (1999). SpikeNET: a simulator for modeling large networks of integrate and fire neurons. *Neurocomputing* 26–27, 989–996.

Dominey, P. F., Arbib, M., and Joseph, J.-P. (1995). A model of corticostriatal plasticity for learning oculomotor associations and sequences. *J. Cogn. Neurosci.* 7, 311–336.

Dominey, P. F., Mallet, A., and Yoshida, E. (2009). Real-time spoken-language programming for cooperative interaction with a humanoid apprentice. *Int. J. HR* 6, 147–171.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Netw.* 15, 495–506.

Fluxe, K., Hokfelt, T., Johansson, O., Jonsson, G., Lidbrink, P., and Ljungdahl, A. (1974). The origin of the dopamine nerve terminals in limbic and frontal cortex. Evidence for mesocortico dopamine neurons. *Brain Res.* 82, 349–355.

Frank, M. J., Doll, B. B., Oas-Terpstra, J., and Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nat. Neurosci.* 12, 1062–1068.

Girard, B., Tabareau, N., Pham, Q. C., Berthoz, A., and Slotine, J. J. (2008). Where neuroscience and dynamic system theory meet autonomous robotics: a contracting basal ganglia model for action selection. *Neural Netw.* 21, 628–641.

Holroyd, C. B., and Coles, M. G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychol. Rev.* 109, 679–709.

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience* 96, 651–656.

Humphries, M. D., Stewart, R. D., and Gurney, K. N. (2006). A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. *J. Neurosci.* 26, 12921–12942.

Kennerley, S. W., Walton, M. E., Behrens, T. E. J., Buckley, M. J., and Rushworth, M. F. (2006). Optimal decision making and the anterior cingulate cortex. *Nat. Neurosci.* 9, 940–947.

Khamassi, M., Martinet, L.-E., and Guillot, A. (2006). "Combining self-organizing maps with mixtures of experts: application to an actor-critic model of reinforcement learning in the basal ganglia," in *From Animals to Animats 9: Proceedings of the Ninth International Conference on Simulation of Adaptive Behavior (SAB)* (Berlin: Springer-Verlag), 394–405. [LNAI 4095].

Khamassi, M., Quilodran, R., Enel, P., Procyk, E., and Dominey, P. F. (2010). "A computational model of integration between reinforcement learning and task monitoring in the prefrontal cortex," in *From Animals to Animats 11: Proceedings of the Eleventh International Conference on Simulation of Adaptive Behavior (SAB)* (Berlin: Springer-Verlag), 424–434. [LNAI 6226].

Khamassi, M., Wilson, C., Rothé, R., Quilodran, R., Dominey, P. F., and Procyk, E. (in press). "Meta-learning, cognitive control, and physiological interactions between medial and lateral prefrontal cortex," in *Neural Basis of Motivational and Cognitive Control*, eds R. Mars, J. Sallet, M. Rushworth, and N. Yeung (MIT Press).

Koechlin, E., and Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends Cogn. Sci.* 11, 229–235.

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

Krichmar, J. L. (2008). The neuromodulatory system – a framework for survival and adaptive behavior in a challenging world. *Adapt. Behav.* 16, 385–399.

Lallée, S., Madden, C., Hoen, M., and Dominey, P. F. (2010). Linking Language with embodied and teleological representations of action for humanoid cognition. *Front. Neurorobot.* 4:8. doi: 10.3389/fnbot.2010.00008

Lee, D., Rushworth, M. F., Walton, M. E., Watanabe, M., and Sakagami, M. (2007). Functional specialization of the primate frontal cortex during decision making. *J. Neurosci.* 27, 8170–8173.

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

McClure, S. M., Gilzenrat, M. S., and Cohen, J. D. (2006). "An exploration–exploitation model based on norepinephrine and dopamine activity," in *Advances in neural information processing systems (NIPS)*, eds Y. Weiss, B. Sholkopf, and J. Platt (Cambridge, MA: MIT Press), 867–874.

Metta, G., Fitzpatrick, P., and Natale, L. (2006). YARP: Yet Another Robot Platform. *Int. J. Adv. Robot. Syst.* 3, 43–48.

Meyer, J.-A., and Guillot, A. (2008). "Biologically-inspired robots," chapter 60, in *Handbook of Robotics*, eds B. Siciliano and O. Khatib (Berlin: Springer-Verlag), 1395–1422.

Miller, E. K., and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202.

Pfeifer, R., Lungarella, M., and Iida, F. (2007). Self-Organization, embodiment, and biologically inspired robotics. *Science* 318, 1088–1093.

Procyk, E., and Goldman-Rakic, P. S. (2006). Modulation of dorsolateral prefrontal delay activity during self-organized behavior. *J. Neurosci.* 26, 11313–11323.

Procyk, E., Tanaka, Y. L., and Joseph, J. P. (2000). Anterior cingulate activity during routine and non-routine sequential behaviors in macaques. *Nat. Neurosci.* 3, 502–508.

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

Redgrave, P., and Gurney, K. N. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.* 7, 967–975.

Redgrave, P., Gurney, K. N., and Reynolds, J. (2008). What is reinforced by dopamine signals? *Brain Res. Rev.* 58, 322–339.

Rushworth, M. F., and Behrens, T. E. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat. Neurosci.* 11, 389–397.

Rushworth, M. F., Behrens, T. E., Rudebeck, P. H., and Walton, M. E. (2007). Contrasting roles for cingulate and orbitofrontal cortex in decision and social behavior. *Trends Cogn. Sci.* 11, 168–176.

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

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599.

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

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

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

Tsagarakis, N. G., Metta, G., Sandini, G., Vernon, D., Beira, R., Becchi, F., Righetti, L., Victor, J. S., Ijspeert, A. J., Carrozza, M. C., and Caldwell, D. G. (2007). iCub – the design and realization of an open humanoid platform for cognitive and neuroscience research. *Adv. Robot.* 21, 1151–1175.

Weitzenfeld, A., Arbib, M. A., and Alexander, A. (2002). *The Neural Simulation Language: A System for Brain Modeling.* Cambridge, MA: MIT Press.

Yu, A. J., and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron* 46, 681–692.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 February 2011; accepted: 11 June 2011; published online: 12 July 2011. Citation: Khamassi M, Lallée S, Enel P, Procyk E and Dominey PF (2011) Robot cognitive control with a neurophysiologically inspired reinforcement learning model. Front. Neurorobot. ***5****:1. doi: 10.3389/fnbot.2011.00001*

APPENDIX

	ACC	ACC	LPFC	PMC	PPC	Visual	VTA	Striatum	SNr	Thal
Number of neurons	9 Qval	1COR, 1ERR, 1 β*	9 Input, 9 out.	9	4 ∗ 9 (4 geometrical shapes)	4 ∗ 9 (4 geometrical shapes)	1 δ	9	9	9
Δt	100 ms	100 ms	100 ms	100 ms	100 ms	100 ms	100 ms	100 ms	100 ms	100 ms
τ (Eq. 1)	0.6	0.1	0.5	0.25	0.5	0.4	N.A.	0.5	0.5	0.5
α	0.9 for Experiment 1 and Experiment 3, 0.5 for Experiment 2									
α_+/α_-		−2.5/0.5								
βinit/η		0.25/0.1								
$\omega_1/\omega_2/\omega_3$		10/−6/1								
Threshold for reset of exploration		−0.25								
Threshold salient event							0.6			
Input threshold	0.75		0.75	0.75						

Parameter table showing the number of neurons and the parameter values of each module. Most modules contain nine neurons (i.e., a $3 * 3$ array topographically encoding different locations in the visual space). The ventral tegmental area (VTA) module contains a single simulated dopaminergic neuron dedicated to reward prediction errors computation (Eq. 2). Δt is the time separating each iteration of the model. t is the time constant of leaky integrator neurons used in Eq. 1. The anterior cingulate cortex (ACC) also contains neurons categorizing feedback (COR: correct; ERR: error) used to estimate the current performance of the agent by means of β* and to regulate the exploration rate β through Meta-Learning. Parameters are separately shown for the part of ACC responsible for Reinforcement learning (RL) and the part of ACC responsible for Meta-learning (ML). Since there is no learning in other parts of the model, RL and ML parameters concern only the ACC. α is the learning rate used for RL processes in Eq. 3. α_+ and α_- are the specific learning rates used for the update of β* (equation 5). β*init* is the value to which the exploration rate is reset in two cases: (1) in Experiments 1 and 2, reset is systematically performed at the beginning of each new problem; (2) in Experiment 3, reset is performed when the robot perceives an object to which it has learned to associate a meta-value below a certain negative threshold ("Threshold for reset of exploration" in the parameter table). η is the update rate of meta-values (Eq. 7). ω_1, ω_2, and ω_3 are the parameters of the sigmoid used to translate β* ($0 < \beta^* < 1$) into β(here, $0 < \beta < 10$). In VTA, a reward prediction error is computed only when a salient event is detected (i.e., when a change concerning perceived objects in the visual space is above a certain threshold, here written as "threshold salient event"). Finally, reinforcement learning and action selection within ACC and LPFC are permitted only when the robot perceives some objects, that is when information about perceived objects in the visual space is above a certain threshold, here written as "input threshold."