Proceedings of the Eighth International Conference on Intelligent Autonomous Systems IAS-8, Amsterdam, The Netherlands, 10-13 March 2004.

Comparing three Critic Models of Reinforcement Learning in the Basal Ganglia Connected to a Detailed Actor in a S-R Task

Mehdi Khamassi[1,2], Benoît Girard[1], Alain Berthoz[2], Agnès Guillot[1]
[1] AnimatLab/Laboratoire d'Informatique de Paris 6, Université Paris 6, Paris
[2] Laboratoire de Physiologie de la Perception et de l'Action, Collège de France, Paris
corresponding author: mehdi.khamassi@poleia.lip6.fr

Abstract. Actor-Critic architectures of reinforcement learning were found to show a strong resemblance with known anatomy and function of a part of the vertebrate's brain: the basal ganglia. Based on this analogy, a large number of Actor-Critic models were simulated to reproduce behaviours of rats performing laboratory tasks. However, most of these models were tested in different tasks and it is often difficult to compare their efficiency. The work presented here concerns the comparison of three Critics, tested with the same Actor part taking into account known basal ganglia anatomy. The specificities of the three Critics connected to this Actor lie on the absence or presence of a temporal representation of stimuli, and on the use of one or more units for the prediction of reward. These architectures are implemented in the same simulated robot performing the same stimulus-response (S-R) experiment: a reward-seeking task. Results show that both temporal representation of stimuli and multiple prediction units are mandatory for the achievement of the task. Improvements for Critic modelling and competing hypotheses for Actor-Critic models are finally discussed.

Introduction

In the frame of Artificial Intelligence, Actor-Critic architectures have been proposed to provide reinforcement learning algorithms by which "embedded agents" can improve performance and show autonomous behaviour while acting in complex dynamic environments [3; 13]. In such architectures, an Actor network learns to select the right actions in the right context so as to maximize the weighted sum of future rewards [3]. The Critic network is designed to compute a prediction of this sum at each timestep based on the current sensory input and the Actor's policy. By means of an iterative process by which the Critic compares its own predictions to the actual rewards obtained by the agent, the Critic can induce a positive or negative reinforcement signal. This signal is used to improve the Actor's policy and to increase the Critic's capacity to estimate correctly the weighted sum of future rewards. The learning rule used by this adaptive Critic is the TD learning rule [13] in which the error between two adjacent predictions (the TD error) is used to update both the Actor's and the Critic's synaptic weights:

$$w^i \leftarrow w^i + \eta.\check{r}^t.E^{it} \qquad (1)$$

where w^i is a synaptic weight, $\eta > 0$ is the learning rate, E^{it} is the input value of the synapse w^i at time t, and $\check{r}^t$ is the reinforcement signal given by the equation of the Temporal Difference error:

$$\check{r}^t = r^t + g.P^t - P^{t-1} \qquad (2)$$

where g is the discount factor ($0 < g < 1$) which determines how far in the future expected rewards are taken into account in the weighted sum of future rewards.

Actor-Critic models of the basal ganglia have been proposed since the TD error has shown a strong resemblance with the dopamine signal projecting to this part of the vertebrate's brain [6]. There is now large evidence that dopamine is a neuromodulator which can control synaptic plasticity of the basal ganglia's input layer - the striatum - and thus can be considered as a neural reinforcement signal. Moreover, electrophysiological studies of monkey dopamine neurons projecting to the striatum have shown that the dopamine signal have the same pattern of activity than the TD error signal described by equation 2 [10]:

- It responds to unexpected rewards (unconditioned stimuli), producing a signal of positive reinforcement when something better than predicted occurs ($[r^t+g.P^t]>P^{t-1}$).
- After learning, when reward is predicted by a stimulus (conditioned stimulus), dopamine neurons do not discharge at reward occurrence but anticipate it by responding to the conditioned stimulus.
- Dopamine neurons depress their activity when an expected reward does not actually occur, producing a signal of negative reinforcement when something is worst than predicted.

This analogy lead neurobiologists to assess that basal ganglia could play the function of an Actor-Critic, and helped them understand the functional distinction between two anatomically distinct substructures of the basal ganglia: one in dorsal striatum selecting motor orders, and the other in ventral striatum, particularly in the nucleus accumbens, controlling reinforcement learning by projections to dopamine neurons [6; 11].

Based on this analogy, a large number of Actor-Critic models of information processing in the basal ganglia have been developed in recent years. A recent review highlights that most of these models describe the Actor part at the striatal level without taking into account known anatomy of this neural structure [7]. These models usually implement an Actor only constituted of a series of competitive components, each one representing a specific action. The action that gives the higher response to a sensorial input wins the competition and is selected. Besides, the Critic parts of these models implement different solutions to reproduce the TD-algorithm. They mainly differ in the way in which the temporal dynamics of dopamine firing are reproduced [7], and in the mechanism implemented to coordinate several units that compute the TD error.

However, it is difficult to compare the efficiency of these models for they were tested in different tasks (e.g., sensorimotor associations, navigation tasks or mental tasks like the Wisconsin Card Sorting Test). It is also difficult to judge their adequacy for modelling reinforcement learning in animals, as they were often tested in non realistic environmental conditions, characterized by a discrete state space and instantaneous reward deliveries.

The objective of this work is to compare, in the same continuous state space task - a S-R learning classically performed by real animals -, the efficiency of three different Critics connected to a same Actor. This Actor is chosen to be more inspired by the dorsal striatum circuitry than the previous ones. The criteria chosen for the comparison are related to the specificities of the Critics, i.e. the accuracy of the predictions computation and the ability to extend learning to the whole experimental environment using one or more prediction units.

1 Methods

1.1 The Environment and the Simulated Robot

The model was implemented in a simulated robot performing a task in a 2D plus-maze (figure 1). This task was chosen in relation with a neurobiological experiment that will serve as a future validation for the model [1]. This is why, notwithstanding the virtual aspect of this work, we will employ terms equivalent to the ones describing the real experiment.

The maze is constituted of four arms, at the end of which are boxes containing a water trough and a lamp. Reward delivery is available only at boxes having their lamp on. Only one box is illuminated at a time (white box on figure 1), and the robot has to learn going to this box to get a non instantaneous reward - i.e. three drops of water, each one being separated from the others by a one second bin. When reward consumption is over, the lamp is turned off, and the robot has to learn to go back to the center of the maze to trigger the onset of another box's lamp. The model will have to build a sequence of stimulus-response associations to allow the robot reaching the lighted box from any starting point in the maze.

At each timestep, the robot's perception is reduced to a 36 colours table, giving a specific colour for 36 different directions (figure 1, upper right). The robot only sees walls (black), the lit box (white), the darkened boxes (dark grey) and the center of the maze which is represented by a grey cross (figure 1, left). Twelve sensorial variables ($0 < var^i < 1$; $0 < i < 12$) are computed out of the colour table to constitute the inputs to both the Actor and the Critic parts of the model (figure 2). Variables are computed as following:

- seeWhite (resp. seeGrey, seeDarkGrey) = 1 if the colour table contains the value 255 (resp. 191, 127).
- angleWhite, angleGrey, angleDarkGrey = (number of boxes in the colour table between the robot's head direction and the desired colour) / 18.
- distanceWhite, distanceGrey, distanceDarkGrey = (number of consecutive boxes in the colour table containing the desired colour) / 18.
- nearWhite (resp. nearGrey, nearDarkGrey) = 1 - distanceWhite (resp. distanceGrey, distanceDarkGrey).

Representing the environment with such continuous variables, and not with discrete events pre-defined by the experimenter, implies for the model to permanently receive a flow of sensorial information and having to learn to detect autonomously events that can be relevant for the task resolution. In this way, the task used for this work is in continuous state space. This characteristic and the use of non instantaneous rewards make the task more realistic than the ones chosen by the previous authors.

The robot has a repertoire of 8 actions, which constitute the outputs of the model. Four of these actions are expected to be relevant for the resolution of the task: *drinking, moving forward, turning to white perception, turning to grey perception*. The model has to learn to associate specific values of sensorial variables with these appropriate actions. For examples, the behaviour *drinking* will have to be associated with *seeWhite* and *nearWhite*, whereas *moving forward* will have to be strongly associated with *seeWhite* and *distanceWhite* and weakly associated to *angleWhite*. The other behaviours - *random exploration, turning to dark grey perception, eating, waiting* - are used to test if learning would avoid reinforcing them.

1.2 The Actor

The architecture we implemented in the Actor part replaces the simple winner-takes-all which usually constitutes Actor models (figure 2). It is a recent model proposed by Gurney, Prescott and Redgrave - the so called GPR model -, taking into account known anatomy and physiology of the basal ganglia [5]. It has been tested successfully in the frame of action selection and has shown to be able to generate coherent and relevant behavioural sequences when implemented in different robots, performing different behavioural tasks, despite its lack of learning ability [4; 8]. Fully described in [5], we will only introduce here its main properties.

Like other Actors, the GPR is constituted of a series of segregated channels, each one representing an action. The specificity of the model is that all these channels are being processed by two different pathways through dorsal striatum: the first implements action selection, whereas the second regulates this selection by enhancing the selectivity under inter-channels competition, and controls the global activity, allowing effective selection irrespective of the number of channels in the model. A cortex-basal ganglia-thalamus-cortex loop allows the model to take into account each channel's persistence (figure 2). The input values of the model are saliences - commitments toward displaying a given action - that are computed out of the twelve sensorial variables, a constant and a persistence factor - equal to 1 if the action was selected at previous timestep.

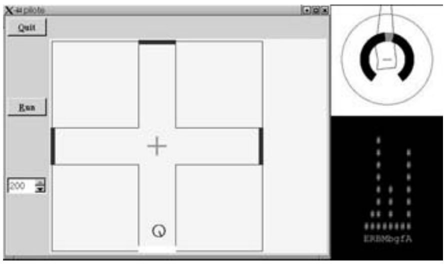

Figure 1: Left: The plus maze task (walls in black; lit box in white; switch-off boxes in dark grey; center of the maze in grey). The robot is represented by a circle with a feature indicating its orientation. Upper right: the robot's perception given by a panoramic camera. Lower right: activation of each channel in the Actor.

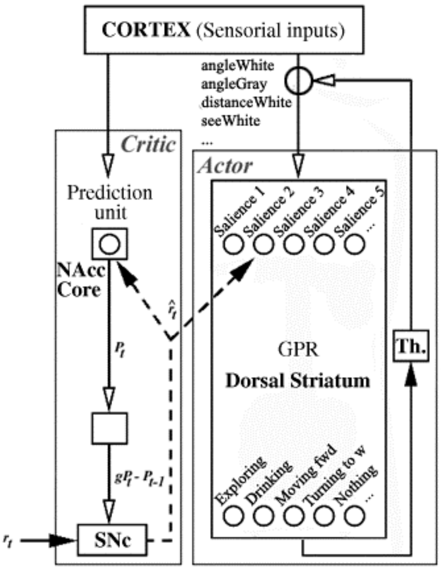

Figure 2: The Actor-Critic model. In dorsal striatum, the GPR model is assumed to process action selection as an Actor (right). Its different layers are constituted of segregated channels (saliences as inputs, actions as outputs). A persistence of last selected action is taken into account in saliences computation via the cortex-basal ganlia-thalamus (Th.)-cortex loop. The Critic (left) computes predictions and signals of reinforcement - symbols are those used in equation 2. It is constituted of one or more prediction units, assumed to be in the nucleus accumbens (Nacc Core), and projecting to the striatum via dopamine neurons in the substantia nigra pars compacta (Snc).

1.3 The Critics

Concerning the Critic part (figure 2), we implemented three different processes of TD-learning derived from three different models [2; 6; 12]. Table 3 shows each Critic model's main characteristics relevant for this work.

	Discrete state space task	Instantaneous rewards	Temporal representation of stimuli	More than 1 prediction unit	Autonomous specialization of prediction units
Houk *et al.*, 1995	●	●			
Suri & Schultz, 2001	●		●	●	
Baldassarre, 2002	●	●		●	●

Table 3: Main characteristics of each of the three Critic models as they were simulated by their authors.

1.3.1 Critic 1 (Houk et al., 1995)

The model proposed by Houk *et al.* was one of the first Actor-Critic models [6]. At each timestep, the Critic computes equation 2 and uses it to update the model's synaptic weights with equation 1 (figure 2). The model does not include any temporal representation of stimuli. It uses a single prediction unit receiving the same input variables than the Actor, except the persistence variable. This unit is supposed to induce learning in the whole environment and no autonomous specialization is required. The prediction computed by the unit at time t is:

$$P^t = \text{f\_sigmo}(w'^1.\text{seeWhite}^t + w'^2.\text{seeGrey}^t + ... + w'^{12}.\text{nearDarkGrey}^t + \text{constant}) \qquad (3)$$

where w'^j is a synaptic weight of the prediction unit, and *f_sigmo* is a sigmoid function to restrict the output domain of the neuron to the interval (0;1) with a discharge threshold set to 0.25. In our simulations, we set to 0.95 the discount factor *g* used in equation 2, and we set to 0.2 the learning rate η used in equation 1.

1.3.2 Critic 2 (Suri & Schultz, 2001)

For the second series of simulations, we have provided the previous model with a solution derived from another Critic [12], which uses a temporal representation of events - such as stimuli and rewards. It allows the Critic to decrease its prediction between the beginning of a non instantaneous reward until its end, even if the sensorial input received by the model does not change during this period of time. This contribution generally provides a timing mechanism to the model, enabling its functioning in the case of non instantaneous rewards. Besides, the model uses several prediction units, but Suri & Schultz do not propose any autonomous specialization mechanism for these units, so that each unit has to be affected *ad hoc* by the experimenter to arbitrary stimuli in the environment.

In this work, as the model has to learn to autonomously discriminate relevant stimuli, we could not do this arbitrary affectation and rather used only one prediction unit. Besides, instead of implementing a special module which has to learn to filter perceived stimuli and to transform them into a temporal representation [12], we have used a single memory cell which cumulates quantities of reward consumed by the robot during a temporal window (10 iterations). The memory cell inhibits the prediction unit of the model, so that the cumulated scalar value c^t is removed to the total prediction computed at time *t*. With this method, equation 3 computed by the prediction unit of the Critic becomes:

$$P^t = \text{f\_sigmo}(w'^1.\text{seeWhite}^t + w'^2.\text{seeGrey}^t + ... + w'^{12}.\text{nearDarkGrey}^t + \text{constant}) - c^t \qquad (4)$$

1.3.3 Critic 3 (Baldassarre, 2002)

In the third series of simulations, we used a model which implements a mixture of experts for the use of several prediction units in the Critic module [2]. In this method, each prediction unit is an expert which can learn autonomously to specialize itself in the prediction of reward in a particular sensorial context. To do so, at each timestep, each expert computes its own prediction of reward, which determines its contribution in the Critic's global prediction - the latter is computed as the sum of each expert's prediction, weighted by a gating network taking into account each expert's prediction error.

The principle of the learning method proposed by Baldassarre's model is: the less an expert makes prediction errors in a region of the maze, the more it will learn so as to specialize in areas of the environment where it does very few errors.

So, the equation used to update the expert's weights depend only on the expert's own error, and is a modification of the Widrow-Hoff rule [15]:

$$w^{ij} \leftarrow w^{ij} + \eta.e^{jt}.h^{jt}.E^{ijt} \qquad (5)$$

where w^{ij} is a synaptic weight of expert *j*, $\eta > 0$ is the learning rate, E^{ijt} is the input value of the synapse w^{ij} at time *t*, e^{jT} is the expert's error signal and h^{jT} is the contribution of expert *j* to the Critic's global prediction error. Detailed equations describing Baldassarre's model can be found in [2].

In Baldassarre's simulations, no temporal representation of stimuli was used because rewards were instantaneous. For our implementation, we added this feature by the use of the same memory cell as described above. Then the Critic's global prediction becomes:

$$P^t = \sum^j [p^{jt}.gate^{jt}] - c^t \qquad (6)$$

where p^{jt} is the prediction of expert *j* and $gate^{jt}$ is a weight associated to expert *j* by the gating network.

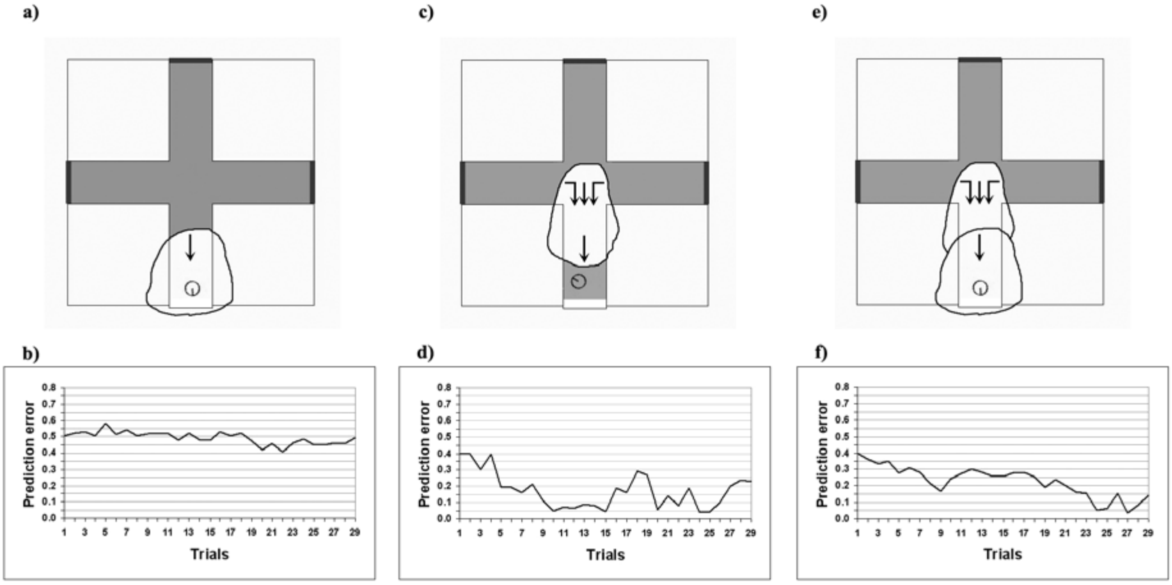

Figure 4: Results of the GPR Actor part connected with Critic 1, 2 and 3 (resp. from left to right). Top: areas were learning has occurred; bottom: prediction error (x-axis: trials; y-axis: prediction error rate).

2 Results

Figure 4 illustrates results corresponding to the three experiments we did, one per model. Each experiment is constituted of 29 trials. A trial begins when a new box is lit, and it ends when the robot has consumed the third drop of water at this box. As a consequence, the lamp of this box is automatically turned off and a new trial begins. Darkened areas in figures 4.a, 4.c and 4.e show, for each experiment, the region of the maze where learning did not occur. Figures 4.b, 4.d and 4.f show, for each trial (x-axis) of each experiment, the cumulated prediction error (y-axis) made by the Critic during reward consumption. This cumulated error is computed over a temporal window of 4 iterations of the model: one for each drop of water plus one iteration after the light has just turned off. According to these figures, we can notice that the three models differ in two main criteria:

- The areas where the robot selects the appropriate actions to solve the task - i.e. the areas where learning has already occurred.
- The way the Critic's prediction errors decrease along the experiment - which is an indication of the efficiency of learning, because the model has to decrease its prediction error at reward location in order to propagate learning to the rest of the maze.

Critic 1: Houk et al's Critic has learned in an area restricted to the neighbourhood of reward location (figure 4.a). It has built a sequence of two actions: when the robot is in front of the lit box, it selects *moving forward*, and when the robot reaches the box, it selects *drinking*. But the behavioural sequence was not extended to the whole maze. This is due to the use of only one prediction neuron, which has a limited computing capacity and cannot extend its domain of correct predictions to a large proportion of the state space. Furthermore, figure 4.b show that prediction errors made by the Critic do not decrease during the experiment. We identified this issue as the fact that the model was first designed to predict instantaneous rewards. In this case, the Critic's prediction was set to zero at the precise time when reward occurred, so that no reward was predicted at the next timestep and no error was thus made by the model. Here, as the reward is constituted of several successive drops of water, if the Critic's prediction is null at first drop, the second drop will not be predicted and the model will thus make a prediction error at next timestep.

Critic 2: The temporal representation of stimuli and rewards used by Suri and Schultz's model is designed to solve this issue [12]. In our experiments, it allowed the Critic to decrease its prediction errors in the neighbourhood of the lit box (figure 4.d). It permitted the model to transfer its learning domain and to start reinforcing the robot's behaviour at the center of the maze. It has started extending the behavioural sequence described for Critic 1 to a three-action sequence: the model has reinforced the action *turning to white stimulus* when the robot arrived at the center of the maze. But as we still used one prediction unit, the Critic's capacity to make correct prediction is limited to a restricted sensorial context input, having as a consequence for the area of correct prediction to dither between reward location and the center of the maze (as illustrated in figure 4.d by the oscillations of the error value after the 16th trial), which sometimes prevents the robot from reaching the reward (figure 4.c).

Critic 3: The mixture of experts used by Baldassarre allows making correct prediction in various sensorial contexts [2]. Figures 4.e and 4.f show the results obtain with the simulation of a two-expert Critic model. Learning has occurred correctly in a large area including the center of the maze and the reward location. The model has built and stabilized a three-action sequence: when the robot reaches the center of the maze, it selects *turning to white* stimulus, when the robot is oriented in the direction of the lit box, it selects *moving forward*, and when the robot is close to the box, it selects *drinking*. Nevertheless, there remains a portion of the maze where learning did not occur (darkened area on figure 4.e).

According to the above-mentioned criteria, Critic 1 is the less appropriate for the achievement of the task, and Critic 3 the most successful. However, the latter’s prediction errors decrease much slower than Critic 2 (figure 4.f), partly because prediction units have to share reinforcement signals. As a consequence, it has a longer learning time to achieve the 29 trials of the experiments: 18h56 for Critic 3, compared to 7h01 for Critic 2 (19h47 for Critic 1). This duration is related to the number of iterations of the TD algorithm (one lasting 800ms) during a trial, as the faster the robot reaches reward location, the less iterations there are in a trial: Critic 3 has an average number of iterations per trial of 2938, compared to 1090 for Critic 2 (3069 for Critic 1). Thus, in our experiments, even if Critic 3 is the only one that could extend learning, its time performance would have to be improved.

3 Discussion and Conclusion

In this work, we tested three models of Critic, connected with the same detailed Actor, in a S-R reward-seeking task. Results demonstrate that learning can occur with all these models, as they never associate non pertinent actions with the task resolution. They also show that both temporal representation of stimuli and coordination of several prediction units that autonomously specialize in a particular context – only present in Critic 3 - are more effective for a Critic to achieve a learning task characterized by a continuous state space and by non instantaneous occurrence of stimuli. However, results also indicate that Critics have still to be improved to extend learning to the whole experimental environment and to prevent a long-lasting learning process when several prediction units are used. In our future work, we will adapt the last Critic model to a mixture of experts method which implements independent experts that can learn to specialize without slowing down the learning process [14].

With such developments, Actor-Critic models could provide a powerful tool for robots that need to adapt autonomously in complex dynamic environments, and with unfamiliar perceptive characteristics. These models constitute the most common hypothesis for the modelling of S-R learning process, which is supposed to be located in the part of the vertebrate's basal ganglia bound to the dorsal striatum. However, another hypothesis, based on recent electrophysiological studies, showing that dopamine has a more complex pattern activity than being a simple reinforcement signal, proposes an alternative model to the TD rule based on a Hebbian learning rule using glutamate reinforcement signals [9]. Another perspective of our work will consist of comparing both hypotheses.

References

[1] Albertin, S. V., Mulder, A. B., Tabuchi, E., Zugaro, M. B. & Wiener, S. I. (2000). Lesions of the Medial Shell of the Nucleus Accumbens Impair Rats in Finding Larger Rewards, but Spare Reward-Seeking Behavior. *Behavioral Brain Research*, 117(1-2):173-83.

[2] Baldassarre, G. (2002). A Modular Neural-Network Model of the Basal Ganglia's Role in Learning and Selecting Motor Behaviours. *Journal of Cognitive Systems Research*, 3(1):5-13.

[3] Barto, A. G. (1995). Adaptive Critics and the Basal Ganglia. In Houk *et al.* (Eds), *Models of Information Processing in the Basal Ganglia* (pp. 215-232). The MIT Press, Cambridge, MA.

[4] Girard, B., Cuzin, V., Guillot, A., Gurney, K. N. & Prescott, T. J. (2003). A Basal Ganglia Inspired Model of Action Selection Evaluated in a Robotic Survival Task. *Journal of Integrative Neuroscience*, 2(2):179-200.

[5] Gurney, K. N., Prescott, T. J. & Redgrave, P. (2001). A Computational Model of Action Selection in the Basal Ganglia. i. A New Functional Anatomy. ii. Analysis and Simulation of Behaviour. *Biological Cybernetics*, 84:401-423.

[6] Houk, J. C., Adams, J. L. & Barto, A. G. (1995). A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement. In Houk *et al.* (Eds), *Models of Information Processing in the Basal Ganglia* (pp. 215-232). The MIT Press, Cambridge, MA.

[7] Joel, D., Niv, Y. & Ruppin, E. (2002). Actor-Critic Models of the Basal Ganglia: New Anatomical and Computational Perspectives. *Neural Networks*, 15:535-547.

[8] Montes-Gonzalez, F. Prescott, T. J., Gurney, K. N., Humphries, M. & Redgrave, P. (2000). An Embodied Model of Action Selection Mechanisms in the Vertebrate Brain. In Meyer *et al.* (Eds), *From Animals to Animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior* (pp.157-166). The MIT Press, Cambridge, MA.

[9] Pennartz, C. M. A, McNaughton, B. L. & Mulder, A. B. (2000). The Glutamate Hypothesis of Reinforcement Learning. *Progress in Brain Research*, 126:231-253.

[10] Schultz, W., Apicella, P. & Ljungberg, T. (1993). Responses of Monkey Dopamine Neurons to Reward and Conditioned Stimuli During Successive Steps of Learning a Lelayed Response Task. *Journal of Neuroscience*, 13(3):900-913.

[11] Schultz, W., Dayan, P. & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, 275:1593-1599.

[12] Suri, R. E. & Schultz, W. (2001). Temporal Difference Model Reproduces Anticipatory Neural Activity. *Neural Computation*, 13:841-862.

[13] Sutton, R. S. & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA.

[14] Tani, J. & Nolfi, S. (1999). Learning to Perceive the World as Articulated: An Approach for Hierarchical Learning in Sensory-Motor Systems. *Neural Networks*, 12(7-8):1131-1141.

[15] Widrow, B. & Hoff, M. E. (1960). Adaptive Switching Circuits. *IRE WESCON Convention Record*, 4:96-104.