Instrumental Conditioning Driven by Neutral Stimuli: A Model Tested with a Simulated Robotic Rat

Vincenzo G. Fiore* **Francesco Mannella*** **Marco Mirolli***
Kevin Gurney+ **Gianluca Baldassarre***
*Istituto di Scienze e Tecnologie della Cognizione,
Consiglio Nazionale delle Ricerche (ISTC-CNR),
Via San Martino della Battaglia 44, I-00185 Roma, Italy
+Department of Psychology, The University of Sheffield,
Sheffield S10 2TP, United Kigdom
vincenzo.g.fiore@gmail.com, k.gurney@sheffield.ac.uk,
{francesco.mannella, marco.mirolli, gianluca.baldassare}@istc.cnr.it

Abstract

Current models of reinforcement learning are based on the assumption that learning must be guided by rewarding (unconditioned) stimuli. On the other hand, there is empirical evidence that dopamine bursts, which are commonly considered as the reinforcement learning signals, can also be triggered by apparently neutral stimuli, and that this can lead to conditioning phenomena in absence of any rewarding stimuli. In this paper we present a computational model, based on an hypothesis proposed in Redgrave and Gurney (2006), in which dopamine release is directly triggered by the superior colliculus (a dorsal midbrain structure) when it detects novel visual stimuli and this supports instrumental conditioning similarly to that usually ascribed to rewarding stimuli. The model incorporates various biological constraints, for example the anatomical and physiological data related to the micro-architecture of the superior colliculus presented in Binns and Salt (1997). The model is validated by reproducing with a simulated robotic rat the results of an experiment with real rats on the role of intrinsically reinforcing properties of apparently neutral stimuli reported in Reed et al. (1996).

1 Introduction

Organisms' capacity for associating a rewarding event with the action that caused it was first studied by Thorndike (*Law of effect*; Thorndike, 1911). Modern physiological techniques now make it possible to record the activity of neurons in ventral midbrain while these associations are being formed. These experiments show that a short burst of dopamine ("phasic dopamine response") takes place following the rewarding event. This signal is now considered to cause the formation of the associations studied by Thorndike (Schultz, 1998).

There is a vast amount of evidence about the important role played by the dopamine (DA) as a signal associated to reward prediction in conditioning learning. For example, Schultz et al. (1997) report a widely accepted set of data showing the correlation between DA phasic activation and primary reward presentations. Nonetheless, this correlation is not static: for instance, when an animal learns to associate a conditioned stimulus (CS) with an unconditioned stimulus (US), the DA burst starts to be triggered by the appearance of the CS instead of the US. This switch in the occurrence of DA activation is considered to be the neural mechanism allowing the process of reward prediction learning. Most of these data can be modeled with the reinforcement learning temporal difference (TD) algorithm developed in the machine learning field by Sutton and Barto (1981, 1987, 1998). The essential feature of the TD learning algorithm consists in producing a signal whenever there is an error in the predictions of the rewards following the actions. This error signal modulates learning in two ways: (a) it affects the probabilities of triggering different actions in different contexts; (b) it modifies the evaluation of the current perceptual state (from which the error itself is computed) expressed in terms of expected future rewards. The hypothesis that phasic dopamine bursts correspond to the reward prediction error signal of the TD learning model has driven the collection and systematization of a wealth of empirical data, and constitutes the currently most influential view of dopamine role in conditioning experiments (Montague et al., 1996; Schultz et al., 1997; Schultz and Dickinson, 2000; Schultz, 2002).

Notwithstanding its important merits, the reward prediction error hypothesis has important limitations. A first limitation is that it does not take into account the role of internal motivations in modulating the effects of external rewards and it seems to conflate the related but different phenomena of classical/Pavlovian and instrumental/operant conditioning (see, for example, Dayan and Balleine (2002); O'Reilly et al. (2007)). A number of biologically-plausible computational models have been recently proposed that try to overcome these limitations of the standard TD model (Balkenius and Moren, 2000; O'Reilly et al., 2007; Mannella et al., 2007, 2008).

A second important limitation of the reward prediction error hypothesis is illustrated by Redgrave and Gurney (2006) who review three classes of empirical findings which seem to be in contrast with the identification of the phasic dopamine burst with the reward prediction error postulated by the TD model: (a) phasic DA responses have been recorded following stimuli with no apparent rewarding value, if these stimuli have not been previously shown to the organism: novelty causes phasic DA independently of the appetitive value of the stimulus; (b) while the time required to establish an association varies depending on both the complexity and the appetibility of the stimulus, phasic DA responses do not show any significant difference depending on these two parameters (furthermore, there is no variation across species in this respect); (c) phasic DA responses temporally precede gaze shifts (latency 70-100 ms and 150-200 ms respectively), meaning that they are too fast to be based on a complex computational analysis of the stimulus, which would be required to evaluate the true "economic value" of reward.

Based on these findings, Redgrave and Gurney (2006) proposed an alternative hypothesis on the role of phasic dopamine bursts in conditioning phenomena: namely, that dopamine represents a *sensory* prediction error signal which is critical for learning the causal relationships between an animal's own actions and their effects on the environment, irrespective of the rewarding value of those effects. Further experimental evidence seems to support this view. In fact, beyond triggering phasic dopamine responses, apparently neutral stimuli like light flashes are also able to support instrumental learning of the actions which cause those stimuli to appear (Reed et al., 1996). The adaptive significance of this dopamine-based action-outcome learning would lie in the fact that it might allow the acquisition of skills and knowledge that might be later exploited to pursue biologically relevant goals. This is explicitly shown in "response preconditioning experiments". In one example of these experiments, the acquisition of a lever-press action on the basis of a light flash, followed by a classical light-food conditioning process which gives appetitive value to the light, is able to subsequently elicit lever press responses with a much higher frequency with respect to situations in which the lever pressing action was associated with stimuli different than the valued light (St Claire-Smith and MacLaren, 1983). This outcome can be explained only in terms of knowledge acquired during the first instrumental conditioning phase in relation to the (lever-press)-(light) association.

While there are several biologically-plausible models of standard classical and instrumental conditioning which also tackle the aforementioned limits of the standard TD learning models, there are no biologically-plausible models on the role and sources of dopamine signals driving learning processes on the basis of "neutral" action outcomes. Recently, several models have been proposed to investigate the intrinsic reinforcing properties of neutral stimuli (e.g. Schmidhuber, 1991; Barto et al., 2004; Oudeyer et al., 2007; Schembri et al., 2007). However, these models have been developed within the machine learning community and so they do not incorporate relevant empirical constraints, both anatomical and physiological, available on these phenomena.

Currently, the most well developed hypothesis regarding the neural basis of instrumental conditioning guided by neutral stimuli is that of Redgrave and Gurney (2006), according to which this kind of learning depends on the triggering of the dopamine learning signal by the superior colliculus (SC), a dorsal midbrain structure. In support of this hypothesis there are four kinds of empirical evidence: (1) the neurons of the SC are specifically sensitive to changes in luminance produced by sudden appearance or disappearance of stimuli in the visual field (Wurtz and Albano, 1980); (2) anatomically, the SC provides a route from retinal ganglion cells to the dopaminergic neurons of the substantia nigra pars compacta (SNc) (Comoli et al., 2003; McHaffie et al., 2006); (3) SC latencies to the appearance of visual stimuli are always shorter than those in SNc (Comoli et al., 2003; Coizet et al., 2003); (4) lesioning the SC stops SNc's responses to luminance changes whilst lesioning visual cortex does not (Dommett et al., 2005; Katsuta and Isa, 2003).

This work proposes, for the first time, a computational model that accounts for some of these data. In particular the model shows how a learning rule which includes influences from phasic dopamine, ongoing motor activity and contextual (visual) input can lead, qualitatively, to the kind of behavioural patterns observed in Reed et al. (1996). Activation of phasic dopamine is triggered via the superior colliculus (SC) which has superficial layers of SC with a microstructure as the one proposed by Binns and Salt (1997) on the basis of neuro-anatomical and neurophysiological data. The whole model is validated by reproducing with a simulated robotic rat (some of) the empirical results of experiments on real rats reported in Reed et al. (1996).

The rest of the paper is organised as follows: Sect. 2 reports the original experiments addressed by the model; Sect. 3 describes the simulated robotic rat and environment used to test the model; Sect. 4 contains a detailed description of the model; Sect. 5 reports the results of the tests; finally, Sect. 6 concludes the paper.

2 Target experiment

The model presented in this paper, described in Sect. 4, is meant to reproduce the results on the role of intrinsically reinforcing properties of apparently neutral stimuli reported in Reed et al. (1996), in particular those of "experiment 4" which was organised as follows. Eight rats were set in an operant conditioning chamber containing a light located on the ceiling (the source of the neutral stimulus) and two levers. No appetitive rewards, such as food or water, were delivered to the rats during the whole experiment. The pressure of lever 1 caused an onset of the light lasting $2s$. In this regards, the experiment used a *variable interval schedule*: the light flash followed the lever-1 pressure only if this was performed after the end of a variable interval (VI). This interval ranged in $(1, 120)s$ and started from the beginning of the test or the last light flash. The whole test lasted eight sessions of $25min$ each (for simplicity, here the experiment consists in only one $25min$ session).

The results of the test shows that the number of pressings of the lever associated with light significantly increase with learning. In particular, despite the absence of primary rewards, the rats clearly change their disposition to pressing the two levers: the pressure ratio changes to approximately 4:1 in favour of the target lever (Reed et al., 1996, p. 43).

3 The simulated environment, robot and experiments

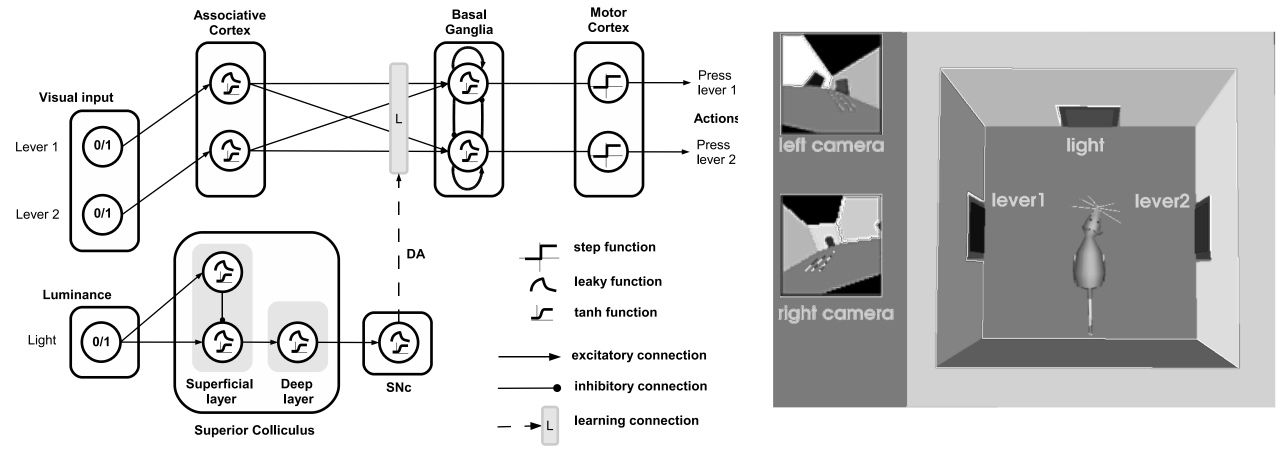
The neural model presented here was tested with a simulated robotic rat implemented on the basis of the 3D physical world simulator Webots^{TM}. The program implementing the model is written in Matlab^{TM} and interfaced with Webots^{TM} through a TCP/IP connection.

The environment is formed by a grey-walled box containing two levers and a light represented with rectangles having different colours. A pressure of lever 1 is followed by a light stimulus, lasting $2s$, in the case the current variable interval has elapsed, otherwise it has no effect (the variable intervals start from the last flash light and have a random duration of $(1, 120)s$. A pressure of lever 2 has no effect.

The simulated robot's chassis roughly reproduces the body of a rat on the basis of cylinders and cones. The robot is endowed with two wheels controlled by two independent motors. The robot is also equipped with two 64×64 pixel cameras each having a 155 degrees pan field (the fields of the two cameras overlap 10 degrees at the robot's front). Three "abstract sensors" are used to encode information about the levers and the light. In particular, the first two ($l1$ and $l2$) encode in a binary fashion the presence/absence of lever 1 and lever 2, and the third encodes in a binary fashion the presence/absence of the light. The rat has also three whiskers on each side of the head, used for obstacle avoidance (see below). Each whisker is implemented with a thin cylinder attached to the robot's head with a spring joint: the angle of the joint is used as the information that the sensor returns to the robot.

The behaviour of the rat is based on three hardwired routines: (1) "Obstacle avoidance": whenever a whisker detects a contact with an obstacle, the routine is invoked and causes the rat to move away from it on the basis of the activation of all whiskers; (2) "Press lever": whenever one of the two units of the motor cortex of the model gets activated (this is the output component of the model, see eq. 3 below), the robot approaches the corresponding lever on the basis of the position of such lever on the robot cameras' images and presses it (by hitting it with any part of the body). (3) "Explore": in all other cases, the rat moves straight ahead (together with the obstacle avoidance routine, this causes the rat to randomly explore the environment).

This work has been carried out within a broad research agenda aiming to produce models that are not only biologically plausible, but are also capable of being embedded in autonomous agents tackling realistic scenarios. The rational for this approach is that closing the agent-environment-agent loop forces us to consider behaviourally relevant time series of inputs to the model and to interpret model outputs in terms of explicit behaviour, rather than abstracting their significance *ad hoc*. Further, we are confronted with the need to design systems having all the components necessary to enable them to function correctly in a complete sensorimotor interaction with the environment. This process may suggest mechanisms that are, perforce, required to perform the function in the model, and whose existence may therefore be predicted in the animal. However, an effect of this strategy is that, for simplicity, much of the overall model (especially that required to evaluate sensory input and produce motor output) may be represented in a way that has only comparatively weak links with the



(a)

(b)

Figure 1: (a) The neural architecture of the model. (b) The simulated environment and rat. The three rectangles on the walls represents the two levers and the light source. The two square images on the left of the graph are the images perceived by the two cameras of the rat.

biology (e.g using localist highly-abstract representations as input signals and hardwired low-level behavioural routines). In spite of these simplifications, the model's core aspects, such as its overall architecture and learning mechanisms, were designed in a biologically relevant form. We think that even this shallow level of embodiment helps framing models in the right perspective. For example, in this paper the variable duration of the experiment phases, dependent on the noisy interplay between the rat and the environment, posed important challenges to the robustness of the model with respect to stimuli's duration and delays.

4 The model

Fig. 1 shows the neural architecture of the model. As mentioned in Sect. 3, the input is formed by three binary sensors encoding the presence/absence of the two levers and the light (L1, L2, L). The output of the model is formed by the two neurons of the motor cortex that trigger the execution of the lever-press routine targeting either one of the two levers. The model is composed of two subsystems: (1) the cortical pathway subsystem which propagates the signals in sequence from the associative cortex (AC) to the basal ganglia (BG) and then to the motor cortex (MC), and implements action selection; (2) the superior colliculus/dopamine subsystem which produces the dopamine learning signal used to update the AC-BG connection weights depending on the light stimulus.

In what follows, τ_x denotes the decay rate of a leaky neuron x, the sub-index x_p denotes the activation potential of neuron x, symbols like $\mathbf{X}$, $\mathbf{x}$, and x denote matrices, column vectors and scalars respectively (transposed matrices and vectors are denoted $\mathbf{X}'$ and $\mathbf{x}'$ respectively). Functions and their derivatives are represented as $f[x]$ and $\dot{f}[x]$, respectively. The functions *pos* and *step* are respectively defined as $pos[x] = \max[0, x]$ and as $step[x] = 1$ if $x \geq 0$ and $step[x] = 0$ otherwise. Hyperbolic tangents are represented as $tanh[x]$. The values of the model's parameters are listed at the end of the section.

4.1 The cortical pathway

The cortical subsystem is conceived as an action selector implementing operant conditioning. This choice is in line with proposing that basal ganglia are the locus of learned rigid sensorimotor S-R associations ("habits"; Yin and Knowlton, 2006).

The cortical pathway receives input signals from the visual sensors related to the levers (L1, L2) and provides as output one of the two possible motor actions, "press lever 1" or "press lever 2". The signals coming from the visual input (*inp*) are processed by the leaky neurons of the AC (*ac*), having an hyperbolic tangent transfer function:

$$\tau_{\mathbf{ac}} \cdot \dot{\mathbf{ac}}_p = -\mathbf{ac}_p + \mathbf{inp} \qquad (1)$$
$$\mathbf{ac} = pos[tanh[\mathbf{ac}_p]]$$

These signals are then propagated from AC to BG via all-to-all connections $\mathbf{W}_{ac-bg}$. BG are formed by two neurons $\mathbf{bg}$ which implement a bias-competition mechanism on the basis of reciprocal inhibitory connections and the "bias signals" (or "votes") from AC:

$$\tau_{\mathbf{bg}} \cdot \dot{\mathbf{bg}}_p = -\mathbf{bg}_p + \qquad (2)$$
$$(\mathbf{W}_{ac-bg} \cdot \mathbf{ac} + \mathbf{bg}_{bl} + \mathbf{n}) + \mathbf{W}_{bg} \cdot \mathbf{bg}$$

$$\mathbf{bg} = pos[tanh[\mathbf{bg}_p]]$$

where $\mathbf{bg}_{bl}$ is a baseline activation, $\mathbf{n}$ is a noise vector with components uniformly drawn in $[-n, n]$ every $4s$, and $\mathbf{W}_{bg}$ is the matrix of the BG's lateral connection weights.

The competition resulting from this dynamics allows only one of the two units of BG to activate the corresponding neuron of MC via one-to-one connections when its threshold th_{mc} is overcome. To this purpose, the units of MC, mc, have a step transfer function with threshold:

$$\mathbf{mc} = step[\mathbf{bg} - th_{mc}] \tag{3}$$

All the connections weights of the cortical pathway are fixed with the exception of AC-BG ones which are modified with an Hebb rule modulated by the dopamine signal from the DA-system (see Sect. 4.2):

$$\Delta\mathbf{W}_{ac-bg} = \eta_{ac-bg} \cdot pos[da - th_{da}] \cdot \mathbf{mc} \cdot \mathbf{ac}' \tag{4}$$

where η_{ac-bg} is a learning coefficient and th_{da} ensures that only phasic DA bursts, and not "background" (tonic) DA, cause learning. The quantities **mc** and **ac** correspond to signals hypothesised in the scheme proposed by Redgrave and Gurney (2006). Thus **mc** represents the "motor-efference copy" – information about the action performed just prior to the light stimulus – and **ac** represents "context".

4.2 The superior colliculus and dopaminergic system

The superior colliculus (SC) is assumed to detect sudden *luminance variations*. In primates, the superficial layer of the SC receives input signals related almost exclusively to visual data, while the deep layers also receive auditory and somatosensory stimuli (Wallace et al., 1998).

The architecture of the simulated SC, which captures the essential features of the micro-anatomy of the SC as reported in Binns and Salt (1997), consists of two layers of neurons. The superficial layer (SCS) receives afferent signals from the robot receptors and the "deep layer" (SCD) sends an efferent signal to the SNc. Both layers are composed of leaky neurons with a hyperbolic tangent transfer function. SCS is formed by two neurons, *sc_si* and *sc_se*, and SCD is formed by one neuron, *sc_d*. The two units of SCS receive a connection from the same luminance sensor L, denoted with l. The first neuron, *sc_si*, sends an inhibitory connection to the second neuron of the same layer, *sc_se*, whereas the latter sends an excitatory connection to the neuron of the deep layer *sc_d*. The interplay between *sc_si* and *sc_se*, having respectively a slow and a fast time constant τ (see parameters below), cause an activation of *sc_d* maximally responsive to luminance increases. More in detail, the inhibitory unit *sc_si* activates as follows:

$$\tau_{sc\_si} \cdot \dot{sc\_si}_p = -sc\_si_p + w_{l-sc\_si} \cdot l \tag{5}$$

$$sc\_si = pos[tanh[sc\_si_p]]$$

where w_{l-sc_si} is the connection weight between the sensor L and *sc_si*.

The excitatory unit *sc_se* activates as follows:

$$\tau_{sc\_se} \cdot \dot{sc\_se}_p = -sc\_se_p + (w_{l-sc\_se} \cdot l - w_{sc\_si-sc\_se} \cdot sc\_si) \tag{6}$$

$$sc\_se = pos[tanh[sc\_se_p]]$$

where w_{l-sc_se} is the connection weight between the sensor L and *sc_se*, and $w_{sc\_si-sc\_se}$ is the lateral inhibitory connection weight of SCS.

The neuron of the deep layer, *sc_d*, receives signals from *sc_se* via the weight $w_{sc\_se-sc\_d}$ and processes them as follows:

$$\tau_{sc\_d} \cdot \dot{sc\_d}_p = -sc\_d_p + w_{sc\_se-sc\_d} \cdot sc\_se \tag{7}$$

$$sc\_d = pos[tanh[sc\_d_p]]$$

The SC output neuron *sc_d* triggers dopamine bursts *da* in SNc:

$$\tau_{da} \cdot \dot{da}_p = -da_p + (w_{sc-da} \cdot sc\_d) \tag{8}$$

$$da = pos[tanh[da_p]]$$

where w_{sc-da} is the connection weight between SC and SNc. The dopamine signal so produced modulates the updating of the connection weights linking AC to BG (see equation 4). As a result of this mechanism, luminance variations detected by the SC can change the way signals are propagated through the cortical pathway, thus influencing action selection in favour of actions which cause the luminance variations themselves.

The parameters of the model were set as follows. Decay coefficients: $\tau_{\mathbf{ac}} = 600$; $\tau_{\mathbf{bg}} = 300$; $\tau_{sc\_si} = 2000$; $\tau_{sc\_se} = 300$; $\tau_{sc\_d} = 300$; $\tau_{da} = 300$. Thresholds: $th_{mc} = 0.6$; $th_{da} = 0.6$. Learning coefficients: $\eta_{ac-bg} = 0.01$. Connection weights related to SC: $w_{l-sc\_se} = 2$; $w_{l-sc\_si} = 3$; $w_{sc\_si-sc\_se} = 2$; $w_{sc\_se-sc\_d} = 1$; $w_{sc-da} = 2.3$; BG lateral connections: $\mathbf{W}_{bg} = \begin{pmatrix} 0 & 0.7 \\ 0.7 & 0 \end{pmatrix}$. The trained connections between AC and BG were initially set to 0. All other connections were set to 1. Other parameters: $bg_{bl} = 0.15$; $n = 0.4$.

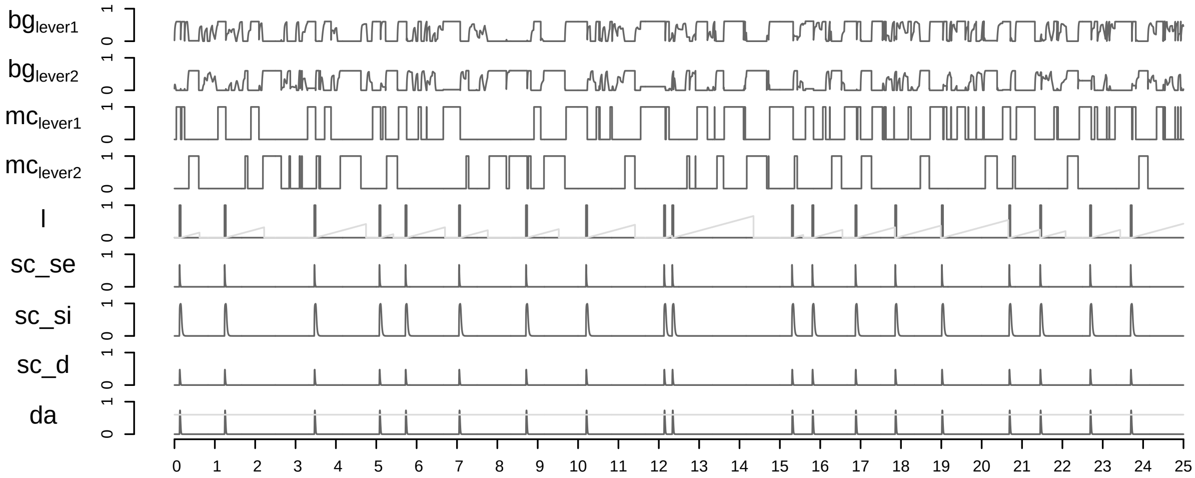


Figure 2: Recordings of the activation of some key units of the model during a test lasting $25min$ and simulating "experiment 4" reported in Reed et al. (1996). The first two rows of the graph show the activation of the two neurons of the BG (**bg**). The third and fourth layers show activations of the MC (**mc**). The fifth layer represents the light stimulus signal and the durations of the VIs ("saw-like" line) starting from the last light signal. The sixth, seventh and eighth rows report the activations of SC (sc_se, sc_si and sc_d). The last row reports the dopamine bursts (da): the gray horizontal line indicates the threshold (th_{da}) that dopamine has to overcome to cause learning.

5 Results

Figure 2 and 3 show the activation of some key neurons of the model during the simulation of experiment 4 reported in Reed et al. (1996), lasting $25min$. Figure 3 shows that the activation that the two BG neurons (first and second row) exhibit in time correspond to the competition between the selection of the two lever-press actions: the first neuron which sends an activation to the corresponding unit in MC, and makes it overcome the threshold th_{mc} (third and fourth row), triggers the corresponding action. Note that the derivative discontinuities of the activation of the BG's units are due either to the noise reset, happening every $4s$, or to the reset of such units, happening after every action execution. The figure shows in particular the triggering of three actions: (1) "Press lever 2": this has no effects on the light; (2) "Press lever 1": this causes a light onset as it is accomplished after the last VI terminates (the delay between the activation of mc_lever1 and the light onset is caused by the fact that the rat takes some time to approach the lever); (3) "Press lever 1": this has no effects on the light as it is accomplished before the current VI terminates.

Figure 2 shows that the rat selects actions randomly during the first minutes of the test but then, as learning goes on, it increases the selection of lever 1 steadily. In particular, the ratio of lever-1 presses vs. lever-2 presses passes from 14:15 in the first five minutes of the test to 34:8 in the last five minutes (average over ten simulated rats). The latter ratio is very similar to that of real rats (Reed et al., 1996), approximately 4:1 (Figure 4).

These results demonstrate that the model succeeds in exploiting what, in a biological setting, may be interpreted as a *neutral stimulus* to trigger learning.

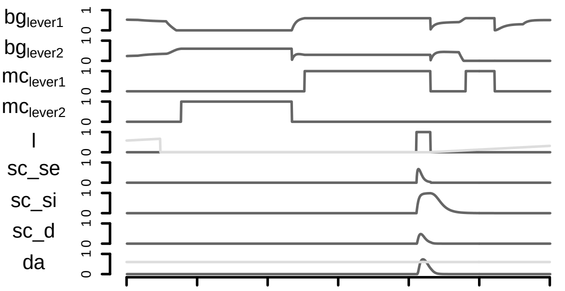


Figure 3: A close-up of Figure 2. X-axis: time in seconds.

This learning process is made possible by the interplay of the SC with the dopamine system (SNc); figure 3 shows this process in more detail. Thus, when the SC detects a variation of the light luminance (see the "light" row of the graph) it causes DA bursts via the SNc (see "da" row in the graph). A light increase causes an activation of sc_se which in turn causes a strong activation of da via sc_d. However, the activation of sc_se is soon inhibited by the activation of sc_si so that the DA production goes below threshold: in this way the SC manages to associate DA only to the *luminance variation* and not to its *level*. Overall, these mechanisms allow the SC to exploit its responsiveness to light variations to trigger dopamine bursts. These signals modify the synapses within the basal ganglia so increasing the probabilities that the same actions are performed in the future.

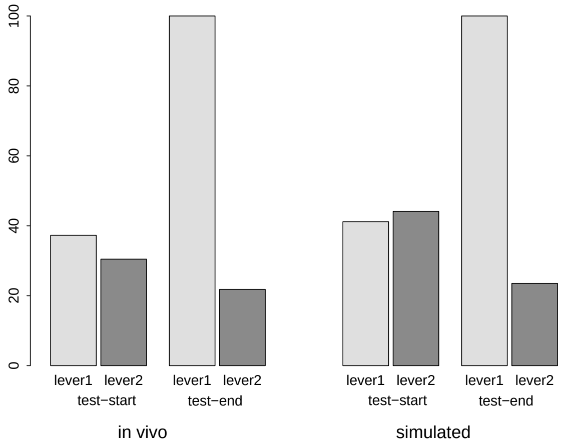


Figure 4: Lever-1 presses vs. lever-2 presses in real and simulated rats at the beginning and end of the test.

6 Conclusions

This paper presented a computational model of how the acquisition of conditioned responses can be driven by neutral stimuli. In particular, the model's architecture incorporates the hypothesis that a phasic dopamine burst (emanating from SNc) is triggered by the detection of luminance changes by the SC associated with an unpredicted stimulus (Redgrave and Gurney, 2006). The model of SC implements the micro-architecture postulated by Binns and Salt (1997) on the basis of anatomical and physiological data, and is indeed capable of detecting rapid increases in luminance in an analogous way to that shown *in vivo* (Wurtz and Albano, 1980). Moreover, the learning rule (equation 4) makes use of excitatory afferent signals to basal ganglia invoked in the hypothesis of Redgrave and Gurney (2006), namely contextual input and motor-efference copy.

The model integrates these assumptions into a complete architecture which has been successfully used to control a simulated 3D robotic rat interacting with a realistic simulated environment. Further, the model was validated by reproducing the results of a biological experiment dealing with action learning in a paradigm which is instrumental/operant in nature but which uses no explicit reward (appetitive stimulus). The results support the hypothesis that the dopaminergic signals triggered by the superior colliculus play a central role in learning the causal relationships between an animal's actions and their environmental outcomes (Redgrave and Gurney, 2006).

Future work will develop the model to demonstrate how learning of action-outcome contingencies might facilitate subsequent accomplishment of biologically relevant goals as in, for example, the response preconditioning experiments of St Claire-Smith and MacLaren (1983), described in the Introduction.

Future work might also tackle two limitations of the current SC model: (1) The SC model is sensitive only to increases of luminance, but not to *decreases*, as the real SC. This simplification was used as a more complex SC was not needed for the targeted experiments. Computationally, it should be possible to build a SC sensitive to both luminance increases and decreases on the basis of the current SC input module and a second similar input module with an inhibitory unit having a time delay *faster* than the one of the excitatory unit. (2) In this work, the light is an abstract representation of the presence/absence of the light stimulus, so it was not necessary to have a SC with a topological architecture responding to different *positions* of the luminance variation in the environment as it happens in the real SC. Computationally, it should be possible to build a SC sensitive to the location of luminance variation on the basis of multiple copies of the current module of the SC so as to form a 2D map of modules topologically corresponding to the retina.

Acknowledgements

This research was supported by the EU Projects *ICEA*, contract no. FP6-IST-027819-IP. We thank Prof. Peter Redgrave for the important feedbacks and discussions on the paper.

References

Balkenius, C. and Moren, J. (2000). Emotional learning: a computational model of the amygdala. *Cybernetics and Systems*, 32(6):611–636.

Barto, A., Singh, S., and Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. In *International Conference on Developmental Learning (ICDL)*, LaJolla, CA.

Binns, K. E. and Salt, T. E. (1997). Different roles for gaba(a) and gaba(b) receptors in visual processing in the rat superior colliculus. *Journal of Physiology*, 504:629–639.

Coizet, V., Comoli, E., Westby, G. M., and Redgrave, P. (2003). Phasic activation of substantia nigra and the ventral tegmental area by chemical stimulation of the superior colliculus: an electrophysiological investigation in the rat. *European Journal of Neuroscience*, 17(1):28–40.

Comoli, E., Coizet, V., Boyes, J., Bolam, J. P., Canteras, N. S., Quirk, R. H., Overton, P. G., and Redgrave, P. (2003). A direct projection from superior colliculus to substantia nigra for detecting salient visual events. *Nature Neuroscience*, 6(9):974–980.

Dayan, P. and Balleine, B. (2002). Reward, motivation and reinforcement learning. *Neuron*, 36:285–298.

Dommett, E., Coizet, V., Blaha, C. D., Martindale, J., Lefebvre, V., Walton, N., Mayhew, J. E. W., Overton, P. G., and Redgrave, P. (2005). How visual stimuli activate dopaminergic neurons at short latency. *Science*, 307(5714):1476–1479.

Katsuta, H. and Isa, T. (2003). Release from gaba(a) receptor-mediated inhibition unmasks interlaminar connection within superior colliculus in anesthetized adult rats. *Neuroscience Research*, 46(1):73–83.

Mannella, F., Mirolli, M., and Baldassarre, G. (2007). The role of amygdala in devaluation: a model tested with a simulated robot. In Berthouze, L., Prince, C. G., Littman, M., Kozima, H., and Balkenius, C., (Eds.), *Proceedings of the Seventh International Conference on Epigenetic Robotics*, pages 77–84. Lund University Cognitive Studies.

Mannella, F., Zappacosta, S., Mirolli, M., and Baldassarre, G. (2008). A computational model of the amygdala nuclei's role in second order conditioning. In *From Animals to Animats 10: Proceedings of the 10th International Conference on the Simulation of Adaptive Behaviour*.

McHaffie, J. G., Jiang, H., May, P. J., Coizet, V., Overton, P. G., Stein, B. E., and Redgrave, P. (2006). A direct projection from superior colliculus to substantia nigra pars compacta in the cat. *Neuroscience*, 138(1):221–234.

Montague, P., Dayan, P., and Sejnowski, T. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, 16(5):1936–1947.

O'Reilly, R. C., Frank, M. J., Hazy, T. E., and Watz, B. (2007). Pvlv: The primary value and learned value pavlovian learning algorithm. *Behavioral Neuroscience*, 121(1):31–49.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(1):265–286.

Redgrave, P. and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nature Reviews Neuroscience*, 7(12):967–975.

Reed, P., Mitchell, C., and Nokes, T. (1996). Intrinsic reinforcing properties of putatively neutral stimuli in an instrumental two-lever discrimination task. *Animal Learning and Behavior*, 24:38–45.

Schembri, M., Mirolli, M., and Baldassarre, G. (2007). Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot. In Demiris, Y., Mareschal, D., Scassellati, B., and Weng, J., (Eds.), *Proceedings of the 6th International Conference on Development and Learning (ICDL)*, pages E1–6, London. Imperial College.

Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. In Meyer, J.-A. and Wilson, S., (Eds.), *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, pages 222–227, Cambridge, MA. MIT Press.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80(1):1–27.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, 36:241–263.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

Schultz, W. and Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Reviews Neuroscience*, 23:473–500.

St Claire-Smith, R. and MacLaren, D. (1983). Response preconditioning effects. *Journal of Experimental Psychology*, 9:41–48.

Sutton, R. S. and Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88:135–140.

Sutton, R. S. and Barto, A. G. (1987). A temporal-difference model of classical conditioning. In *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, pages 355–378, Mahwah, NJ. Lawrence Erlbaum Associates.

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

Thorndike, E. L. (1911). *Animal Intelligence.* Transaction Publishers, Rutgers, NJ.

Wallace, M. T., Meredith, M. A., and Stein, B. E. (1998). Multisensory integration in the superior colliculus of the alert cat. *The Journal of Neurophysiology*, 80:1006–1010.

Wurtz, R. H. and Albano, J. E. (1980). Visual-motor function of the primate superior colliculus. *Annual Review of Neuroscience*, 3:189–226.

Yin, H. H. and Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, 7:464–476.