Neuronal Mechanisms Underlying Exploration-Exploitation Strategies in Operant Learning

Sergio Lew, Hernan G. Rey, B. Silvano Zanutto, Member, IEEE

Abstract — One of the most valuable mechanisms in animal self-adaptation is the ability to switch between exploration and exploitation strategies. In this work, we present a computational model that learns visual discrimination paradigms and adapts its behavior according to rules change. In the model, dopamine and norepinephrine neurons are proposed as detectors of changes in the environment. Dopamine modulates the excitability and plasticity of artificial neurons in the prefrontal cortex and motor-related structures. These neurons change their synaptic weights following a Hebbian or anti-Hebbian rule depending on the amount of released dopamine and, as the reward rate increases, they induce exploitative behaviors. On the other hand, tonic levels of norepinephrine modulate both, information flows towards motor structures and the excitability of dopaminergic neurons, facilitating the switch from exploitation to exploration strategies. The computational model predicts behavioral and physiological results and provides a computational framework to the exploration-exploitation dilemma in self-adaptive agents.

I. INTRODUCTION

BEHAVIORAL adaptation is a key ability that allows to maximize reward and to minimize punishment in changing environments. In reinforcement learning, animals select behavioral responses by means of two well-differentiated strategies: exploration and exploitation. In the artificial intelligence (AI) literature, this issue has been solved using static or dynamic linear combinations of these strategies or with “selective attention” mechanisms for switching between them [1]. However, none of these mechanisms have biological plausibility. On the other hand, understanding how animals control their behavior might provide new ideas about how to build intelligent robots [2][3][4][5][6].

There is experimental evidence obtained from behaving monkeys that shows that the prefrontal cortex (PFC) is a key component in learning [7] and that dopamine (DA) and norepinephrine (NE) modulate cortical and subcortical neuronal activity. In the PFC, dopamine acts as a neuromodulator of the excitability of neuronal clusters.

There are two main effects concerning the action of dopamine over pyramidal neurons in the PFC: firing inhibition via GABAergic interneurons and synergism between D1 dopaminergic receptors and NMDA receptors [8][9].

In behaving monkeys, dopaminergic neurons of the ventro tegmental area (VTA) and of the Substantia Nigra pars Compacta (SNc) fire when appetitive unconditioned stimuli (US), such as food or sweet liquids, are delivered. In classical and operant conditioning, after a few trials of stimuli pairing, those neurons start to fire when conditioned stimuli (CS) are presented. Also an inhibition can be observed if predicted rewards are omitted [10].

Visual and somatosensory cortical neurons are modulated by noradrenergic neurons of the Locus Coeruleus (LC) [11]. The LC function has at least two distinguishable modes [12]. In the phasic mode, neuron bursts are closely coupled with behavioral responses that are generally highly accurate. In a tonic mode, LC activity is increased when performance is more erratic, with higher response times and error rate. The increased tonic firing reflects increased environmental uncertainty. When reward persistently becomes weaker, an increase in the LC-NE tonic activity facilitates alternative behaviors in order to obtain reward. In monkeys, the strong projection of the medial prefrontal cortex (mPFC) onto the LC [12] suggests its importance in generating reward-related patterns of LC activity. Changes in LC firing rate serves to optimize the balance between exploitative and explorative behaviors, increasing the likelihood of obtaining reward and thereby maximizing utilities. In addition to the main effects of NE at cortical structures, in midbrain areas, noradrenergic innervation of DA neurons has been shown to be a mechanism of excitability modulation [13][14]

In previous works, we developed a neural network model that learns many different tasks: the matching law, partial reinforcement extinction, blocking, learned helplessness, response selection, successive negative contrast effect, modulation of the avoidance response, transfer of control between conditioned stimuli and spontaneous recovery [15][16][17][18], as well as all the experiments explained by other the computational models [19]. Our modeling work has been successfully applied in the robotics area. Gutnisky and Zanutto [4] implemented the control of a vehicle in an avoidance task by a previously developed operant learning model that has better performance when compared with the Q-Learning algorithm. In Gutnisky and Zanutto [5] the phenomenon of cooperation was studied. A modified operant learning model was compared against other strategies in the Iterated and Evolutionary Prisoner’s Dilemma task. The proposed model outperforms the other
strategies, especially when (even slightly) noisy responses were allowed.

Moreover, using our model, we developed an animat approach to dynamic team formation in a group of distributed robots [6]. Here we will show how during operant learning, exploration and exploitation strategies emerge as a consequence of concurrent dopaminergic and noradrenergic modulation in the PFC and motor related structure.

II. THE MODEL

The activity of each neuron in the model represents the activity of a certain functional cluster of neurons. The time is discretized in steps representing 100 ms each. As shown in Fig. 1, the input layer of the model is constituted by a set of cue selective neurons. Each time a conditioned stimuli $CS'$ is present, $CS' = 1$ , otherwise $CS' = 0$ . The unconditioned stimulus, which is an appetitive reward, $(US)$ is set to 0 unless it is present $US=6$. In behaving monkeys, short-term memory activity (STM) is observed in the last stages of the visual ventral pathway, that is, in visual cortex V4 and inferotemporal cortex (ITC) [20] and also in mPFC, where reward related stimuli are processed. As neuroanatomical data reveals [21], those STM's would be inputs to the VTA/SNc, the basal ganglia (BG) and premotor cortex (PMC) and the IPFC. In the model, STMs of conditioned and unconditioned stimuli are computed as:

\[ \tau_c(S') = (1 - \alpha) \tau_{-c}(S') + \alpha S' \quad \text{if } S' > 0 \]
\[ \tau_c(S') = (1 - \beta) \tau_{-c}(S') \quad \text{if } S' = 0 \]

(1)

where $S'$ represents the $CS$ or $US$ stimuli. Short term memories are reset during the inter-trial intervals.

Fig. 1. Scheme of the neural network model. The first layer generates short term memories of the stimuli as a result of the interaction between different structures such as medial PFC, inferotemporal cortex, hippocampus and amygdala. In the model there are 80 neurons in the IPFC, 3 in the BG-PMC and a single neuron in the VTA/SNc. The parameters used during the simulations are: $\alpha = 0.35; \beta = 0.01; \delta = 0.0007$ after execution of any response, otherwise $\delta = 0.00001$; $\theta_{bg} = 0.005$; $\theta_{bg} = \theta_{bg} - \theta_{pnc} = 0.35$; $basal_{pfc} = 0.15$; $basal_{pfc} = 0.15$; $a_e = 0.063$; $H = 0.65$; $\mu_{pfc} = 0.09996$; $\mu_{pfc} = 0.0014$; $\mu_{pfc} = 0.99997$; $\mu_{pfc} = 0.0013$.

Considering that activity changes in the BG and the PMC occur at the same time course during visual discrimination learning [22], we model the BG and the PMC as a single layer of neurons.

If throughout a trial, none of the BG-PMC neurons activity exceeds the activation threshold, a random response is executed with probability 1/3. When a response is executed, the output of the associated neuron is forced to 1 through a period of four time steps, while others outputs are forced to 0 along the same time period.

In a VD task these responses represent saccadic movements to the right ($R^1$), left ($R^2$) and to any other non-rewarded direction ($R^3$). All of them are codified at the motor related structures layer. If the executed response leads to reward, dopaminergic neurons at the VTA/SNc modify their synaptic weights, strengthening the prediction of $US$ based on the presence of the $CS$. Synaptic weights at the VTA/SNc are updated by using a Rescorla-Wagner rule [23]. This rule allows associating previously paired $CSS$-$US$ preserving blocking and overshadowing effects [23]. Although time difference models (TD) fit better the firing pattern of dopaminergic neurons during classical and operant conditioning [24], we are interested in postsynaptic effects of DA on the frontal lobe and motor-related areas. In monkeys, pairing the $CS$ and the $US$ produces phasic firing of dopaminergic neurons of VTA/SNc when the $CS$ is presented [24]. Although this phasic activation lasts a few hundred of milliseconds, postsynaptic effects of DA on the PFC and motor-related structures persist for longer periods [25][26]. The magnitude to quantify post-synaptic effects of dopamine ($P$) and the modification of VTA/SNc synaptic weights is computed as:

\[ v_i(X) = v_{-i}(X) + \delta(X(US_i - P)) \cdot X = \tau_c(CS'), M^k \]

(2)

\[ P_t = \frac{1 - 0.3 \cdot Ic_i}{1 + e^{-10(\sum_{S' \in S} n_i(S') \cdot \tau(S') + \sum_{M' \in M} n_i(M') \cdot \tau(M') - 0.3)}} \]

(3)

where $v_i(X)$ belongs to the interval [-1,1], $M^k$ represents a cluster of neurons in IPFC and $Ic_i$ represents the inhibition exerted by the LC over the excitability of dopaminergic neurons, observed in vitro and in vivo experiments [13] and whose computation will be explained below in detail. Each time the reward is delivered, $US=6$ for the following 10 steps, otherwise $US=0$. When an error occurs $P_t$ is set to 0.1, which stands for the below baseline activity observed in VTA/SNc neurons when the predicted reward is omitted [24]. VTA stimulation decreases the spontaneous firing of PFC pyramidal neurons, mainly by exciting interneurons [27]. In our model, such inhibition is represented by clamped negative synaptic weight $u_i(P)$ from the VTA to the IPFC. However, due to the synergy between NMDA and D1 receptors [8], we postulate that initially inhibited pyramidal neurons in the PFC will strongly fire when afferent inputs release large amounts of glutamate. This activated cluster will then inhibit other clusters [28]. For this reason a “winner takes all” mechanism is applied at the IPFC output [9].
following equations show the calculation for the outputs of neurons at the IPFC \((O_k^i)\).

\[
O_k^i = \sum_{CS} w_i^j (CS') \tau_i (CS') + u_i (P)P_i + B_{\text{winner}} P_i + \text{basal}_{\text{PFC}} \quad \text{if } O_k^i > 0, \text{ else } O_k^i = 0
\]

\[
M_k^i = \begin{cases} 
O_k^i \delta(k - k^*) & \text{if } O_k^i \geq \theta_{\text{PFC}} \\
O_k^i & \text{if } O_k^i < \theta_{\text{PFC}} 
\end{cases}
\]

basal_{PFC} is the baseline firing rate, \(\theta_{\text{PFC}}\) is an activation threshold, \(\delta(x)\) is the Kronecker delta function and \(k^* = \arg \max_i O_k^i\) represents the index of the winner neuron. For this neuron, \(B_{\text{winner}} = 0.14\), else \(B_{\text{winner}} = 0\), which stands for the synergism between D1 dopamine receptors and NMDA receptors.

It has been hypothesized that DA modulates the excitability of striatal neurons allowing the BG to inhibit competing programs and to release the correct one [29]. As in the IPFC, in our model the released DA inhibits the motor area through clamped negative synaptic weight \(w_i (P)\), and, in contrast to this general inhibition, the winner neuron is excited proportionally to the amount of released DA. The effect of this mechanism is to apply a “brake” over all possible motor programs and to release the one whose activity surpasses a fixed threshold. The output of response neurons is computed as:

\[
R_i^j = \sum_{CS} w_i^j (CS') \tau_i (CS') l_{ci} + \sum_{M^j} w_i^j (M^j) M^j + w_i (P)P_i + B_{\text{winner}} P_i + \text{basal}_{BG-PMC}
\]

where \(\text{basal}_{BG-PMC}\) is the baseline firing rate and \(l_{ci}\) represents a modulation exerted by noradrenergic neurons of the Locus Coeruleus (LC) over visual and somatosensory cortical effects. Effects of NA on the modulation of glutamate evoked responses has been proved to have an inverted U shape [30], that is, low and high doses of NE produce a decrease in neuron excitability while moderate doses increase it.

In behaving monkeys, tonic firing of LC neurons show a defined correlation with performance [30]. Tonic frequencies of 2-3 Hz are associated with good performance periods while frequencies >3 Hz are related with periods where erratic performance and distractibility are observed. This gives a hint about the function of the noradrenergic system in the regulation of exploratory behavior [12]. We model the tonic firing of LC neurons as a function of the received reward through a time window that includes many trials.

\[
\tau_i (US_{long}) = (1 - \alpha_c) \tau_i (US_{long}) + \alpha_c US_i
\]

\[
lc_i = 1 - 0.4 \tau_i (US_{long})
\]

Short term memories for the response neurons are computed according to

\[
\tau_i (R') = (1 - \alpha) \tau_i (R') + \alpha R_i^j
\]

and as in eq. (5) for the IPFC area, a “winner takes all” rule is applied when a response is executed. In addition to excitability, dopamine effects over PFC pyramidal neurons are also related to modifications of synaptic efficacy via LTP and LTD [31][32]. For this reason, previous models have used the DA signal in the modulation of synaptic weights modifications [15][16][18][19]. In our model IPFC and BG-PMC neurons update their synaptic weights, which belong to the interval \([0,1]\), as follows:

\[
u_i^k (CS') = \begin{cases} 
\mu_{\text{PFC}} w_i^k (CS') + \nu_{\text{PFC}} \tau_i (CS') O_k^i & \text{if } k = \text{winner} \& P_i > H \\
\mu_{\text{PFC}} w_i^{k-1} (CS') & \text{if } k = \text{winner} \& P_i < H
\end{cases}
\]

\[
w_i^k (M^j) = \mu_{BG-PMC} w_i^{k-1} (M^j) + \nu_{BG-PMC} \Phi (CS') \tau_i (R') l_{ci}
\]

\[
\Phi = \begin{cases} 
0.6 & \text{if } P_i \geq H \\
-0.6 & \text{if } P_i < H
\end{cases}
\]

where \(H\) controls the Hebbian or anti-Hebbian learning depending on the amount of released DA.

III. SIMULATIONS AND RESULTS

Fig. 2 shows the temporal sequence of stimuli presentation, response execution and reward delivery in the VD task. The subject has to associate two different cues (conditioned stimuli CS1 and CS2) with saccadic movements in opposite directions (R1 and R2). A stimulus is presented for 500 ms.

![Fig. 2. Schematic diagram of the visual discrimination task. In both cases the delay period can be set to zero. For conditioned stimuli CS’ and CS”, R' and R” represent rewarded (US) responses respectively. During a reversal process, the response indexes are exchanged.](Image)

After the stimulus offset, a delay period of 1 sec is introduced. Then the subject is allowed to respond, and if the response is correct, the reward (US) is delivered.

During acquisition, the subject has to find the appropriate responses for each CSs through an exploration process, covering the space of all possible responses. This space could be quite dense in a complex environment. Two different visual stimuli need to be associated with two
different responses. Presentation of stimulus $CS^1$ is rewarded after execution of response $R^1$ while the same occurs for stimulus $CS^2$ and $R^2$. At the beginning of the experiment all responses are executed with the same probability. As can be seen in Fig. 3, exclusion of response $R^3$ occurs during the first 50 trials of the experiment because of the Hebbian or anti-Hebbian learning occurred in BG-PMC neurons. Those curves are the result of 499 out of 500 (99.8%) averaged experiments that reach a performance higher than 90%. Although the performance at the time of $R^3$ elimination is near 50%, the subset of responses related to reward has already been selected by the motor structures. Meanwhile, synaptic weights of BG-PMC neurons acquired a weak mapping of the stimulus-response relation. This mapping is strengthened in subsequent trials where correct responses are executed and reward is delivered.

In Fig. 4 we show simulation results for performance, levels of norepinephrine ($NE$) and dopamine ($DA$) and reaction times during acquisition of the task. As the performance increases, the tonic level of norepinephrine decreases and the path from input to output layers is attenuated. The prediction of reward increases the dopamine level and promotes learning. The DA effect induces a decrease in exploration and an increase in exploitation. It is a delicate balance of information flow from the IPFC and the short-term memories layer to motor structures what is proposed here as the mechanism to modulate exploration-exploitation strategies.

Once the stimulus-response (S-R) mappings are acquired (and successfully exploited), the contingencies can be reversed (i.e., if the original ones are $CS_1-R_1$ and $CS_2-R_2$, now the reinforced ones are $CS_1-R_2$ and $CS_2-R_1$) and thus a new exploration process will be required to solve the task. To force a new exploration process, a reversal procedure is applied when the model executes 30 correct consecutive responses and the previous pairing between CSs and responses were inverted [33].

The results in Fig 5 show that due to previous learning, a time period of about 15 trials of behavioral perseveration can be observed. As the averaged received reward diminishes, tonic activity in the locus coeruleus augments. Thus, there is an increment in the information flow from the short-term memories layer to motor structures and an inhibition of the dopaminergic firing. As a consequence, this process modifies the signal to noise ratio of direct input-output path and increases the probability of executing responses randomly, restarting the exploration process.
IV. DISCUSSION

The presented computational model explains the roles and interactions between dopamine and noradrenaline in the prefrontal cortex in an operant learning task. Based on neurophysiological and neuroanatomical hypothesis, the model provides a new framework for understanding how exploration and exploitation could emerge.

In cortical and subcortical structures, dopamine is involved in the response of exploitation through two different mechanisms: neuron excitability and synaptic plasticity. In the IPFC, as dopamine is released, local interneurons inhibit pyramidal neurons. However, due to the NMDA-D1 synergism, the firing rate of pyramidal neurons that receive strong glutamatergic inputs, differentiates from the rest of the population. In motor structures, dopamine allows to reduce the search space by extinguishing responses that are not associated with reward and increasing the probability of executing rewarded responses. In the model, dopamine is also involved in learning; synaptic changes are driven by Hebbian or anti-Hebbian law depending on the level of released dopamine. Responses associated with both, reward and prediction of reward, are exploited and the synaptic paths connecting input structures, the IPFC and motor structures are reinforced.

While dopamine is involved in response exploitation, experimental data suggest the role of noradrenaline in exploration. We propose that, due to the strong afferent input from the orbitofrontal cortex (which is part of the mPFC), the modulation of the LC tonic firing can be done by reward related information. Noradrenaline released by this mechanism can modulate the signal flow from visual and somatosensory neurons to more cognitive and motor structures. In addition, in vitro and in vivo experiments show that noradrenaline modulates the excitability of dopaminergic neurons. Depending on the complexity of task, noradrenaline actions can be found at different structures and task stages. In simple operant paradigms its principal effect could be seen as a neuromodulation of dopaminergic excitability, allowing fast changes in S-R mapping as required during a reversal.

We propose that, together, dopaminergic and noradrenergic mechanisms allow switching between exploitation and exploration stages depending on the changes in the environment. Giving the experimental results obtained in simulations with changing environments, the model provides a computational framework to the exploration-exploitation dilemma for intelligent machines in changing environments.

The computational theory presented here explains how exploration and exploitation strategies emerge during operant learning and the way they are selected by concurrent dopaminergic and noradrenergic modulation at the PFC and motor related structures.

ACKNOWLEDGMENT

Supported by grants from Agencia Nacional de Promoción Científica y Tecnológica (PICT 2485), Consejo Nacional de Investigaciones Científicas y Técnicas (PIP 112-200801-02851, PIP 112 201101 01054) and Universidad de Buenos Aires (UBACYT 200 20 100 100 978).

REFERENCES


