Actor-Critic Models of Reinforcement Learning in the Basal Ganglia:

From Natural to Artificial Rats

Preprint : accepted for publication in *Adaptive Behavior 13(2):131-148, Special Issue Towards Aritificial Rodents, 2005.*

Mehdi Khamassi[1,2], Loïc Lachèze[1], Benoît Girard[1,2], Alain Berthoz[2] and Agnès Guillot[1]

[1]AnimatLab, LIP6, 8 rue du capitaine Scott, 75015 Paris, France

[2] LPPA, Collège de France, 11 place Marcellin Berthelot, 75005 Paris, France

Since 1995, numerous Actor-Critic architectures for reinforcement learning have been proposed as models of dopamine-like reinforcement learning mechanisms in the rat's basal ganglia. However, these models were usually tested in different tasks, and it is then difficult to compare their efficiency for an autonomous animat. We present here the comparison of four architectures in an animat as it performs the same reward-seeking task. This will illustrate the consequences of different hypotheses about the management of different Actor sub-modules and Critic units, and their more or less autonomously determined coordination. We show that the classical method of coordination of modules by mixture of experts, depending on each module's performance, did not allow solving the task. Then we address the question of which principle should be applied to efficiently combine these units. Improvements for Critic modeling and accuracy of Actor-critic models for a natural task are finally discussed in the perspective of our Psikharpax project – an artificial rat having to survive autonomously in unpredictable environments.

Keywords animat approach - TD learning - Actor-Critic model - S-R task - taxon navigation

1. Introduction

This work aims at adding learning capabilities in the architecture of action selection introduced by Girard *et al.* in this issue. This architecture will be implemented in the artificial rat Psikharpax, a robot that will exhibit at least some of the capacities of autonomy and adaptation that characterize its natural counterpart (Filliat *et al.*, 2004). This learning process capitalizes on Actor-Critic architectures, which have been proposed as models of dopamine-like reinforcement learning mechanisms in the rat's basal ganglia (Houk *et al.*, 1995). In such models, an Actor network learns to select actions in order to maximize the weighted sum of future rewards, as computed on line by another network, a Critic. The Critic predicts this sum by comparing its estimation of the reward with the actual one by means of a Temporal Difference (TD) learning rule, in which the error between two successive predictions is used to update the synaptic weights (Sutton and Barto, 1998). A recent review of numerous computational models, built on this principle since 1995, highlighted several issues raised by the inconsistency of the detailed implementation of Actor and Critic modules with known basal ganglia anatomy and physiology (Joel *et al.*, 2002). In the first section of this paper, we will consider some of the main issues, updated with anatomical and neurophysiological knowledge. In the second section, we will illustrate the consequences of alternative hypotheses concerning the various Actor-Critic designs by comparing animats that perform the same classical instrumental learning (S-R task). During the test, the animat freely moves in a plus-maze with a reward placed at the end of one arm. The reward site is chosen randomly at the beginning of each

trial and it refers to site-specific local stimuli. The animat has to autonomously learn to associate continuous sensory information with certain values of reward and to select sequences of behaviors that enable it to reach the goal from any place in the maze. This experiment is more realistic than others used to validate Actor-Critic models, often characterized by an a priori fixed temporal interval between a stimulus and a reward (e.g., Suri and Schultz, 1998), by an unchanged reward location over trials (e.g., Strösslin, 2004), or by a discrete state space (e.g., Baldassarre, 2002).

We will compare, in this task, four different principles inspired by Actor-Critic models trying to tackle the issues evocated in the first section. The first one is the seminal model proposed by Houk *et al.* (1995), which uses one Actor and a single prediction unit (*Model AC* – one Actor, one Critic), which is supposed to induce learning in the whole environment. The second principle implements one Actor with several Critics *(Model AMC1* – one Actor, Multiple Critics). The Critics are combined by a mixture of experts where a gating network is used to decide which expert – which Critic – is used in each region of the environment, depending on its performance in that region. The principle of mixture of experts is inspired from several existing models (Jacobs *et al.*, 1991; Baldassarre, 2002; Doya *et al.*, 2002). The third one is inspired by Suri and Schultz (2001) and uses also one Actor with several Critic experts. However, the decision of which expert should work in each sub-zone of the environment is independent from the experts' performances, but rather depends on a partition of the sensory space perceived by the animat (*Model AMC2* – one Actor, Multiple Critics). The fourth one (*Model MAMC2* – Multiple Actors, Multiple Critics) proposes the same principle as the previous Critic, combined with several Actors, which latter principle is one of the features of Doya *et al.*'s model (2002), particularly designed for continuous tasks, and is also a feature of Baldassarre's model (2002). Here we will implement these principles in four models using the same design for each Actor component. Their comparison will be made on the learning speed and on their ability to extend learning to the whole experimental environment.

The last section of the paper will discuss the results on the basis of acquired knowledge in reinforcement learning tasks in artificial and natural rodents.

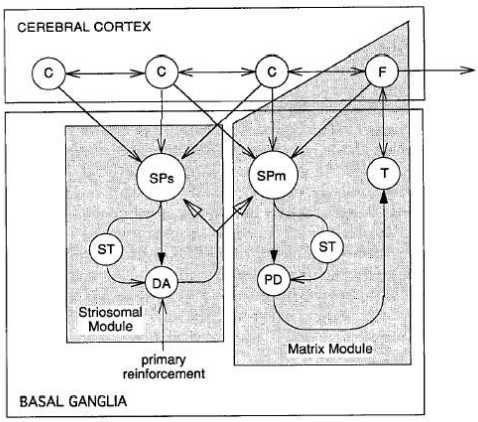

Figure 1 Schematic illustration of the correspondence between the modular organization of the basal ganglia including both striosomes and matrix modules and the Actor-Critic architecture in the model proposed by Houk et al, (1995). F, columns in the frontal cortex; C, other cortical columns; SPs, spiny neurons striosomal compartments of the striatum; SPm, spiny neurons in matrix modules; ST, subthalamic sideloop; DA, dopamine neurons in the substantia nigra pars compacta; PD, pallidal neurons; T, thalamic neurons. (adapted from Houk et al., 1995).

2. **Actor-Critic designs: the issues**

The two main principles of Actor-Critic models that lead to consider them as a good representation of the role of the basal ganglia in reinforcement learning of motor behaviors are (i): the implementation of a Temporal Difference (TD) learning rule which leads to translate progressively reinforcement signals from the time of reward occurrence to environmental contexts that precede the reward.; (ii): the separation of the model in two distinct parts, one for the selection of motor behaviors (actions) depending on the current sensory inputs (the Actor), and the other for the driving of the learning process via dopamine signals (the Critic).

Schultz's work on the electrophysiology of dopamine neurons in monkeys showed that dopamine patterns of release are similar to the TD learning rule (see Schultz, 1998 for a review). Besides, the basal ganglia are a major input to dopamine neurons, and are also a privileged target of reinforcement signals sent by these neurons (Gerfen *et al.*, 1987). Moreover, the basal ganglia appears to be constituted of two distinct sub-systems, related to two different parts of the striatum – the major input nucleus of the basal ganglia –, one projecting to motor areas in the thalamus, the other projecting to dopamine neurons, influencing the firing patterns of these neurons at least to some extent (Joel and Weiner, 2000).

These properties lead the first Actor-Critic model of the basal ganglia to propose the matrisomes of the striatum to constitute the Actor, and the striosomes of this very structure to be the Critic (Houk *et al.*, 1995, figure 1). The classical segregation of 'direct' and 'indirect' pathways from the striatum to the dopaminergic system (SNc, substantia nigra pars compacta, and VTA, ventral tegmental area; Albin *et al.*, 1989) was used in the model to explain the timing characteristics of dopamine neurons' discharges.

Numerous models were proposed to improve and complete the model of Houk *et al*. However, most of these computational models have neurobiological inconsistencies and lacks concerning recent anatomical hypotheses on the basal ganglia (Joel *et al.*, 2002).

An important drawback is that the Actor part of these models is often simplistic compared to the known anatomy of the basal ganglia and does not take into account important anatomical and physiological characteristics of the striatum. For example, recent works showed a distinction between neurons in the striatum having different dopamine receptors (D1-receptors or D2-receptors; Aizman *et al.*, 2000). This implies at least two different pathways in the Actor, on which tonic dopamine has opposite effects, going beyond the classical functional segregation of 'direct' and 'indirect' pathways in the striatum (Gurney *et al.*, 2001).

Likewise, some constraints deriving from striatal anatomy restrict the possible architectures for the Critic network. In particular, the striatum is constituted of only one layer of medium spiny neurons – completed with 5% of interneurons (Houk *et al.*, 1995). As a consequence, Critic models cannot be constituted of complex multilayer networks for reward prediction computation. This anatomical constraint lead several authors to model the Critic as a single-neuron (Houk *et al.*, 1995; Montague *et al.*, 1996), which works well in relatively simple tasks. For more complicated tasks, several models assign one single Critic neuron to each subpart of the task. These models differ in the computational mechanism used to coordinate these neurons. Baldassarre (2002) and Doya et al. (2002) propose to coordinate Critic modules with a mixture of experts method: the module that has the best performance at a certain time during the task becomes expert in the learning process of this subpart of the task. Another model proposes an affectation of experts to subparts of the task (such as

stimuli or events) in an a priori manner, independently from each expert's performance (Suri and Schultz, 2001). It remains to assess the efficiency of each principle, as they have been at work in heterogeneous tasks (*e.g.* Wisconsin Card Sorting Test, Discrete Navigation Task, Instrumental Conditioning).

These models also question the functional segregation of the basal ganglia in 'direct' and 'indirect' pathways (see Joel *et al.*, 2002 for a review). These objections are built on electrophysiological data (for review see Bunney *et al.*, 1991) and anatomical data (Joel and Weiner, 2000) which show that these two pathways are unable to produce the temporal dynamics necessary to explain dopamine neurons patterns of discharge. These findings lead to question the localization of the Critic in the striosomes of the dorsal striatum, and several models capitalized on its implementation in the ventral striatum (Brown *et al.*, 1999; Daw, 2003). These works are supported by recent fMRI data in humans, showing a functional dissociation between dorsal striatum as the Actor and ventral striatum as the Critic (O'Doherty *et al*., 2004), but they may be controversial for the rat, as electrophysiological data (Thierry *et al.*, 2000) showed that an important part of the ventral striatum (the nucleus accumbens core) does not project extensively to the dopamine system in the rat brain.

We can conclude that the precise implementation of the Critic remains an open question, if one takes also into account a recent model assuming that a new functional distinction of striosomes in the dorsal striatum – based on differential projections to GABA-A and GABA-B receptors in dopamine neurons – can explain the temporal dynamics expected (Frank *et al.*, 2001).

Besides these neurobiological inconsistencies, some computational requirements on which numerous Actor-Critic models have focused seem unnecessary for a natural reward-seeking task. For example, as Houk *et al.*'s model could not account for temporal characteristics of dopamine neurons firing patterns, most of the alternative models focused on the simulation of the depression of dopamine at the precise time where the reward is expected when it eventually does not occur. To this purpose, they concentrated on the implementation of a temporal component for stimulus description – which is computed outside of the model and is sent as an input to the model via cortical projections (Montague *et al.*, 1996; Schultz *et al.*, 1997). These models were tested in the same tasks chosen by Schultz *et al.* (1993) to record dopamine neurons in the monkey, using a fixed temporal bin between a stimulus and a reward. However, in natural situations where a rodent needs to find food or any other type of reward, temporal characteristics of the task are rarely fixed but rather depend on the animal's behavior and on the environment's changes/evolution.

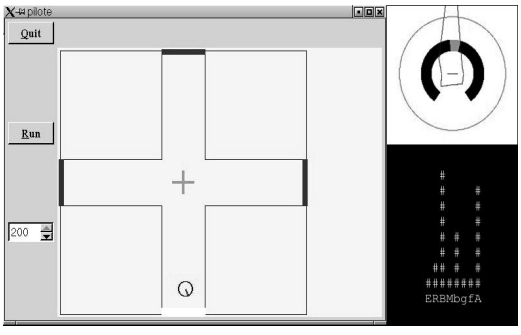

Figure 2 Left: the robot in the plus maze environment. A white arm extremity indicates the reward location. Other arm extremities do not deliver any reward and are shown in black. Upper right: the robot's visual perceptions. Lower right: activation level of different channels in the model.

3. **Method**

The objective of this work is to evaluate the efficiency of the main principles on which current Actor-Critic models inspired by the basal ganglia are designed, when they are implemented in the same autonomous artificial system. The main addressed issues are:

- The implementation of a detailed Actor, whose structure would be closer to the anatomy of the dorsal striatum, assessing whether reinforcement learning is still possible within this architecture.
- The comparison of the function of one Critic unit, versus several alternative ways to coordinate different Critic modules for solving a complex task where a single-neuron is not enough.
- The test of the models in a natural task involving taxon navigation where events are not predetermined by fixed temporal bins. Instead, the animat perceives a continuous sensory flow during its movements, and has to reactively switch its actions so as to reach a reward.

3.1. The simulated environment and task

Figure 2 shows the experimental setup simulated, consisting in a simple 2D plus-maze. The dimensions are equivalent to a 5m * 5m environment with 1m large corridors. In this environment, walls are made of segments colored on a 256 grayscale. The effects of lighting conditions are not simulated. Every wall of the maze is colored in black (luminance = 0), except walls at the end of each arm and at the center of the maze, which are represented by specific colors: the cross at the center is gray (191), three of the arm extremities' walls are dark gray (127) and the fourth is white (255), indicating the reward location (equivalent to a water trough delivering two drops – non instantaneous reward – not a priori known by the animat).
The plus-maze task mimicks the neurobiological and behavioral studies that will serve as future validation for the model (Albertin *et al*., 2000). In this task, at the beginning of each trial, one arm extremity is randomly chosen to deliver reward. The associated wall is colored in white whereas walls at the three other extremities are dark gray. The animat has to learn that selecting the action *drinking* when it is near the white wall (distance < 30 cm) and faces it (angle < 45 degrees) gives it a reward. Here we assume that reward = 1 for n iterations (n = 2), without considering how the hedonic value of this reward is determined.
We expect the animat to learn a sequence of context-specific behaviors, so that it can reach the reward site from any starting point in the maze:

- When not seeing the white wall, face the center of the maze and move forward.
- As soon as arriving at the center (the animat can see the white wall), turn to the white stimulus.
- Move forward until being close enough to reward location.
- Drink.

The trial ends when reward is consumed: the color of the wall at reward location is changed to dark gray, and a new arm extremity is chosen randomly to deliver reward. The animat has then to perform again the learned behavioral sequence. Note that there is no break between two consecutive trials: trials follow each other successively.

The more efficiently and fluently the animat performs the above described behavioral sequence, the less time it will take to reach the reward. As a consequence, the criterion chosen to validate the models is the time to goal, plotted along the experiment as the learning curve of the model.

3.2. The animat

The animat is represented by a circle (30 cm diameter). Its translation and rotation speeds are 40 $cm.s^{-1}$ and $10°.s^{-1}$. Its simulated sensors are:

- An omnidirectional linear camera providing every 10° the color of the nearest perceived segment. This results in a 36 colors table that constitute the animat's visual perception (see figure 2),
- Eight sonars with a 5m range, an incertitude of ±5 degrees concerning the pointed direction and an additional ±10 cm measurement error,

The sonars are used by a low level obstacle avoidance reflex which overrides any decision taken by the Actor-Critic model when the animat comes too close to obstacles.

The animat is provided with a visual system that computes 12 input variables $(\forall i \in [1;12], 0 < \mathrm{var}_i < 1)$ out of the 36 colors table at each time step. These sensory variables constitute the state space of the Actor-Critic and so will be taken as input to both the Actor and the Critic parts of the model (figure 3). Variables are computed as following:

- *seeWhite*(resp. *seeGray*, *seeDarkGray*) = 1 if the color table contains the value 255 (resp. 191, 127), else 0.
- *angleWhite*, *angleGray*, *angleDarkGray* = (number of boxes in the color table between the animat's head direction and the desired color) / 18.
- *distanceWhite*,*distanceGray*, *distanceDarkGray* = (maximum number of consecutive boxes in the color table containing the desired color) / 18.
- *nearWhite* (resp. *nearGray*, *nearDarkGray*) = 1 – *distanceWhite* (resp. *distanceGray*, *distanceDarkGray*).

Representing the environment with such continuous variables will imply for the model to permanently receive a flow of sensory information and having to learn autonomously the events (sensory contexts) that can be relevant for the task resolution.

The animat has a repertoire of 6 actions: *drinking*, *moving forward*, *turning to white perception*, *turning to gray perception*, *turning to dark gray perception*, and *waiting*. These actions constitute the output of the Actor model (described below) and the input to a low-level model that translates it into appropriate orders to the animat's engines.

3.3. The model: description of the Actor part

The Actor-Critic model is inspired by the rat basal ganglia. As mentioned in section 2, the Actor can be hypothesized as implemented in the matrix part of the basal ganglia, while striosomes in the dorsal striatum are considered as the anatomical counterpart for the Critic. The Critic produces dopamine-like reinforcement

signals that help it learn to predict reward during the task, and that make the Actor learn to select appropriate behaviors in every sensory context experienced during the task.

The architecture implemented in the Actor is a recent model proposed by Gurney, Prescott and Redgrave (2001a,b) – henceforth called GPR model - that replaces the simple winner-takes-all which usually constitutes Actor models and is supposed to be more biologically plausible.

Like other Actors, the GPR is constituted of a series of parallel channels, each one representing an action (in our implementation, we used 6 channels corresponding to the 6 actions used for the task). This architecture constitutes an alternative view to the prevailing functional segregation of the basal ganglia into 'direct' and 'indirect' pathways discussed in section 1 (Gurney *et al.*, 2001). All these channels are composed by two different circuits through dorsal striatum: the first is the 'selection' pathway, implementing action selection properly via a feed-forward off-center on-surround network, and mediated by cells in the dorsal striatum with D1-type receptors. The second is the 'control' pathway, mediated by cells with D2-type receptors in the same area. Its role is to regulate the selection by enhancing the selectivity inter-channels, and to control the global activity within the Actor. Moreover, a cortex-basal ganglia-thalamus loop in the model allows it to take into account each channel's persistence in the process of selection (see Gurney *et al.*, 2001, for detailed description and mathematical implementation of the model). The latter characteristic showed some interesting properties that prevented a robot from performing behavioral oscillations (Montes-Gonzalez *et al.*, 2000; Girard *et al.*, 2003).

In our implementation, the input values of the Actor model are saliences – i.e. the strength of a given action – that are computed out of the 12 sensory variables, a constant implementing a bias, and a persistence factor – equal to 1 for the action that was selected at previous timestep (figure 3). At each timestep t (timesteps being separated by a 1 sec bin in our simulations), the action that has the highest salience is selected to be performed by the animat, the salience of action i being:

$$sal_i(t) = \left[\sum_{j-1}^{13} \mathrm{var}_j(t) \cdot w_{i,j}(t) \right] + persist_i(t) \cdot w_{i,14}(t) \qquad (1)$$

where $\mathrm{var}_{13}(t) = 1, \ \forall t$, and the $w_{i,j}(t)$ are the synaptic weights representing, for each action *i*, the association strength with input variable *j*. These weights are initiated randomly $(\forall i,j, -0.02 < w_{i,j}(t=0) < 0.02)$ and the objective of the learning process will be to find a set of weights allowing the animat to perform the task efficiently.

An exploration function is added that would allow the animat to try an action in a given context even if the weights of the Actor do not give a sufficient tendency to perform this action in the considered context.

To do so, we introduce a clock that triggers exploration in two different cases:

- When the animat has been stuck for a large number of timesteps (*time* superior to a fixed threshold α) in a situation that is evaluated negative by the model (when the prediction *P*(t) of reward computed by the Critic is inferior to a fixed threshold).
- When the animat has remained for a long time in a situation where *P*(t) is high but this prediction doesn't increase that much ($|P(t+n) - P(t)| < \epsilon$) and no reward occurs.

If one of these two conditions is true, exploration is triggered: one of the 6 actions is chosen randomly. Its salience is being set to 1 (Note that: when exploration = false, $sal_i(t) < 1, \ \forall i,t, w_{i,j}(t)$) and is being

maintained to 1 for a duration of 15 timesteps (time necessary for the animat to make a 180° turn or to run from the center of the maze until the end of one arm).

3.4. The model: description of the Critic part

For the Critic part of the model, different principles based on existing techniques are tested. The idea is to test the hypothesis of one single Critic unit first, but also to provide the Critic with enough computational capacities so that it can correctly estimate the value function over the whole environment of the task. In other words, the Critic will have to deal with several different sensory contexts – corridors, maze center, extremity of arms, etc. equivalent to different stimuli –, and will have to associate a correct reward prediction to these contexts.

One obvious possibility would be a multilayer perceptron with several hidden layers but, as mentioned before in section 2, there are anatomical constraints which prevent us from adopting this choice: our Critic is supposed to be situated in the striosomes of dorsal striatum, which structure is constituted of only one layer of medium spiny neurons (Houk *et al.*, 1995). Thus we need a more general method that combines several Critic modules, each one being constituted of a single neuron and dealing with a particular part of the problem space.

The method adopted here is the mixture of experts, which was proposed to divide a non-linearly separable problem into a set of linearly separable problems, and to affect a different expert to each considered sub-problem (Jacobs *et al.*, 1991).

The Critics tested in this work differ mainly in two following manners:

- The first (*Model AMC1*) implements a mixture of experts in which a gating network is used to decide which expert is used in each region.
- The second (*Model AMC2*) implements a mixture of experts in which a hand-determined partition of the environment based on a categorization of visual perceptions is used to decide which expert works in each sub-zone.

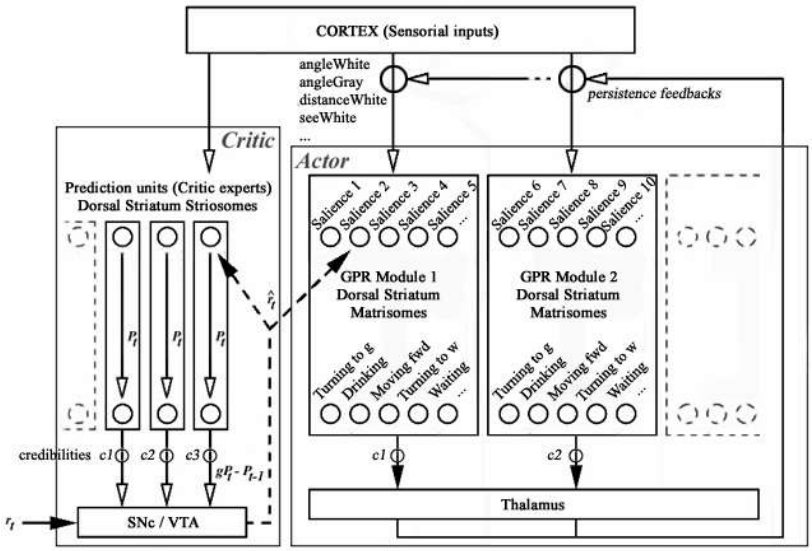

Figure 3 General scheme of the models tested in this work. The Actor is a group of GPR modules with saliences as inputs and actions as outputs. The Critic (involving striosomes in the dorsal striatum, and the substantia nigra compacta (SNc))

propagates towards the Actor an estimate $\hat{r}$ of the instantaneous reinforcement triggered by the selected action. The particularity of this scheme is to combine several modules for both Actor and Critic, and to weight the Critic experts' predictions and the Actor modules' decisions with credibilities. These credibilities can be either computed by a gating network (*Model AMC1*) or in a context-dependent manner (*Models AMC2 and MAMC2*).

Moreover, since the animat has to solve a task in continuous state space, there could be interferences between reinforcement signals sent by different Critic experts to the same single Actor. In this way, whereas one model will employ only one Actor (*Model AMC2*), another one will use one Actor module associated to each expert (*Model MAMC2*). Figure 3 shows the general scheme with different modules employed as suggested by the models presented here.
Performances of *Models AMC1*, *AMC2* and *MAMC2* will be compared, together with the one of the seminal Actor-Critic model inspired by the basal ganglia, proposed by Houk, Adams and Barto (1995), and using a single cell Critic with a single Actor (*Model AC*).
We will start by the description of the simplest Critic, the one belonging to *Model AC*.

3.4.1. Model AC

In this model, at each timestep, the Critic is a single linear cell that computes a prediction of reward based on the same input variables than the Actor, except the persistence variable:

$$P(t) = \sum_{j=1}^{13} \mathrm{var}_j(t) \cdot w'_j(t) \qquad (2)$$

where $w'_j(t)$ are the synaptic weights of the Critic.
This prediction is then used to calculate the reinforcement signal by means of the TD-rule:

$$\hat{r}(t) = r(t) + gP(t) - P(t-1) \qquad (3)$$

where $r(t)$ is the actual reward received by the animat, and g is the discount factor ($0 < g < 1$) which determines how far in the future expected rewards are taken into account in the sum of future rewards.
Finally, this reinforcement signal is used to update both Actor's and Critic's synaptic weights according to the following equations respectively:

$$w_{i,j}(t) \leftarrow w_{i,j}(t-1) + \eta \cdot \hat{r}(t) \cdot \mathrm{var}_j(t-1) \qquad (4)$$

$$w'_j(t) \leftarrow w'_j(t-1) + \eta \cdot \hat{r}(t) \cdot \mathrm{var}_j(t-1) \qquad (5)$$

where $\eta > 0$ is the learning rate.

3.4.2. Model AMC1

As this Critic implements N experts, each expert *k* computes its own prediction of reward at timestep *t*:

$$p_k(t) = \sum_{j=1}^{13} w'_{k,j}(t) \cdot \mathrm{var}_j(t) \qquad (6)$$

where the $w'_{k,j}(t)$ are the synaptic weights of expert *k*.

Then the global prediction of the Critic is a weighted sum of experts' predictions:

$$P(t) = \sum_{k=1}^{N} cred_k(t) \cdot p_k(t) \qquad (7)$$

where $cred_k(t)$ is the credibility of expert *k* at timestep *t*. These credibilities are computed by a gating network which learns to associate, in each sensory context, the best credibility to the expert that makes the smaller prediction error. Following Baldassarre's description (2002), the gating network is constituted of N linear cells which receive the same input variables than the experts and compute an output function out of it:

$$o_k(t) = \sum_{j=1}^{13} w''_{k,j}(t) \cdot \mathrm{var}_j(t) \qquad (8)$$

where $w''_{k,j}(t)$ are the synaptic weights of gating cell *k*.

The credibility of expert *k* is then computed as the softmax activation function of the outputs $o_f(t)$:

$$cred_k(t) = \frac{o_k(t)}{\sum_{f=1}^{N} o_f(t)} \qquad (9)$$

Concerning learning rules, whereas equation (3) is used to determine the global reinforcement signal sent to the Actor, each Critic's expert has a specific reinforcement signal based on its own prediction error:

$$\hat{r}_k(t) = r(t) + gP(t) - p_k(t-1) \qquad (10)$$

The synaptic weights of each expert *k* are updated according to the following formula:

$$w''_{k,j}(t) \leftarrow w''_{k,j}(t-1) + \eta \cdot \hat{r}_k(t) \cdot \mathrm{var}_j(t-1) \cdot h_k(t) \qquad (11)$$

where $h_k(t)$ is the contribution of expert *k* to the global prediction error of the Critic, and is defined as:

$$h_k(t) = \frac{cred_k(t-1) \cdot corr_k(t)}{\sum_{f=1}^{N} cred_f(t-1) \cdot corr_f(t)} \qquad (12)$$

where $corr_k(t)$ is a measure of the « correctness » of the expert *k* defined as:

$$corr_k(t) = \exp\left(\frac{-\hat{r}_k(t)^2}{2\sigma^2}\right) \quad (13)$$

where σ is a scaling parameter depending on the average error of the experts (see parameters table in the appendix section).
Finally, to update the weights of the gating network, we use the following equation:

$$w''_{k,j}(t) \leftarrow w''_{k,j}(t-1) + m \cdot diff(t) \cdot \mathrm{var}_j(t-1) \quad (14)$$

with $diff(t) = h_k(t) - cred_k(t-1)$ (15)

where *m* is a learning rate specific to the gating network.
So the credibility of expert *k* in a given sensory context depends on its performance in this context.

3.4.3. Model AMC2

The Critic also implements N experts. However, it differs from *Model AMC1* in the way the credibility of each expert is computed.
The principle we wanted to bring about here is to dissociate credibilities of experts from their performance. Instead, experts would be assigned to different subregions of the environment – these regions being computed as windows in the perceptual space –, would remain enchained to their associate region forever, and would progressively learn to accurate their performance along the experiment. This principle is declined from Houk et al. (1995) for the improvement of their model, assuming that different striosomes may be specialized in dealing with different behavioral tasks. This proposition was implemented by Suri and Schultz (2001) in using several TD models, each one computing predictions for only one event (stimulus or reward) that occurs in the simulated paradigm.
To test this principle, we replaced the gating network by a hand-determined partition of the environment (e.g. a coarse representation of the sensory space): At timestep *t*, the current zone β depends on the 12 sensory variables computed by the visual system. *Example: if (seeWhite = 1 and angleWhite < 0.2 and distanceWhite > 0.8) then zone = 4 (e.g.* β*=4).* Then $cred_\beta(t) = 1$, $cred_k(t) = 0$ for all other experts, and expert β has then to compute a prediction of reward out of the 12 continuous sensory variables. Predictions and reinforcement signals of the experts are determined by the same equations than Critic of *Model AMC1*.
This was done as a first step in the test of the considered principle. Indeed, we assume that another brain region such as the parietal cortex or the hippocampus would determine the zone (sensory configuration) depending on the current sensory perception (McNaughton, 1989; Burgess *et al.*, 1999), and would send it to the Actor-Critic model of the basal ganglia. Here, the environment was partitioned into N=30 zones, an expert being associated to each zone. The main difference between this scheme and the one used by Suri and Schultz

is that, in their work, training of experts in each sub-zone was done in separated sessions, and the global model was tested on the whole task only after training of all experts. Here, experts will be trained simultaneously in a single experiment.
Finally, one should note that this method is different from applying a coarse coding of the state space that constitutes the input to the Actor and the Critic (Arleo and Gerstner, 2000). Here, we implemented a *coarse coding of the credibility space* so as to determine which expert is the most credible in a given sensory configuration, and kept the 12 continuous sensory variables, plus a constant described above, as the state space for the reinforcement learning process. This means that within a given zone, the concerned expert has to learn to approximate a continuous reward value function, based on the varying input variables.

3.4.4. Model MAMC2

The Critic of this Model is the same as in *Model AMC2* and only differs from its associated Actor.
Instead of using one single Actor, we implemented N different Actor modules. Each Actor module has the same structure than the simple Actor described in section 3.4 and is constituted of 6 channels representing the 6 possible actions for the task. The difference resides in the fact that only actions of the Actor associated with the zone in which the animat is currently are competing to determine the animat's current action.
As a consequence, if the animat was in zone β at time *t* and performed action *i*, the reinforcement signal $\hat{r}(t+1)$ computed by the Critic at next timestep will be used to update only weights of action *i* from the Actor β according to the following equation:

$$w_{k,i,j}(t) \leftarrow w_{k,i,j}(t-1) + \eta \cdot \hat{r}(t) \cdot \mathrm{var}_j(t-1) \qquad (16)$$

Other equations are the same than those used for Critic of *Model AMC2*. As mentioned above, this principle – using a specific controller or a specific Actor for each module of the Actor-Critic model – is inspired by the work of Doya *et al.*, (2002).

3.5. Results

In order to compare the learning curves of the four simulated models, and so as to evaluate which models manage to solve the task efficiently, we adopt the following criterion: after 50 trials of training (out of 100 for each experiments), the animat has to achieve an equivalent performance to a hand-crafted model that can already solve the task (Table 1). To do so, we simulated the GPR action selection model with appropriate hand-determined synaptic weights and without any learning process, so that the animat can solve the task as if it had already learned it. With this model, the animat performed a 50 trials experiment with an average performance of 142 iterations per trial. Since each iteration lasted approximately 1 sec, as mentioned above, it took a little bit more than 2 min per trials to this hand-craft animat to reach the reward.

Table 1. Performances of each model.

Model	GPR	AC	AMC1	AMC2	MAMC
Performance	142	587	623	3240	97

Table 1 shows the performance of each model, measured as the average number of iterations per trial after trial #50. Figure 4 illustrates results to the four experiments performed in the 2D environment, one per model. The x-axis represents the successive trials along the experiments. For each trial, y-axis shows the number of iterations needed for the animat to reach the reward and consume it. Figure 4.a shows the learning curve of *Model AC*. We can first notice that the model increased rapidly its performance until trial 7, and stabilized it at trial 25. However, after trial 50, the average duration of a trial is still 587 iterations, which is nearly 4 times higher than the chosen criterion. We can explain this limitation by the fact that *Model AC* is constituted of only one single neuron in the Critic, which can only solve linearly separable problems. As a consequence, the model could learn only a part of the task – in the area near the reward location –, but it was unable to extend learning to the rest of the maze. So the animat has learned to select appropriate behaviors in the reward area, but it still performs random behaviors in the rest of the environment.

Model AMC1 is designed to mitigate the computational limitations of *Model AC*, as it implies several Critic units controlled by a gating network. Figure 4.b shows its learning curve after simulation in the plus-maze task. The model has also managed to decrease its running time per trial at the beginning of the experiment. However, we can notice that the learning process is more unstable than the previous one. Furthermore, after the 50th trial, the model has a performance of 623 iterations, which is not better than *Model AC*. Indeed, the model couldn't extend learning to the whole maze either. We can explain this failure by the fact that the gating network did not manage to specialize different experts in different subparts of the task. As an example, figure 5 shows the reward prediction computed by each Critic's expert during the last trial of the experiment. It can be noticed that the first expert (dark curve) has the highest prediction throughout the whole trial. This is due to the fact that it is the only one the gating network has learned to consider as credible – its credibility remains above 90% during the whole experiment. As a consequence, only one expert is involved in the learning process and the model becomes computationally equivalent to *Model AC*: it cannot extend learning to the whole maze, which is confirmed by the absence of any reward prediction before the perception of the reward site (stimulus occurrence) in Figure 5.

Figure 4.c shows the learning curve of *Model AMC2* which implements another principle for experts coordination. This model cannot suffer from the same limitations than *Model AMC1*, since each expert was a priori assigned to a specific area of the environment. As a consequence, it quickly managed to extend learning to the whole maze. However, the consequence of this process is to produce interferences in the Actor's computations: the same Actor receives all experts' teaching signals, and it remains unable to switch properly between reinforced behaviors. For example, when the action '*drinking*' is reinforced, the Actor starts selecting this action permanently, even when the animat is far from reward location. These interferences explain the very bad performances obtained with *Model AMC2*.

The last simulated model (*Model MAMC2*) performed best. Its learning curve is shown on figure 4.d. This model implements several Actor modules (an Actor module connected to each Critic expert). As a consequence, it avoids interferences in the learning process and rapidly converged to a performance of 97 iterations per trial. This good performance cannot been reached with the multi-Actor only, since we tried to combined several Actor modules to model *AMC1* and got a performance of 576 iterations per trial. So the achievement of the task implies the combination of a multi-Actor and a good specialization of experts.

For checking the ability of *Model MAMC2* to learn the same task in more realistic conditions, we simulated it a 3D environment, working in real time and implementing physical dynamics (Figure 7). This experiment constituted an intermediary step favoring the implementation into an actual Pekee robot (Wany Robotics). The animat is still able to learn the task in this environment and gets good performances after 35 trials (Figure 6; corresponding average performance of the animat between trials 35 and 65: 284 iterations per trial).

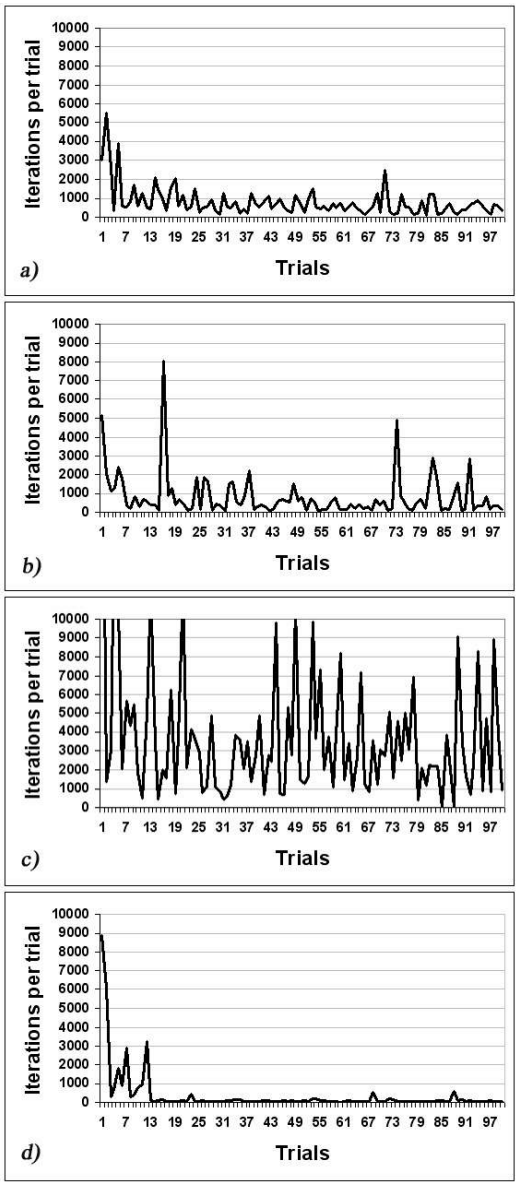

Figure 4 Learning curves of the four models simulated in the 2D plus-maze task over 100 trials experiments. X-axis: trials. Y-axis: number of iterations per trial (truncated to 10000 it. for better readability). **a)** *Model AC*. **b)** *Model AMC1*. **c)** *Model AMC2*. **d)** *Model MAMC2*.

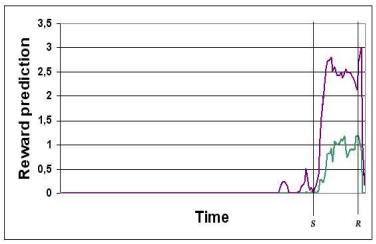

Figure 5 Reward prediction computed by each Critic's expert of Model AMC1 during trial #100 of the experiment. Time 0 indicates the beginning of the trial. S: perception of the stimulus (the white wall) by the animat. R: beginning of reward delivery. The dark curve represents the prediction of expert 1. The other experts' predictions are melted into the light curve or equal to 0.

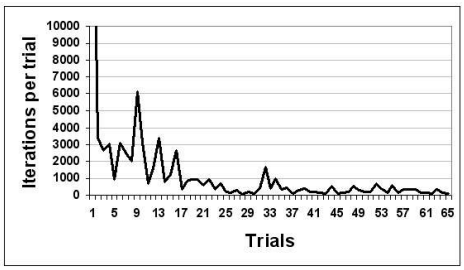

Figure 6 Learning curve in the 3D environment. X-axis: trials. Y-axis: number of iterations per trial.

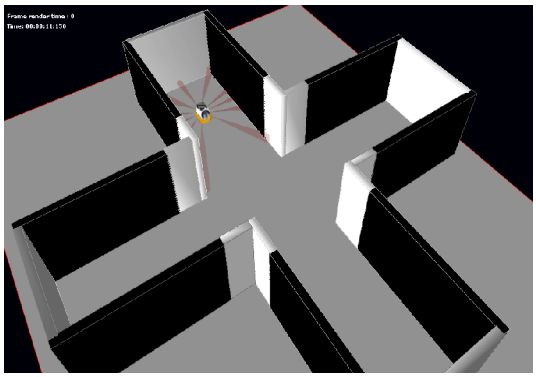

Figure 7 Simulation of the plus-maze task in a 3D environment. Like the 2D environment, one random arm extremity is white and delivers reward. The animat has to perform taxon navigation so as to find and consume this reward. Gray stripes arising from the animat's body represent its sonar sensors used by its low level obstacle avoidance reflex.

4.Discussion and future work

In this work, we compared learning capabilities on a S-R task of several Actor-Critic models of the basal ganglia based on distinct principles. Results of simulations with *models AC, AMC1, AMC2* and *MAMC2* demonstrated that:

- A single-component Critic cannot solve the task (*Model AC*);
- Several Critic modules controlled by a gating network (*Model AMC1*) cannot provide good specialization, and the task remains unsolved.
- Several Critic modules a priori associated with different subparts of the task (*Model AMC2*) and connected to a single Actor (an Actor component being composed of a 6 channels GPR) allow learning to extend to areas that are distant from reward location, but still suffer from interferences between signals sent by the different Critic to the same single Actor.

Model MAMC2, combining several Critic modules with the principle of *Model AMC2*, and implementing several Actor components produces better results in the task at matter, spreading learning in the whole maze and reducing the learning duration. However, there are a few questions that have to be raised concerning the biological plausibility and the generalization ability of this model.

4.1. Biological plausibility of the proposed model

When using a single GPR Actor, each action is represented in only one channel – an Actor module being constituted of one channel per action (Gurney *et al.*, 2001) – and the structural credit assignment problem – which action to reinforce when getting a reward – can be simply solved: the action that has the highest salience inhibits its neighbors via local recurrent inhibitory circuits within D1 striatum (Brown and Sharp, 1995). As a consequence, only one channel in the Actor will have enough pre- and post-synaptic activity to be eligible for reinforcement.

When using several Actor modules, this property is not true anymore: even if only one channel per Actor module may be activated at a given time, each Actor module will have its own activated channel, and several concurring synapses would be eligible for reinforcement within the global Actor. To solve this problem, we

considered in our work that only one channel in the entire Actor is eligible at a given time. However, this implies for the basal ganglia to have one of the two following characteristics: it should either exist non-local inhibition between Actor modules within the striatum, or there should be some kind of selectivity in the dopamine reinforcement signals so that even if several channels are activated, only those located in the target module receives dopamine signals.
To the best of our knowledge, these characteristics were not found in the basal ganglia, and a few studies tend to refute the dopamine selectivity (Pennartz, 1996).

4.2. Computational issues

Several computational issues need also to be addressed. First, the results presented here show that the learning process was not perturbed by the fact to use an Actor detailing the action selection process in the basal ganglia. This Actor has the property to take into account some persistence provided by the cortex-basal ganglia-thalamus-cortex loops. The way this persistence precisely influence the learning process in the different principles compared in this work was not thoroughly studied here. However we suspect that persistence could probably challenge the way different Actors interact with Critic's experts, as switching between actions does not exactly follow switches in sensorimotor contexts with this model. This issue should be examined in a future work.

Generalization ability of the multi-module Actor: Another issue that needs to be addressed here is the generalization ability of the multi-module Actor model used in this experiment. Indeed, Model MAMC2 avoids interferences in the Actor because hand-determined subzones of the maze are absolutely disjoint. In other words, learned stimulus-response associations in a given zone cannot be performed in another zone, and do not interfere with the learning process is this second zone even if visual contexts associated to each of them are very similar. However, this leads also to an inability to generalize from one zone to the other: even if the distinction we made between two zones seemed relevant for the plus-maze task, if these two zones are similar and would imply similar motor responses in another task, the animat would have to learn twice the same sensorimotor association – one time in each zone. As a consequence, the partition we set in this work is task-dependent.
Instead, the model would need a partitioning method that autonomously classifies sensory contexts independently from the task, can detect similarities between two different contexts and can generalize learned behaviors in the first experienced context to the second one.

About the precise time of reward delivery:
In the work presented here, the time of reward delivery depends exclusively on the animat's behavior, which differs from several other S-R tasks used to validate Actor-Critic models of the basal ganglia. In these tasks, there is a constant duration between a stimulus and a reward, and several Actor-Critic models were designed so as to describe the precise temporal dynamics of dopaminergic neurons in this type of task (Montague *et al.*, 1996). As a consequence, numerous Actor-Critic models focused on the implementation of a time component for stimulus representation, and several works capitalized on this temporal representation for the application of Actor-Critic models of reinforcement learning in the basal ganglia to robotics (Perez-Uribe, 2001; Sporns and

Alexander, 2002). Will we need to add such a component to our model to be able to apply it to certain type of natural tasks, or survival tasks?

In the experiments presented here, we didn't need such a temporal representation of stimuli because there was sufficient information in the continuous sensory flow perceived by the animat during its moves, so that the model can dynamically adapt its reward predictions, as observed also in another work (Baldassarre and Parisi, 2000). For example, when the animat is at the center of the maze, perceives the white wall (stimulus predicting reward) and moves towards reward location, the latter stimulus becomes bigger in the visual field of the animat, and the model can learn to increase its reward prediction, as shown in figure 8. We didn't aim at explaining the depression of dopamine neurons' firing rates when a reward doesn't occur, nevertheless we were able to observe this phenomenon in cases where the animat was approaching the reward site, was about to consume it, but finally turned away from it (R events in figure 8).

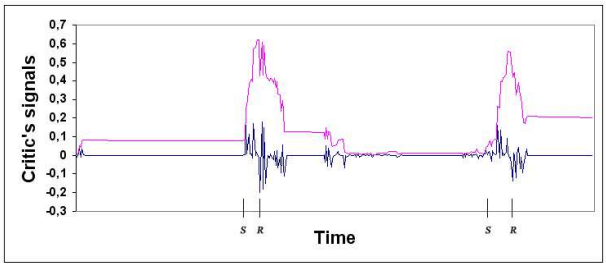

Figure 8 Reward prediction (light curve) and dopamine reinforcement signal (dark curve) computed by Critic of Model *MAMC2* in the 3D environment. X-axis: time. Y-axis: Critic's signals. S : perception of the stimulus (white wall) by the animat; R: Reward missed by the animat.

Using Critics dependent or independent from the performance: In our experiments, *Model AMC1*, implementing a gating network for experts' credibilities computation, did not solve the task. We saw in section 2 that, during the simulations, one expert became rapidly the most credible, which forced the model to use only one neuron to solve the task. The use of gating networks in the frame of mixture of experts methods has already being criticized (Tang *et al.*, 2002). According to these authors, this approach works well on problems composed of disjoint regions but does not generalize well, suffering from effects on boundaries of regions.

In our case, we explain the failure in the experts' specialization with *Model AMC1* by the observation that until the model has started to learn the task, and so can propagate teaching signals to the rest of the maze, only reward location has a value. As a consequence, it is the only area where the gating network tries to train an expert, and the latter rapidly reaches a high credibility. Then, as reward value starts to be extended to a new zone, this same expert still has the best credibility while getting bad performances. Other experts do not have significantly better performances – since they were not trained yet and since the new area and the first one are not disjoint. As a consequence, they remain non credible and the model starts having bad performances.

In his work, Baldassarre managed to obtain a good specialization of experts (Baldassarre, 2002). This may be partly explained by the fact that his task involved three different rewards located in three different sensory contexts. The simulated robot had to visit all rewards alternatively since the very beginning of the task. This may have helped the gating network to attribute good credibilities to several experts. However, reward

locations in Baldassarre's task are not perfectly disjoint, which result in a difficult specialization: one of the experts is the most credible for two of the three rewards (see Baldassarre, 2002).
Another model (Tani and Nolfi, 1999) proposes a different mixture of experts where the gating network is replaced with a dynamical computation of experts' credibilities. Their model managed to categorize the sensori-motor flow perceived by a simulated robot during its movements. However, their method does not use any memory of associations between experts' credibilities and different contexts experienced during the task. As a consequence, experts' specialization is even more dependent to each expert's performance than Baldassarre's gating network, and suffers from the same limitation when applied to reinforcement learning in our plus-maze task - as we experimented in unpublished work.

Combining self-organizing maps with mixture of expert: To test the principle of dissociating the experts credibility from their performance, we partitioned the environment into several sub-regions. Yet, this method is ad hoc, lacks autonomy, and suffers generalization abilities if the environment is changed or becomes more complex. We are currently implementing Self-Organizing Maps (SOM) as a method of autonomous clustering of the different sensory contexts will be used to determine these zones. Note that this proposition differs from the traditional use of SOM to cluster the state space input to experts or to Actor-Critic models (Smith, 2002; Lee *et al.*, 2003). It is rather a clustering of the credibility space, which was recently proposed by Tang *et al.* (2002). We also would like to compare the use of SOM with the use of place cells. Indeed models of hippocampal place cells have already been used for coarse coding of the input state space to the Actor and the Critic (Arleo and Gerstner, 2000; Foster *et al.*, 2000; Strösslin, 2004) but, in our case, we would like to use place cells to determine experts' credibilities.

4.3. Future work

As often mentioned in the literature, and as confirmed in this work, the application of Actor-Critic architectures to continuous tasks is more difficult than their use in discrete tasks. Several other works have been done on the subject (Doya, 2000). However, these architectures still have to be improved so as to decrease their learning time:
Particularly, the learning performance of our animat seems still far from the learning speed that real rat can reach in the same task (Albertin *et al.*, 2000), even if the high time constant that we used in our model does not allow a rigorous comparison yet (cf. parameters table in the appendix). This could be at least partly explained by the fact that we implemented only S-R learning (or habit learning), whereas it has recently been known that rats are endowed with two distinct learning systems related to different cortex-basal ganglia-thalamus loops: a habit learning system and a goal-directed learning one (Ikemoto and Panksepp, 1999; Cardinal *et al.*, 2002). The latter would be fast, used at the early stages of learning, and implying an explicit representation of rewarding goals or an internal representation of action-outcome contingencies. The former would be very slow and takes advantage of the latter when the animat reaches good performances and becomes able to solve the task with a reactive strategy (S-R) (Killcross and Coutureau, 2003; Yin *et al.*, 2004).
Some theoretical work has already been started to extend Actor-Critic models to this functional distinction (Dayan, 2001). In the practical case of our artificial rat, both such systems could be useful in two different manners.

First, it could be useful to upgrade the exploration function. This function could have an explicit representation of different places of the environment, and particularly of the reward site. Then, when the animat gets reward for the first time, the exploration function would guide it trying behaviors that can allow it to reach the explicitly memorized reward location. The function could also remember which behaviors have already been tried unsuccessfully in the different areas, so that untried behaviors are selected instead of random behaviors in the case of exploration. This would strengthen the exploration process and is expected to increase the animat's learning speed.

The second possible use of a goal-directed behavior component is to represent the type of reward the animat is working for. This can be useful when an animat has to deal with different rewards (food, drink) so as to satisfy different motivations (hunger, thirst). In this case, a component that chooses explicitly the current reward the animat takes as an objective can select sub-modules of the Actor that are dedicated to the sequence of behaviors that leads to the considered reward. This improvement would serve as a more realistic validation of the artificial rat Psikharpax when it has to survive in more natural environments, satisfying concurrent motivations.

Acknowledgments

This research has been granted by the LIP6 and the Project *Robotics and Artificial Entities* (ROBEA) of the Centre National de la Recherche Scientifique, France. Thanks for useful discussions to Drs. Angelo Arleo, Gianluca Baldassarre, Francesco Battaglia, Etienne Koechlin and Jun Tani.

References

Aizman, O., Brismar, H., Uhlen, P., Zettergren, E., Levey, A. I., Forssberg, H., Greengard, P. & Aperia, A. (2000). Anatomical and Physiological Evidence for D1 and D2 Dopamine Receptors Colocalization in Neostriatal Neurons. *Nature Neuroscience*, 3(3):226-230.

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.

Albin, R. L., Young, A. B. & Penney, J. B. (1989). The functional anatomy of basal ganglia disorders. *Trends in Neuroscience*, 12:366-375.

Arleo, A. & Gerstner, W. (2000). Spatial Cognition and Neuro-Mimetic Navigation: A Model of the Rat Hippocampal Place Cell Activity. *Biological Cybernetics*, Special Issue on Navigation in Biological and Artificial Systems, 83:287-299.

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

Baldassarre, G. & Parisi, D. (2000). Classical and instrumental conditioning: From laboratory phenomena to integrated mechanisms for adaptation. In Meyer *et al.* (Eds), *From Animals to Animats 6: Proceedings of the*

Sixth International Conference on Simulation of Adaptive Behavior, Supplement Volume (pp.131-139). The MIT Press, Cambridge, MA.

Brown, J., Bullock, D. & Grossberg, S. (1999). How the Basal Ganglia Use Parallel Excitatory and Inhibitory Learning, or Incentive Salience ? *Brain Research Reviews*, 28:309-369.

Brown, L. & Sharp. F. (1995). Metabolic Mapping of Rat Striatum: Somatotopic Organization of Sensorimotor Activity. *Brain Research*, 686:207-222.

Bunney, B. S., Chiodo, L. A. & Grace, A. A. (1991). Midbrain Dopamine System Electrophysiological Functioning: A Review and New Hypothesis. *Synapse*, 9:79-84.

Burgess, N., Jeffery, K. J. & O'Keefe, J. (1999). Integrating Hippocampal and Parietal Functions: a Spatial Point of View. In Burgess, N. *et al.* (Eds), *The Hippocampal and Parietal Foundations of Spatial Cognition*, pp.3-29, Oxford University Press, UK.

Cardinal, R. N., Parkinson, J. A., Hall, J. & Everitt, B. J. (2002). Emotion and Motivation: The Role of the Amygdala, Ventral Striatum and Prefrontal Cortex. *Neuroscience Biobehavioral Reviews*, 26(3):321-352.

Dayan., P. (2001). Motivated Reinforcement Learning. *NIPS*, 14: 11-18. The MIT Press.

Daw, N. D. (2003). *Reinforcement Learning Models of the Dopamine System and Their Behavioral Implications*. PhD Thesis, Carnegie Mellon University, Pittsburgh, PA.

Doya, K. (2000). Reinforcement Learning in Continuous Time and Space. *Neural Computation*, 12:219-245.

Doya, K., Samejima, K., Katagiri, K. & Kawato, M. (2002) Multiple Model-based Reinforcement Learning. *Neural Computation*. 14(6):1347-69.

Filliat, D., Girard, B., Guillot, A., Khamassi, M., Lachèze, L., Meyer, J.-A. (2004). State of the artificial rat Psikharpax. In Schaal *et al.* (Eds), *From Animals to Animats 8: Proceedings of the Eighth International Conference on Simulation of Adaptive Behavior*, pp.2-12. The MIT Press, Cambridge, MA.

Foster, D., Morris, R. & Dayan, P. (2000). Models of Hippocampally Dependent Navigation using the Temporal Difference Learning Rule. *Hippocampus*, 10:1-16.

Frank, M. J., Loughry, B. & O'Reilly, R. C. (2001). Interactions Between Frontal Cortex and Basal Ganglia in Working Memory: A Computational Model. *Cognitive, affective and behavioral neuroscience*, 1(2):137-160.

Gerfen, C. R., Herkenham, M. & Thibault, J. (1987). The Neostriatal Mosaic. II. Patch- and Matrix- Directed Mesostriatal Dopaminergic and Non-Dopaminergic Systems. *Journal of Neuroscience*, 7:3915-3934.

Girard, B., Cuzin, V., Guillot, A., Gurney, K. & Prescott, T. (2003). A Basal Ganglia inspired Model of Action Selection Evaluated in a Robotic Survival Task. *Journal of Integrative Neuroscience*, 2(22), 179-200.

Gurney, K. N., Prescott, T. J. & Redgrave, P. (2001a). A Computational Model of Action Selection in the Basal Ganglia. I. A new functional anatomy. *Biological Cybernetics*. 84, 401-410.

Gurney, K. N., Prescott, T. J. & Redgrave, P. (2001b). A Computational Model of Action Selection in the Basal Ganglia. II. Analysis and simulation of behavior. *Biological Cybernetics*. 84, 411-423.

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.* The MIT Press, Cambridge, MA.

Ikemoto, S. & Panksepp., J. (1999). The Role of the Nucleus Accumbens Dopamine in Motivated Behavior: A Unifying Interpretation with Special Reference to Reward-Seeking. *Brain Research Reviews*, 31:6-41.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. & Hinton, G.E. (1991). Adaptive Mixture of Local Experts. *Neural Computation*, 3:79-87.

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

Joel, D. & Weiner, I. (2000). The Connections of the Dopaminergic System with Striatum in Rats and Primates: An Analysis with respect to the Functional and Compartmental Organization of the Striatum. *Neuroscience*, 96:451-474.

Killcross, A. S. & Coutureau, E. (2003). Coordination of Actions and Habits in the Medial Prefrontal Cortex of Rats. *Cerebral Cortex*, 13(4):400-408.

Lee, J. K. & Kim, I. H. (2003). Reinforcement Learning Control Using Self-Organizing Map and Multi-Layer Feed-Forward Neural Network. In *International Conference on Control Automation and Systems, ICCAS 2003*.

McNaughton, B. L. (1989). Neural Mechanisms for Spatial Computation and Information Storage. In Nadel *et al.* (Eds), *Neural Connections, Mental Computations*, chapter 9, pp.285-350, MIT Press, Cambridge, MA.

Montague, P. R., Dayan, P. & Sejnowski, T. J. (1996). A framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning. *Journal of Neuroscience*, 16:1936-1947.

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.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K. & Dolan, R. (2004). Dissociable Roles of Dorsal and Ventral Striatum in Instrumental Conditioning. *Science*, 304:452-454.

Pennartz, C. M. A. (1996). The Ascending Neuromodulatory Systems in Learning by Reinforcement: Comparing Computational Conjectures with Experimental Findings. *Brain Research Reviews*, 21:219-245.

Perez-Uribe, A. (2001). Using a Time-Delay Actor-Critic Neural Architecture with Dopamine-like Reinforcement Signal for Learning in Autonomous Robots. In Wermter *et al.* (Eds), *Emergent Neural Computational Architectures based on Neuroscience: A State-of-the-Art Survey* (pp. 522-533). Springer-Verlag, Berlin.

Schultz, W. (1998). Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology*, 80(1):1-27.

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

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

Smith, A. J. (2002). Applications of the Self-Organizing Map to Reinforcement Learning. *Neural Networks*, 15 (8-9):1107-1124.

Sporns, O. & Alexander, W. H. (2002). Neuromodulation and Plasticity in an Autonomous Robot. *Neural Networks*, 15:761-774.

Strösslin, T. (2004). *A Connectionnist Model of Spatial Learning in the Rat.* PhD thesis, EPFL, Swiss Federal Institute of Technology, Swiss.

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

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

Tang, B., Heywood, M. I. & Shepherd, M. (2002). Input Partitioning to Mixture of Experts. In *International Joint Conference on Neural Networks*, pp. 227-232, Honolulu, Hawaii.

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.

Thierry, A.-M., Gioanni, Y., Dégénetais, E., and Glowinski, J. (2000). Hippocampo-prefrontal cortex pathway: anatomical and electrophysiological characteristics. *Hippocampus*, 10:411-419.

Yin, H. H., Knowlton, B. J. & Balleine, B. W. (2004). Lesions of Dorsolateral Striatum Preserve Outcome Expectancy but Disrupt Habit Formation in Instrumental Learning. *European Journal of Neuroscience*, 19 (1):181-189.

Appendix : Parameters Table

Symbo	***Valu***	***Description***
Δt	1 sec.	Time constant – time between two successive iterations of the model.
α	40 iter.	Time threshold to trigger the exploration function.
g	0.98	Discount factor of the Temporal Difference learning rule.
η	0.01	Learning rate of the Actor and Critic modules.
N	30	Number of experts in the Critic of Models *AMC1*, *AMC2* and *MAMC2*.
σ	2	Scaling parameter in the mixture of experts of Model *AMC1*.
m	0.1	Learning rate of the gating network in Model *AMC1*.