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 whereupon 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, it induces 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

Sergio Lew is with Instituto de Ingeniería Biomédica, FI-Universidad de Buenos Aires and with IBYME-CONICET, Buenos Aires, Argentina (email: slew @fi.uba.ar)

Hernan G. Rey is with Instituto de Ingeniería Biomédica, FI-Universidad de Buenos Aires and with IBYME-CONICET, Buenos Aires, Argentina and with NeuroEngineering Lab, University of Leicester, UK (email: hgr3@le.ac.uk)

B. Silvano Zanutto is with Instituto de Ingeniería Biomédica, FI-Universidad de Buenos Aires and with IBYME-CONICET, Buenos Aires, Argentina (email: silvano@fi.uba.ar)

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^i is present, $CS^i = 1$, otherwise $CS^i = 0$. 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 STMs would be inputs to the VTA/SNc, the basal ganglia (BG) and premotor cortex (PMC) and the lPFC. In the model, STMs of conditioned and unconditioned stimuli are computed as:

$$\begin{aligned} \tau_t(S^i) &= (1-\alpha)\,\tau_{t-1}(S^i) + \alpha\, S^i \quad \text{if } S^i > 0 \\ \tau_t(S^i) &= (1-\beta)\,\tau_{t-1}(S^i) \qquad\quad \text{if } S^i = 0 \end{aligned} \tag{1}$$

where S^i represents the *CS* or *US* stimuli. Short term memories are reset during the inter-trial intervals.

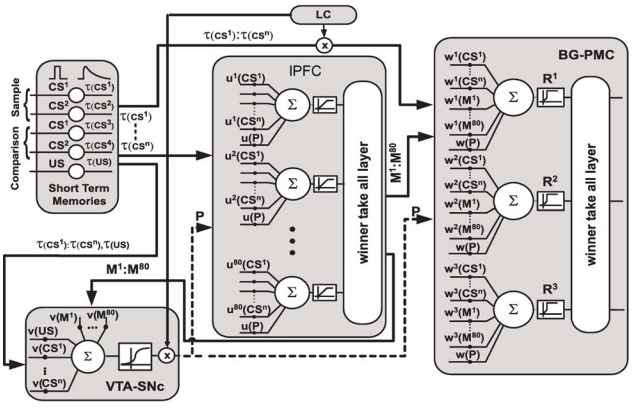

Fig. 1. Scheme of the neural network model. The first layer generates short term memories of the stimuli as a result of the interaction between different structures such as medial PFC, inferotemporal ctx., hippocampus and amygdala. In the model there are 80 neurons in the lPFC, 3 in the BG-PMC and a single neuron in the VTA/SNc. The parameters used during the simulations are: $\alpha = 0.35$; $\beta = 0.01$; $\delta_v = 0.0007$ after execution of any response, otherwise $\delta_v = 0.00001$; $u_t(P) = w_t(P) = -0.05$; $\theta_{PFC} = \theta_{BG-PMC} = 0.35$; $basal_{BG-PMC} = 0.15$; $basal_{PFC} = 0.15$; $\alpha_{lc} = 0.003$; $H = 0.6$; $\mu_{PFC} = 0.9996$; $v_{PFC} = 0.0014$; $\mu_{BG-PMC} = 0.9997$; $v_{BG-PMC} = 0.0013$.

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 *CS*s-*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:

$$v_t(X) = v_{t-1}(X) + \delta_v X(US_t - P_t), \quad X = \tau_t(CS^i), M^k \tag{2}$$

$$P_t = \frac{1 - 0.3\, lc_t}{1 + e^{-10(\sum_{\forall S^i} v_t(S^i)\tau(S^i) + \sum_{\forall M^k} v_t(M^k)\tau(M^k) - 0.3}} \tag{3}$$

where $v_t(X)$ belongs to the interval [-1,1], M^k represents a cluster of neurons in lPFC and lc_t 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 $u_t(P)$ from the VTA to the lPFC. However, due to the synergism 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 lPFC output [9]. The

following equations show the calculation for the outputs of neurons at the lPFC (O_t^k).

$$O_t^k = \sum_{\forall CS^i} u_t^k(CS^i)\tau(CS^i) + u_t(P)P_t + \\ + B_{winner}P_t + basal_{PFC} \qquad if\ O_t^k > 0, else\ O_t^k = 0 \tag{4}$$

$$M_t^k = \begin{cases} O_t^k \delta(k^* - k) & if\ O_t^{k^*} \geq \theta_{PFC} \\ O_t^k & if\ O_t^{k^*} < \theta_{PFC} \end{cases} \tag{5}$$

$basal_{PFC}$ is the baseline firing rate, θ_{PFC} is an activation threshold, $\delta(x)$ is the Kronecker delta function and $k^* = \arg\max_k O_t^k$ represents the index of the winner neuron. For this neuron, $B_{winner} = 0.14$, else $B_{winner} = 0$, 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 lPFC, in our model the released DA inhibits the motor area through clamped negative synaptic weight $w_t(P)$, 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:

$$R_t^j = \sum_{\forall CS^i} w_t^j(CS^i)\tau_t(CS^i)\, lc_t + \\ + \sum_{\forall M^k} w_t^j(M^k) M^k + w_t(P)P_t + B_{winner}P_t + basal_{BG-PMC} \tag{6}$$

where $basal_{BG-PMC}$ is the baseline firing rate and lc_t represents a modulation exerted by noradrenergic neurons of the Locus Coeruleus (LC) over visual and somatosensory cortical neurons. 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 medium 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.

$$\tau_t(US_{long}) = (1-\alpha_{lc})\,\tau_{t-1}(US_{long}) + \alpha_{lc} US_t \\ lc_t = 1 - 0.4\,\tau_t(US_{long}) \tag{7}$$

Short term memories for the response neurons are computed according to

$$\tau_t(R^j) = (1-\alpha)\,\tau_{t-1}(R^j) + \alpha R_t^j \tag{8}$$

and as in eq. (5) for the lPFC 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 lPFC and BG-PMC neurons update their synaptic weights, which belong to the interval [0,1], as follows:

$$u_t^k(CS^i) = \begin{cases} \mu_{PFC}\, u_{t-1}^k(CS^i) + \nu_{PFC}\, \tau_t(CS^i)\, O_t^k \\ \qquad if\ k = winner\ \&\ P_t > H \\ u_{t-1}^k(CS^i) \\ \qquad if\ k = winner\ \&\ P_t < H \end{cases} \tag{9}$$

$$w_t^j(CS^i) = \mu_{BG-PMC}\, w_{t-1}^j(CS^i) + \nu_{BG-PMC}\, \Phi\, \tau_t(CS^i)\, \tau_t(R^j)\, lc_t \\ w_t^j(M^k) = \mu_{BG-PMC}\, w_{t-1}^j(M^k) + \nu_{BG-PMC}\, \Phi\, M^k\, \tau_t(R^j) \tag{10}$$

$$\Phi = \begin{cases} 0.6 & if\ P_t \geq H \\ -0.6 & if\ P_t < H \end{cases} \tag{11}$$

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.

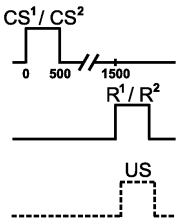

Fig. 2. Schematic diagram of the visual discrimination task. In both cases the delay period can be set to zero. For conditioned stimuli CS^1 and CS^2, R^1 and R^2 represent rewarded (*US*) responses respectively. During a reversal process, the response indexes are exchanged.

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.

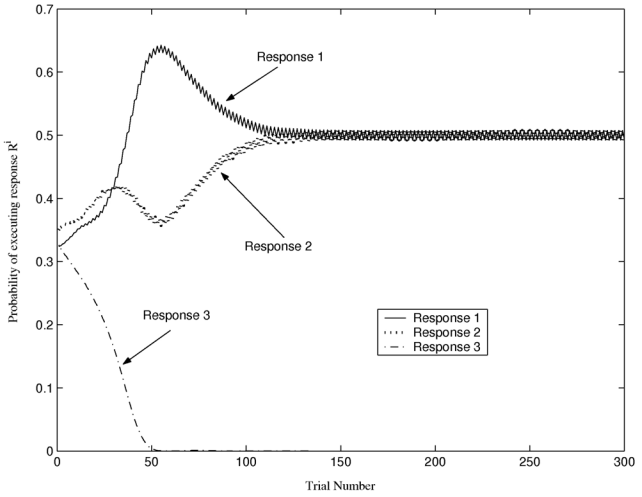

Fig. 3. Average values of relative frequency of response execution in a visual discrimination task. Response *R3* is eliminated from the set of possible responses at trial 50. As imposed in the VD task, responses *R1* and *R2* stabilize in a probability of execution of 50%

In Fig. 4 we show simulation results for performance, levels of norepinephrine (lc_t) and dopamine (P_t) 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 lPFC 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 CS1-R1 and CS2-R2, now the reinforced ones are CS1-R2 and CS2-R1) 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].

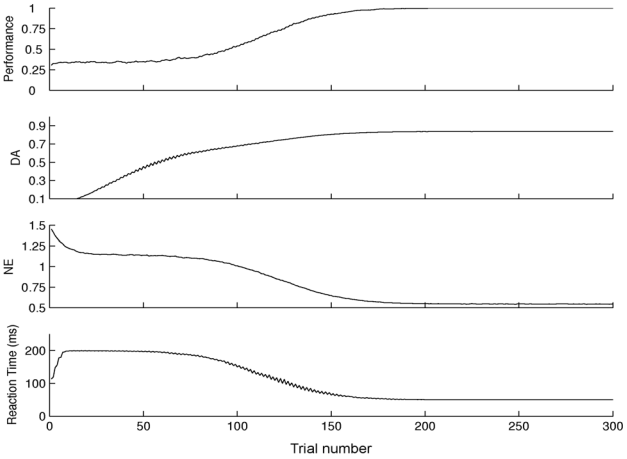

Fig. 4. For the VD task, average values from 499 out of 500 (99.8%) simulations that reached a performance of 90% were computed. Performance acquisition, DA (*Pt*) and NE (*lct*) levels, and reaction times are shown.

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.

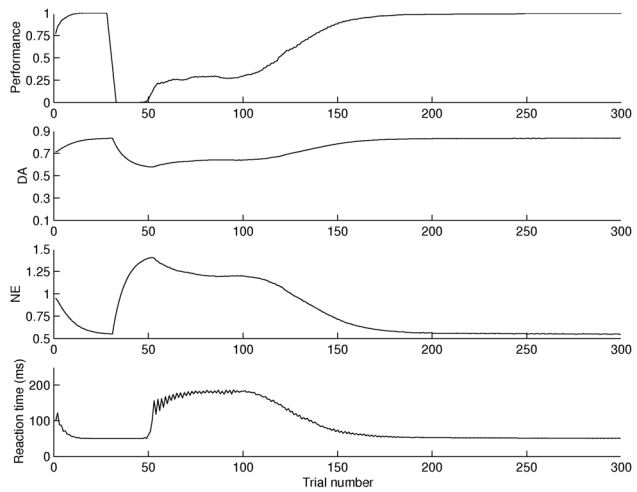

Fig. 5. For the VD task, average values from 490 out of 500 (98%) experiments that reach a performance higher than 90% were computed. Reversal procedure was applied after 30 correct trials were executed. Those average values were calculated in a range that starts 30 trials before reversal. Performance acquisition, DA (*Pt*) and NE (*lct*) levels, and reaction times are shown. A period of response preservation can be seen after reversal was applied. As performance deteriorates, NE level increases until it reaches its maximum value. LC inhibition over DA neurons forces DA level below the Hebbian/anti-Hebbian threshold. Between trials 50 and 100 the exploration process takes place, after that, the transition from exploration and exploitation is observed.

IV. Discussion

The presented computational model explains the roles and interactions between dopamine and norepinephrine 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 lPFC, as dopamine is released, local interneurons inhibit pyramidal neurons. However, due 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 lPFC and motor structures are reinforced.

While dopamine is involved in response exploitation, experimental data suggest the role of norepinephrine 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. Norepinephrine released by this mechanism can modulate the signal flow from visual and somatosensorial neurons to more cognitive and motor structures. In addition, in vitro and in vivo experiments show that norepinephrine modulates the excitability of dopaminergic neurons. Depending on the complexity of task, norepinephrine 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 Cientificas y Técnicas (PIP 112-200801-02851, PIP 112 201101 01054) and Universidad de Buenos Aires (UBACYT 200 20 100 100 978).

References

[1] S.B. Thrun, "The role of exploration in learning control", In, *Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches*, DA White and DA Sofge, Ed. New York, NY: Van Nostrand Reinhold. 1992.

[2] P. Gaudiano, and C. Chang, "Adaptive obstacle avoidance with a neural network for operant conditioning: Experiments with real robots," In *Proceedings of the 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)* pp. 13–18. Monterey, CA, 1997.

[3] C. Chang, and P. Gaudiano, "Application of biological learning theories to mobile robot avoidance and approach behaviors," *Journal of Complex Systems*, vol 1, pp. 79–114, 1998.

[4] D.A. Gutnisky, and B.S. Zanutto, " Learning obstacle avoidance with an operant behavior model," *Artificial Life*, vol. 10(4), pp. 65-81, 2004.

[5] D.A. Gutnisky, and B.S. Zanutto, "Cooperation in the iterated prisoner's dilemma is learned by operant conditioning mechanisms," *Artificial Life*, vol. 10(4), pp. 433-461, 2004b.

[6] D.A.Gutnisky, R. Zelmann, and B.S. Zanutto, "Multiagent team formation performed by operant learning: an animat approach,". *IEEE World Congress on Computat. Intelligence*, Vancouver, Canada, 2006.

[7] D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller, "Visual categorization and the primate prefrontal cortex: neurophysiology and behavior," *J Neurophysiol*., vol. 88(2), pp. 929-941, 2002.

[8] J. Wang, and P. ODonnell, "D(1) dopamine receptors potentiate NMDA-mediated excitability increase in layer V prefrontal cortical pyramidal neurons," *Cereb. Cortex*, 11(5), 452-462, 2001.

[9] A. Lavin, L. Nogueira, C.C. Lapish, R.M. Wightman, P.E. Phillips, and J.K. Seamans, "Mesocortical dopamine neurons operate in distinct temporal domains using multimodal signaling," *J. Neurosci*., vol. 25(20), pp. 5013-5023, 2005.

[10] W. Schultz, "Getting formal with dopamine and reward. Neuron," vol. 36(2), pp. 241-263, 2002.

[11] C.W. Berridge, and B.D.Waterhouse, "The locus coeruleus-noradrenergic system: modulation of behavioral state and state-dependent cognitive processes," *Brain Res. Rev*., vol. 42(1), pp. 33-84, 2003.

[12] G. Aston-Jones, and J.D. Cohen, "An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance," *Annual Review of Neuroscience*, vol. 28, pp. 403-450, 2005.

[13] C.A. Paladini, and J.T. Williams, "Noradrenergic inhibition of midbrain dopamine neurons," *J Neurosci*. Vol. 24(19), pp. 4568-4575, 2004.

[14] J. Grenho, M. Nisell, S. Ferre, G. Aston-Jones, and T.H. Svensson, "Noradrenergic modulation of midbrain dopamine cell ¯ring elicited by stimulation of the locus coeruleus in the rat*." J Neural Transm Gen Sect*., vol. 93(1), pp. 11-25, 1993.

[15] S.E. Lew, C. Wedemeyer, and B. S. Zanutto, "Role of unconditioned stimulus prediction in the operant learning: a neural network model," *Proceedings of IJCNN '01*, Washington DC, USA, pp. 331-316, 2001.

[16] S.E. Lew, H.G. Rey, D. Gutnisky, B.S. Zanutto, "Differences in prefrontal and motor structures learning dynamics depend on task complexity: a neural network model," *Neurocomputing*, vol. 71, pp. 2782–2793, 2008.

[17] M. Rapanelli , S.E. Lew , L. R. Frick , B. S. Zanutto "Plasticity in the Rat Prefrontal Cortex: Linking Gene Expression and an Operant Learning with a Computational Theory," *PLoS ONE* 5(1): e8656. doi:10.1371/journal.pone.0008656, 2010.

[18] H. G. Rey, S.E. Lew, B.S. Zanutto, "Dopamine and Norepinephrine Modulation of Cortical and Sub-cortical Dynamics During Visuomotor Learning," In *Monoaminergic Modulation of Cortical*

Excitability, K. Y. Tseng and M. Atzori, Ed. Springer, 2007, pp. 247-260.

[19] N.A. Schmajuk, and B.S. Zanutto, "Escape, avoidance, and imitation: a neural network approach," *Adaptive Behavior*, vol. 6(1), pp. 63 – 129, 1997.

[20] J.M. Fuster, and J.P. Jervey, "Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task," *J Neurosci*, vol. 2(3), pp. 361-375, 1982.

[21] C. Cavada, T. Company, J. Tejedor, R. J. Cruz-Rizzolo, and F. Reinoso-Suarez, "The anatomical connections of the macaque monkey orbitofrontal cortex," A review. *Cereb Cortex*., vol. 10(3), pp. 220-242, 2000.

[22] P. J Brasted, and S. P. Wise, "Comparison of learning-related neuronal activity in the dorsal premotor cortex and striatum," *Eur. J. Neurosci*., vol. 19(3), pp. 721-740, 2004.

[23] R. A. Rescorla, and A. R. Wagner, "A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement," In *Classical Conditioning II: Current Research and Theory*, A. H. Black and W. F. Prokasy, ED. New York: Appleton Century Crofts, 1972, pp. 64-99.

[24] P.R. Montague, S.E. Hyman, and J. D. Cohen, "Computational roles for dopamine in behavioural control," *Nature*, 431(7010), pp. 760-767, 2004.

[25] P.A. Garris, and R. M. Wightman, "Different kinetics govern dopaminergic transmission in the amygdala, prefrontal cortex, and striatum: an in vivo voltammetric study," *J. Neurosci*., vol. 14(1), pp. 442-450, 1994.

[26] F. Gonon, "Prolonged and extrasynaptic excitatory action of dopamine mediated by D1 receptors in the rat striatum in vivo," *J. Neurosci*., vol. 17(15), pp. 5972-5978, 1997.

[27] K. Y. Tseng, N. Mallet, K. L. Toreson., C. Le Moine, F. Gonon, and P. O'Donnell, "Excitatory response of prefrontal cortical fast-spiking interneurons to ventral tegmental area stimulation in vivo. *Synapse* 59, 412–417, 2006

[28] A. Compte, N. Brunel, P.S. Goldman-Rakic, and X.J. Wang, "Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model," *Cereb. Cortex*, vol. 10(9), pp. 910-923, 2000.

[29] W. Mink, "The basal ganglia: focused selection and inhibition of competing motor programs," *Prog. Neurobiol*., vol. 50(4), pp. 381-425, 1996.

[30] M. Usher, J.D. Cohen, D. Servan-Schreiber, J. Rajkowski, and G. Aston-Jones, "The role of locus coeruleus in the regulation of cognitive performance," *Science*, 283(5401), 549-554, 1999.

[31] S. Otani, H. Daniel, M.P. Roisin, and F. Crepel "Dopaminergic modulation of long-term synaptic plasticity in rat prefrontal neurons," *Cereb. Cortex*, vol. 13(11), pp. 1251-1256, 2003.

[32] J.N.J. Reynolds, and J.R Wickens, "Dopamine-dependent plasticity of corticostriatal synapses," Neural Networks, vol. 15(4-6), pp. 507-521, 2002.

[33] A. Pasupathy, and E.K. Miller, "Different time courses of learning-related activity in the prefrontal cortex and striatum," *Nature*, 433(7028), pp. 873-876, 2005.