Biological cumulative learning through intrinsic motivations : A simulated robotic study of the development of visually-guided reaching

Vieri G. Santucci, Gianluca Baldassarre, Marco Mirolli
Istituto di Scienze e Tecnologie della Cognizione, CNR
Via san Martino della Battaglia 44, 00185 Roma
{vieri.santucci, gianluca.baldassarre, marco.mirolli}@istc.cnr.it

Abstract

This work aims to model the ability of biological organisms to achieve cumulative learning, i.e. to learn increasingly more complex skills on the basis of simpler ones. In particular, we studied how a simulated kinematic robotic system composed of an arm and an eye can learn the ability to reach for an object on the basis of the ability to systematically look at the object, which, in our set-up, represented a prerequisite for the reaching task. We designed the system by following several biological constraints and investigated which kind of sub-task reinforcements might facilitate the development of the final skill. We found that the performance in the reaching task was optimized when the reinforcement signal included not only the extrinsic reinforcement provided by touching the object but also an intrinsic reinforcement given by the error in the prediction of fovea activation. We discuss how these results might explain biological data regarding the neural basis of action discovery and reinforcement learning, in particular with respect to the neuromodulator dopamine.

1. Introduction

One of the characteristics of biological organisms is the ability to achieve cumulative learning, i.e. the possibility to learn different skills that are dependent on each other. For example, in human infants the development of visually-guided reaching seems to depend on the development of the ability to orient the eyes towards the objects (Georgopoulos, 1986; Land, 2006). What are the characteristics of the brain of natural organisms that might support such cumulative learning processes?

In organisms action selection and learning seems to take place in the basal ganglia (BG: Doya, 2000; Graybiel, 2005). Many studies suggest that the dorsal regions of the BG implement the actor-critic reinforcement learning architecture that learns through the temporal difference (TD) learning algorithm (Sutton and Barto, 1998). In particular, the phasic burst of the neuromodulator dopamine has been proposed to represent the TD error learning signal (Houk et. al., 1995; Schultz et al., 1997; Schultz, 2002). Moreover, the two classes of BG input neurons have been proposed to represent, respectively, the critic and the actor of the actor-critic reinforcement learning model (Barto, 1995, Joel et. al., 2002; Khamassi et. al., 2005): (a) striosome neurons, which project to dopaminergic (DA) neurons in the substantia nigra pars compacta (SNc) and in the ventral tegmental area (VTA), and are supposed to implement the critic which evaluates the current state and provides the learning signal; (b) matrix neurons, which project to the cortical sensorymotor areas through the pallidal neurons and the thalamus, and are supposed to implement the actor which selects the actions to be performed.

The ability of cumulative learning might be supported in natural organisms by a certain level of modularity in animals action control systems. In particular, there is evidence (Romanelli et al., 2005) that different effectors (e.g. the eye and the hand) are controlled by different basal-ganglia–thalamo–cortical pathways. On the other hand, it seems likely that the reinforcement signals that determine the phasic activation of dopaminergic neurons are unique for all the sensory-motor subsystems. This might generate a non-trivial problem, since the reinforcement signals caused by the actions of one controller might interfere with the learning of other controllers. In computational terms, the problem is given by the fact that reinforcement learning algorithms (including TD learning) have been developed for solving Markovian problems (Sutton and Barto, 1998) in which both the transitions between states and the reinforcements depend exclusively on the currently perceived state and on the system's actions, while the described biological organization seems to imply a non-Markovian situation in which both the transitions and the reinforcements for one sub-controller depend also by the actions and the states of the others.

A possible solution to this problem might be found in another important biological phenomenon related to dopamine. There is ample evidence (Horvitz, 2000) that phasic DA is triggered not only by extrinsic rewards (food, sex, etc.) but also by any kind of salient stimuli: in particular, by unexpected changes in the environment. These and other findings led Redgrave and Gurney (2006) to criticize the parallel between DA and TD signals, and to suggest that the phasic dopamine could represent a “novelty” signal that would permit the discovery and learning of new actions.

Here, we propose that the two opposing positions regarding phasic dopamine might in fact be reconciled by considering DA as a sort of TD signal of a reinforcement learning system that is determined not only by extrinsic reinforcements but also by "intrinsic" reinforcements provided by unexpected events. In particular, we propose that the activation of dopaminergic neurons by unexpected and novel stimuli might constitute (part of) the neural basis of what psychologists have been called "intrinsic motivations" (White, 1959; Ryan and Deci, 2000), i.e. motivations that are not related to external rewards but rather to the agent's knowledge and/or competence.

Recently, the topic of intrinsic motivations has been gaining increasing interest in the robotics and machine learning communities (Schmidhuber, 1991a-b; Huang and Weng, 2002; Kaplan and Oudeyer, 2003; Barto et al., 2004; Oudeyer et al., 2007; Schembri et al., 2007a-c; Lee et al., 2009), but in general this kind of works only consider computational issues (but see also Kaplan and Oudeyer, 2007). On the contrary, in this paper we address the topic of intrinsically-motivated cumulative learning from the point of view of biological systems.

In order to model biological cumulative learning, we investigated how a learning system that follows the aforementioned biological constraints, and that controls a simulated arm and an eye, might acquire the ability to reach for objects on the basis of the ability to appropriately look at them. In particular, we compared the results of different experiments in which we varied the sources of the reinforcement signals and we found that the performance in the reaching task was optimized when the reinforcement signal included not only the extrinsic reinforcement provided by reaching the object but also an intrinsic reinforcement given by the error in the prediction of the activation of the fovea.

The rest of the paper is structured as follows. Section 2 presents the experimental set-up, section 3 shows the results and section 4 concludes by discussing the relevance of the results, in particular with respect to the biological basis of the intrinsically-motivated cumulative learning of skills.

2. Set up

2.1. The task

The simulated robotic system is composed of an arm and an eye working on a two-dimensional plane and its task is to learn and reach for an object randomly placed on a table (figure 1).

The arm is composed of two segments (arm and fore-arm) which are 4.85 and 3.0 units long, respectively; each of the two joints (shoulder and elbow) can move within the interval [0 180] degrees, within a maximum step of 25 degrees in either direction. The object to be reached, which is a circle with a diameter of 0.5, is randomly placed in front of the robot on a rectangular table whose dimensions are 4 and 7 respectively, so that every point of the table is reachable by the robot's hand. The eye of the robot can move on both the x and y axes with a maximum step of 8 units in each direction. The visual field is a square with a size of 14, so that the eye can always perceive the object, even when it foveates outside the table.

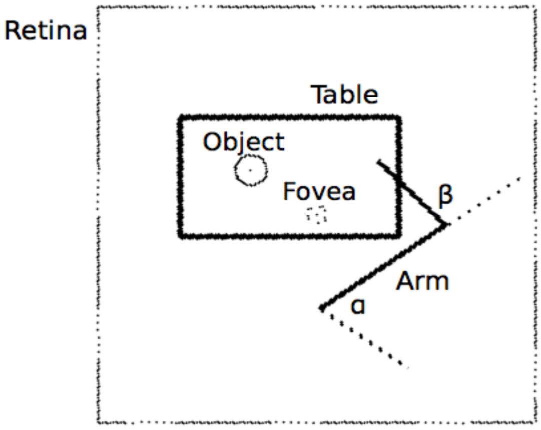

Figure 1: The experimental set-up, with the retina, the fovea, the arm and the table with the object.

The sensory system of the robot includes the proprioception (the angles of the two joints of the arm) and the visual perception of its hand and the object. Furthermore, the robot has also a rudimentary "fovea", which consists of a single sensor that is activated if (a part of) the object is perceived in the centre of the visual field, and a touch sensor, which is activated if the hand touches the object (i.e., if it is within the circle represented by the object, since for simplicity collisions are not simulated and objects are penetrable). The goal of the robot is to reach for the object and touch it as much as possible. The robot is trained through a classical reinforcement learning algorithm (Sutton and Barto, 1998) where a reinforcement of 1 is given whenever the hand touches the object.

Since we were interested in the cumulative learning of different skills in cases where more complex skills depend on simpler ones, we devised the set up so that the skill of reaching crucially depends on the skill of foveating the object, as it seems to happen in the development of human reaching. In particular, in the set-up the controller of the arm receives visual information regarding the position of the hand with respect to the eye, but does not receive any information regarding the position of the object. In this way, the ability to reach the object can only be developed after the eye has learnt to look at the object in a systematic way, so that the information regarding the position of the hand with respect to the centre of the visual field indirectly provides information about the relationship between the hand and the object.

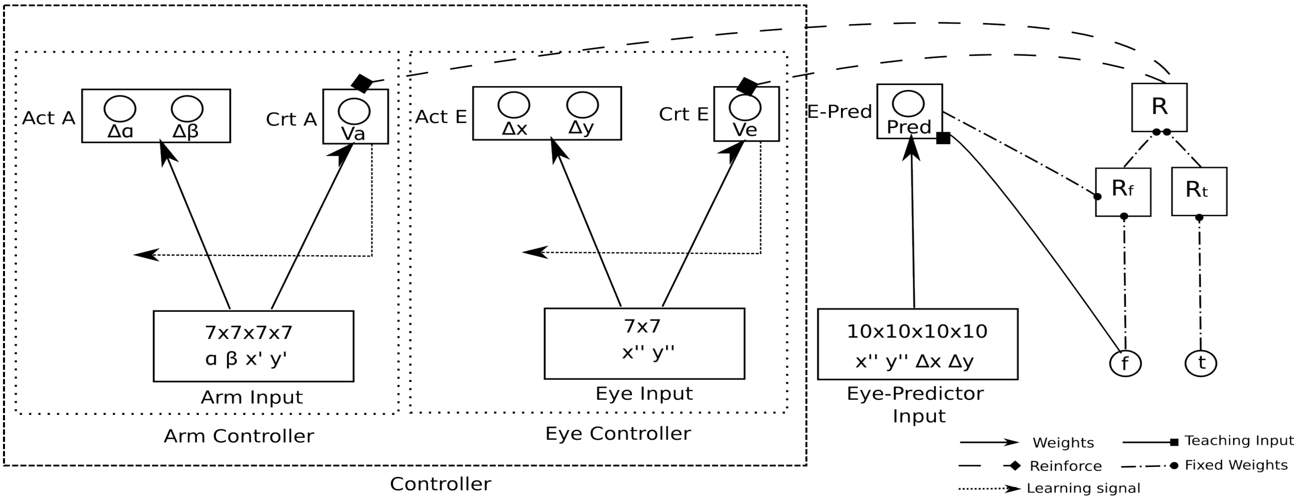

Figure 2 : The controller with its two sub-components (arm and eye controllers), the eye-predictor, and the reinforcement system. α and β are the angles of the two arm joints; x' and y' are the distances of the hand with respect to the center of the fovea on the x and y axes, respectively; $\Delta\alpha$ and $\Delta\beta$ are the variations of angles α and β, respectively, as determined by the arm's actor; *Va* is the evaluation of arm's critic; x'' and y'' are the distances of the object with respect to the fovea on the x and y axes, respectively, Δx and Δy are the displacements of the eye on the x and y axes, respectively, as determined by eye's actor; *Ve* is the evaluation of eye's critic; *Pred* is the prediction of the eye-predictor; *f* is the activation of the fovea sensor; *t* is the activation of the touch sensor; *Rf* and *Rt* are the reinforcements related to foveating and to touching the object, respectively; *R* is the total reinforcement. See text for details.

As we wanted to test the idea that the cumulative learning of skills in biological agents is improved by intrinsic motivations, we confront the learning of our system under three conditions, differing only with respect to the reinforcement signal that drives learning: (A) the one just described, in which the reinforcement signal is given only by the touch sensor; (B) one in which a further reinforcement is given also for the sub-goal of foveating the object; (C) one in which the further reinforcement consists in a *surprise* signal (prediction error) relative to the prediction of the activation of the foveal input (see section 2.3).

2.2. The controller

Fig. 2 shows the controller of the system. As we described in the introduction, we tried to model some of the areas and mechanisms (fig. 3) that are involved in the biological cumulative learning of actions. The controller consists of two sub-controllers, one dedicated to the control of the eye (eye-controller) and the other to the control of the arm (arm-controller). Both sub-controllers are neural network implementations of the actor-critic architecture (Sutton and Barto, 1998) adapted to work with continuous state and action spaces (Doya, 2000; Schembri et al., 2007a), in discrete time.

For all the inputs to the system we use population coding through Gaussian radial basis functions (RBF) (Pouget and Snyder, 2000):

$$a_i = e^{-\sum_d \left(\frac{c_d - c_{id}}{2\sigma_d^2}\right)^2}$$

where a_i is the activation of input unit i, c_d is the input value of dimension d, c_{id} is the preferred value of unit *i* with respect to dimension *d*, and σ_d^2 is the width of the Gaussian along dimension *d* (widths are parametrized so that when the input is equidistant, along a given dimension, to two contiguous neurons the activation of both of these is 0.5).

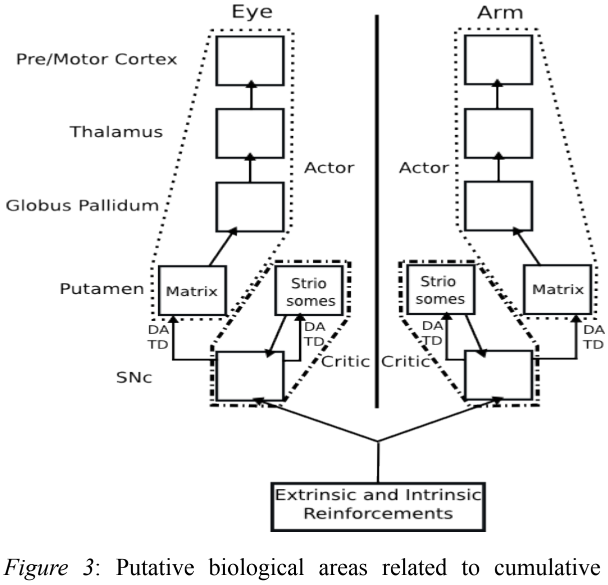

Figure 3: Putative biological areas related to cumulative learning corresponding to the components of the model.

The inputs to the eye-controller are the coordinates (x and y) of the object with respect to the centre of the visual field, uniformly distributed on a 7x7 grid (in the range [-7, 7], i.e., there are 49 units whose preferred inputs are uniformly distributed on the two-dimensional space whose origin is the centre of the visual field).

The eye-controller actor has two output units which receive connections from all its input units and have a sigmoidal activation function:

$$o_j = \Phi\left(b_j + \sum_i^N a_i w_{ij}\right) \qquad \Phi(x) = \frac{1}{1-e^{-x}}$$

where b_j is the bias of output unit j, N is the number of input units, and w_{ji} is the weight of the connection linking input unit i to output unit j. The eye-controller motor commands correspond to the activations of the two eye output units plus noise:

$$o_j^n = o_j + r$$

where o_j^n is the motor command i and r is a random value uniformly drawn in [– 0.02, 0.02]. The two eye motor commands (in [0, 1]) are then remapped in [–8, 8] and determine the displacement of the eye in the two dimensions (x and y, respectively).

The inputs to the arm-controller are the angles of the two arm joints α and β (in [0, 180]) and the coordinates x and y of the hand with respect to the centre of the visual field (in [-7, 7]), uniformly distributed in a 7x7x7x7 grid. The arm-controller has two output units that are fully connected with the input units and have a sigmoidal activation function as the eye output units. The arm-controller motor commands correspond to the activations of the two arm output units plus random value uniformly drawn in [-0.2, 0.2]. The two motor commands (in [0, 1]) are then remapped in [-25, 25] and determine the displacement of the two arm joints (Δα and Δβ).

Each of the sub-controllers is also endowed with a critic, whose output unit is a linear combination of the respective input units.

In the condition in which learning is driven not only by the touch sensor but also by the system "surprise" with respect to the foveal input, the controller is also endowed with a predictor whose task is to predict the activation of the foveal sensor. The input to the predictor is another population of RBF units that encode the coordinates of the object with respect to the centre of the visual field (in the range of [-7, 7]) and the motor commands determined by the eye-controller output units (in [0, 1]), uniformly distributed in a 10x10x10x10 grid. The output of the predictor is a single sigmoidal unit with activation p fully connected with the input units.

2.3. Learning

Following biological constraints, both the sub-controllers receive the same reinforcement signal (R), which is defined as:

$$R = R_t + R_f$$

where R_t is the activation of the touch sensor t (1 if the hand touches the object, 0 otherwise). R_f varies according to the three experimental conditions: in condition A (normal) it is always 0; in condition B (fovea) it equals the activation of the fovea sensor F (1 if the fovea perceives the object, 0 otherwise); in condition C (surprise) it is defined as $max[0, F-p]$ where p is the activation of the foveal predictor output.

Learning depends on the TD reinforcement learning algorithm (Sutton and Barto, 1998). For each sub-controller k, the TD error δ_k is calculated as usual:

$$\delta_k = R^t + \gamma_k V_k^t - V_k^{t-1}$$

where V_k^t is the output of the critic of controller k at time step t and γ_k is the discount factor, set to 0.9 for both the eye and the arm controllers. The weights of the critic k are updated in the standard way:

$$\Delta w_{ki} = \eta_k^c \delta_k a_i$$

where η_k^c is the learning rate , set to 0.01 for both the eye and the arm controllers.

The weights of the actor k are updated as follows:

$$\Delta w_{kji} = \eta_k^a \delta_k \left(o_j^n - o_j\right)\left(o_j\left(1-o_j\right)\right) a_i$$

where η_k^a is the learning rate (set to 0.1 for both the eye and the arm controller), and $\left(o_j\left(1-o_j\right)\right)$ is the derivative of the sigmoid function.

The learning of the predictor is supervised, with the actual activation of the fovea sensor (at time step $t+1$) acting as the teaching input for the prediction (at time t). The change in the weights is performed according to the standard delta rule:

$$\Delta w_i = \eta^p (F-p)\left(p(1-p)\right) a_i$$

where η^p is the learning rate, set to 0.01, F is the activation of the fovea sensor (at t+1), p is the activation of the predictor's output (at t) and a_i is the activation of input i.

3. Results

Each experiment was run for 150000 trials, with each trial lasting 40 time steps, after which the object, the eye, and the arm were repositioned randomly, so the system could get several rewards in one trial. The object and the eye were always repositioned inside the table, while the arm joint angles were set randomly in [0, 180] (so that the hand could be outside the table). Every 100 trials we performed 100 test trials (during which learning is switched off), in which we recorded several data. For each condition we ran five replications of the experiment: all the presented data are average results of the five replications.

Fig. 4 shows the performance on the reaching task in the three experimental conditions. In the first condition, the arm increased its performance quite soon, but reached only sub-optimal values. In the second condition, where the reinforcement for the sub-task of foveation was introduced, the results got worse: although the system obtained the same sub-optimal performance as condition A, it took more trials to reach those values. In condition C, where the reinforcement related to foveation is a prediction error signal, the learning of the reaching ability speeded-up, and reached optimal performance (100%) after about 40000 trials.

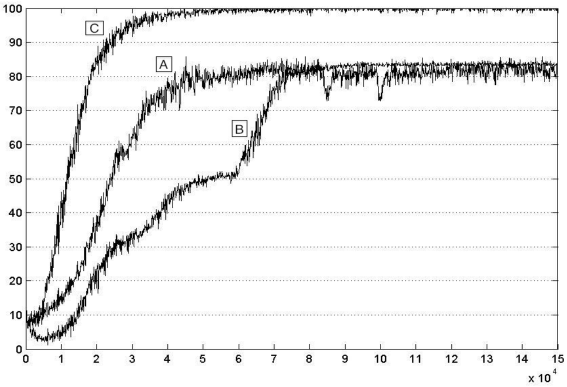

Figure 4: Average number of test trials in which the arm reaches the object (at least once) under the three conditions. A: normal condition; B: fovea condition; C: surprise condition. See text for details

Let us now consider the three conditions more in detail. In fig. 5 the performance of the system with respect to the sub-task of foveation is shown. In condition A, where the sub-task of foveation was not reinforced, the eye did not learn to foveate the object. Nonetheless, its behaviour was such that the arm could learn its task: the behaviour of the eye was to move always to a specific position with respect to the object, thus providing to the arm a constant input that indirectly brings the information on the position of the object, that is required for learning to reach (data not shown).

In condition B an explicit reinforcement for the sub-task of foveation was introduced. Although the eye very rapidly reached high performance in its task, the system ability in reaching for the object did not improve compared to condition A; rather the learning process was slowed down. This is because the new reinforcement signal interfered with the learning of the final task. Indeed, as described in section 2.3, the reinforcement is composed by the sum of the signals coming from the fovea and the touch sensors; so, because the eye has learnt its task very quickly (after only 5000 trials it reaches 100% performance value), the learning of the final task was impaired by the frequent reinforcement signal coming from the eye foveating the object. The critic of the arm has no information to predict the reinforcement signal coming from the eye activity: as we said in the introduction, this is a typical non-Markovian problem, which causes problems to the RL algorithm used, not devised to tackle this situation. Moreover, despite the reaching for the object provided an increase in the reinforcement signal, this was not sufficient to rapidly drive the learning of the arm: in fact in a world where there is a frequent signal coming from the eye, the information carried by the reinforcement provided by the arm is lowered.

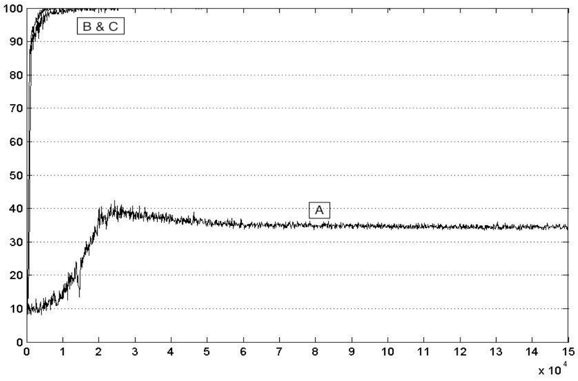

Figure 5: Average number of test trials in which the eye foveates the object (at least once) under the three conditions. A: normal condition; B: fovea condition; C: surprise condition. See text for details.

In condition C, the reinforcement for the eye reaching the object was determined by the prediction error (surprise) of the fovea activation, based on the activity of the predictor. As shown in fig. 5, also in this condition the eye very quickly learnt to foveate the object, at least once per trail. Fig. 6 shows that also the average number of steps in which the eye foveated the object raised very quickly. The learning of the predictor followed the learning of the eye-controller: when the eye started to foveate the object with continuity, the prediction error increased a little but as soon as the ability of the eye improved the predictor reliably learnt to predict the activation of the fovea sensor and the prediction error decreased again. As a consequence, after about 20000 trials, the reinforcement provided by the fovea prediction error (Rf, fig. 6), which drove the learning so far, had lowered again to chance value. At this time the arm was still not able to reach the object with satisfying continuity, so the total reinforcement R lowered as well. However, the reliable behaviour of the eye permitted the arm to start learning to reach the object: the reinforcement coming from the touch sensor (Rt, fig. 6) and the global reinforcement R increased and soon the system reached an almost optimal performance (fig. 4). Note that during the period in which reaching was learnt, the average time spent by the eye foveating the object kept increasing (from about 0.85 to about 0.95), while the reinforcement coming from the fovea prediction error was very low: hence, the further improvement of the foveating ability was likely driven not only by the reinforcement relative to the activation of the fovea but also from the one provided by the touch sensor.

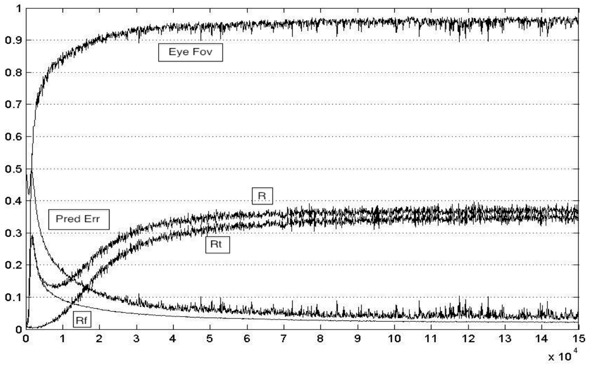

Figure 6: Several data of condition C: total reinforcement (R), fovea reinforcement (Rf), touch reinforcement (Rt), prediction error (Pred Err) of the predictor of fovea activation, and percentage of time-steps in which the eye foveates the object (Eye Fov).

4. Discussion and future works

This work presented a simulated robotic system that acquires the capacity to reach for an object after having learnt to systematically look at it, as required by the fact that the former ability depends on the latter. Biological considerations drove the design of the learning system and suggested the following assumptions: (1) actions are learnt through a reinforcement learning process that takes place in the basal ganglia (Doya, 2000; Graybiel, 2005); (2) the dorsal regions of the basal ganglia implement actor-critic reinforcement learning architectures (Barto, 1995; Joel et al. 2002; Daw et al. 2005); (3) learning is driven by the TD learning algorithm, with the phasic activation of the neuromodulator dopamine playing the role of the learning signal (Reynolds and Wickens, 2002) in analogy with the TD error (Houk et al., 1995; Schultz et al. 1997; Schultz, 2002); (4) different effectors (e.g., eye and hand) are controlled by different controllers (Romanelli et al., 2005); (5) there is a single reinforcement signal for the different controllers (Schultz, 2002).

We compared the results of three experiments in which we varied the sources of reinforcement. The system was able to develop the reaching ability when only the touching of the object was reinforced (condition A) but performance was sub-optimal because of the complexity due to the dependency of the behaviour of the arm on the one of the eye. Adding an explicit reinforcement for the sub-task of foveating the object (condition B) gave no improvement in the final performance of the system while in fact slowing down the learning process. The reason is that although the additional reinforcement drove the acquisition of the ability of looking at the object, it interfered with the learning of the reaching ability by providing irrelevant and unpredictable learning signals to the arm controller. The best results were obtained in the experiment in which the further reinforcement consisted in the “unpredicted” activation of the fovea (condition C). The reason is that such an intrinsic reinforcement signal is well suited for driving cumulative learning processes because it is present only when the intermediate skill has still to be acquired but fades away as soon as that ability has been learnt. In our model, for example, as soon as the skill to foveate had been learnt, the predictor learnt to predict the activation of the fovea sensor and started inhibiting the intrinsic reinforcement signal so that the system could focus on learning the reaching task on the basis of the reinforcement provided by touching the object, without the interference of the reinforcement coming from the eye activity.

The model can explain the otherwise puzzling neuroscientific evidence showing that dopamine is not only activated by biological rewards such as food, but also by other salient events like lights, tones and other novel or unexpected events (Horvitz, 2000, Dommett et al., 2005; Lisman and Grace, 2005). Furthermore, behavioural experiments have shown that apparently neutral stimuli like a light can be used for training animals to perform certain actions in instrumental conditioning tasks (Reed et al., 1996; see also Fiore et al., 2008 for a model that reproduces these data). In order to explain these and other evidence that contrasted with the interpretation of dopamine as reward prediction error (Schultz, 2002), Redgrave and Gurney (2006) proposed that phasic dopamine is in fact an intrinsic learning signal that allows the discovery and development of novel actions. The model presented in this paper can be considered as lying between these two opponent views of phasic dopamine. From the one hand, the model assumes that dopamine is in fact a form of reinforcement prediction error, playing the role of the TD signal in computational reinforcement learning; on the other, the model also assumes that a fundamental role of the dopamine signal is to drive the acquisition of new actions (skills) on the basis of the occurrence of unexpected events. In this respect, the model has shown how intrinsic reinforcements provided by unexpected events can lead to the acquisition of new skills, which in turn can be used for learning other abilities in a cumulative fashion.

In the computational literature on intrinsically-motivated learning, the idea of using a *prediction error* as an intrinsic reinforcement has been first proposed by Schmidhuber (1991a) and used in other subsequent models (e.g. Huang and Weng, 2004). However, Schmidhuber (1991b) argued that a prediction error is not a good intrinsic reinforcement signal as it can generate problems if the environment is unpredictable: in such a case, the reinforcement provided by the prediction error would never decrease and the system would get stuck in trying to reproduce unpredictable outcomes. To avoid this problem, the use of the *progress in the predictions* was proposed as a better intrinsic reinforcement, a solution that has been adopted in developmental robotic systems (e.g. Oudeyer et al., 2007). In contrast to this, in our model, as in biological systems, the intrinsic reinforcement signals that drive action learning depend on unpredicted events, not on progress in predictions.

How can we reconcile this with the problem of getting stuck on unpredictable events? We think that the problem of unpredictability might be solved (in real as well as in artificial systems) by further intrinsic motivation signals that work not at the level of the single skills, but at a higher level of the hierarchical organization of action, a level that is in charge of deciding which is the skill that has to be trained in each context. If such a level is trained on the basis of intrinsic rewards related to the *learning progress in skill acquisition*, as it happens in the work of Schembri et al. (2007a,b,c), then unpredictable events would not lead the system to get stuck in trying to reproduce them: if a skill cannot be learnt, the learning progress will be zero, and the system will move on and try to learn something else.

In future work, we plan to investigate this hypothesis by merging the use of unexpected events as intrinsic reinforcements for skill acquisition (as in the model of this paper) with the use of a hierarchical system in which intrinsic rewards are based on a measure of progress in skill acquisition (as in the models of Schembri et al., 2007b). We plan to test such a more complex system in a richer world, for example in which more than one object can be present and more complex sequences of skills can be performed. The goal of this future research is to investigate whether such a richer system in a richer environment can cumulatively learn a higher number of skills on the basis of intrinsic motivation signals.

Acknowledgements

This research was supported by the EU Project IM-CLeVeR, contract no. FP7-IST-IP-231722.

References

Barto, A. (1995). Adaptive critics and the basal ganglia. In Houk, J.C., Davis, J.L., Beiser, D.G. (Eds.), *Models of Information Processing in the Basal Ganglia*. MIT Press, Cambridge, MA, pp. 215-232.

Barto, A.; Singh, S., Chentanez, N. (2004). Intrinsically Motivated Learning of Hierarchical Collections of Skills. In *International Conference on Developmental Learning (ICDL).*

Berlyne, D. E. (1960). *Conflict, Arousal and Curiosity*. New York, McGraw Hill.

Daw, N.; Niv, Y. & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*. Vol. 8, pp. 1704-1711.

Dommett, E., Coizet, V., Blaha, C. D., Martindale, J., Lefebvre, V., Walton, N., Mayhew, J. E., Overton, P. G. & Redgrave, P. (2005). How Visual Stimuli Activate Dopaminergic Neurons at Short Latency. *Science*. Vol. 307, 5714, pp. 1476-1479.

Doya, K. (1999). What are the Computations of the Cerebellum, the Basal Gangila, and the Cerebral Cortex? *Neural Networks*. Vol. 12, 961-974.

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Computation.* Vol.12, 1, pp. 219-245.

Fiore, V. G., Mannella, F., Mirolli, M., Gurney, K., Baldassarre, G. (2008). Instrumental Conditioning Driven by Neutral Stimuli: A Model Tested with a Simulated Robotic Rat. In Schlesinger M., Berthouze L., Balkenius C. (eds.), *Proceedings of the Eight International Conference on Epigenetic Robotics.* Lund University Cognitive Studies 139, Lund: University of Lund. pp. 13-20.

Georgopoulos, A. P., (1986). On Reaching. *Annual Review of Neuroscience*. Vol. 9, pp. 147-170.

Graybiel, A. M. (2005). The basal ganglia: learning new tricks and loving it. *Current Opinion in Neurobiology.* Vol. 15, 6, pp. 638-644.

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non- reward events. *Neuroscience.* Vol. 96, 4, pp. 651-656.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use reward signals that predict reinforcement. In Houk, J.C., Davis, J.L., Beiser, D.G. (Eds.), *Models of information processing in the basal ganglia*. Cambridge: MIT Press. pp. 249-270.

Huang, X., Weng, J. (2002). Novelty and Reinforcement Learning in the Value System of Developmental Robots. In Prince, C.G., Demiris, Y., Marom, Y., Kozima, H., Balkenius, C. (Eds.), *Proceedings Second International Workshop on Epigenetic Robotics*. Vol. 94, pp. 47-55.

Joel, D., Niv, Y., Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*. Vol. 15, 4-6, pp. 535-547.

Kaplan, F., Oudeyer, P.-Y. (2003). Motivational principles for visual know-how development, In Prince, C.G., Berthouze, L., Kozima, H., Bullock, D., Stojanov, G.; Balkenius, C. (Eds.) *Proceedings of the Third International Workshop on Epigenetic Robotics*. Lund University Cognitive Studies, Lund, pp. 73-80.

Kaplan F. and Oudeyer P-Y. (2007). In search of the neural circuits of intrinsic motivation. *Frontiers in Neuroscience.* Vol.1, 1, pp.225-236.

Khamassi, M., Lacheze, L., Girard, B., Berthoz, A., Guillot, A. (2005). Actor-Critic Models of Reinforcement Learning in the Basal Ganglia: From Natural to Artificial Rats. *Adaptive Behavior*. Vol. 13, 2, pp. 131-148.

Land, M. F., (2006). Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research*. Vol. 25, 3, pp. 296-324.

Lee, R., Walker, R., Meeden, L., Marshall, J. (2009). Category-based intrinsic motivation, In Canamero, L., Oudeyer, P.-Y., Balkenius, C. (Eds.) *Proceedings of the Ninth International Conference on Epigenetic Robotics*. Lund University Cognitive Studies, Lund, pp. 81-88.

Lisman, J. E., Grace, A. A. (2005). The Hippocampal-VTA Loop: Controlling the Entry of Information into Long-Term Memory. *Neuron*. Vol. 46, pp. 703-713.

Oudeyer, P-Y., Kaplan, F., Hafner, V. V. (2007). Intrinsic Motivation System for Autonomous Mental Development. *IEEE Transactions on Evolutionary Computation.* Vol. 11, 2, pp. 265-286.

Pouget and Snyder, (2000) Computational approaches to sensorimotor transformations. *Nature Neuroscience*. Vol. 3, pp. 1192-1198.

Redgrave, P., Gurney, K. (2006). The Short-Latency Dopamine Signal: a Role in the Discovering Novel Actions?. *Nature Reviews Neuroscience.* Vol. 7, 12, pp. 967-975.

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

Reynolds, J. N. & Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*. Vol. 15, 4-6, pp. 507-521.

Romanelli, P., Esposito, V., Schaal, D. W., Heit, G. (2005). Somatotopy in the basal ganglia: experimental and clinical evidence for segregated sensorimotor channels. *Brain Research Reviews*. Vol. 48, pp. 112-128.

Ryan, R. M., Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology.* Vol. 25, pp. 54-67.

Schembri, M., Mirolli, M., Baldassarre, G. (2007a). Evolution and learning in an intrinsically motivated reinforcement learning robot. In Almeida y Costa, F., Rocha, L. M., Costa, E., Harvey, I., Coutinho, A., (Eds.) *Advances in Artificial Life*. Berlin, Springer, pp. 294-333

Schembri, M.; Mirolli, M. & Baldassarre, G. (2007b), Evolving childhood's length and learning parameters in an intrinsically motivated reinforcement learning robot. In Berthouze, L., Dhristiopher, G.P., Littman, M., Kozima, H., Balkenius, C. (Eds.) *Proceedings of the Seventh International Conference on Epigenetic Robotics.* Lund University Cognitive Studies, Lund, pp. 141-148.

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

Schmidhuber, J. (1991a). A possibility for implementing curiosity and boredom in model-building neural controllers. In Meyer, J. A. and Wilson, S. W., (Eds.), *Proc. of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, Cambridge, Massachusetts/London, England. MIT Press/Bradford Books, pp. 222-227.

Schmidhuber, J. (1991b). Curious model-building control system. *Proc. of International Joint Conference on Neural Networks*. Singapore, Vol. 2, pp- 1458-1463.

Schultz, W., Dayan, P., Montague, P. R. (1997). A neuronal substrate of prediction and reward. *Science*. Vol. 275, 5306, pp. 1593-1599.

Schultz, W. (2002), Getting Formal with Dopamine and Reward. *Neuron*, Vol. 36, pp. 241-263.

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

White, R. W. (1959). Motivation Reconsidered: The Concept of Competence. *Psychological Review.* Vol 66, 5, pp. 297-333.