A bio-inspired learning signal for the cumulative learning of different skills

Vieri G. Santucci[1,2] Gianluca Baldassarre[1] Marco Mirolli[1]

[1] Istituto di Scienze e Tecnologie della Cognizione (ISTC), CNR
Via San Martino della Battaglia 44, 00185, Roma, Italia
[2] School of Computing and Mathematics, University of Plymouth
Plymouth PL4 8AA, United Kingdom
{vieri.santucci, gianluca.baldassarre, marco.mirolli}@istc.cnr.it

Abstract. Building artificial agents able to autonomously learn new skills and to easily adapt in different and complex environments is an important goal for robotics and machine learning. We propose that providing artificial agents with a learning signal that resembles the characteristic of the phasic activations of dopaminergic neurons would be an advancement in the development of more autonomous and versatile systems. In particular, we suggest that the particular composition of such a signal, determined both by intrinsic and extrinsic reinforcements, would be suitable to improve the implementation of cumulative learning. To validate our hypothesis we performed some experiments with a simulated robotic system that has to learn different skills to obtain rewards. We compared different versions of the system varying the composition of the learning signal and we show that only the system that implements our hypothesis is able to reach high performance in the task.

1 Introduction

Building artificial agents able to autonomously form ample repertoires of actions and to easily adapt in different and complex environments is an important gola for robotics and machine learning. These characteristics are typical of biological agents that have the ability to autonomously learn new skills that can be useful for optimizing their survival probabilities. Moreover, these new skills can be combined together to generate complex sequences that can lead an agent to discover novel ways of interaction with the environment in a cumulative fashion. If we want to develop artificial agents with the ability of cumulatively learning different skills to improve their adaptive behaviour, a crucial issue [1] is to provide a proper signal to guide agents in the discovery and acquisition of novel actions and to deploy them in the appropriate situations.

The neuromodulator dopamine (DA) has long been recognized to play a fundamental role in motivational control and reinforcement learning processes [2–5]. In particular, phasic DA activations have been related to the presentation of unpredicted rewards [6–9] but also to other phasic, not reward-related, unexpected stimuli [10–13]. These data led to the formulation of two main hypotheses on the

functional role of DA signal. One hipothesis[14–16] looks at the similarities of DA activations with the temporal-difference (TD) error of computational reinforcement learning [17], and suggests that phasic DA represents a *reward prediction error* signal with the role of guiding the maximisation of future rewards through the selection of the appropriate actions. The second hypothesis [18–20] focuses on the activations for unexpected events and states that phasic DA is a *sensory prediction error* signal with the function of guiding the discovery and acquisition of novel actions.

As we pointed out in another work [21], we consider these two hypotheses both partially true, but at the same time not capable of taking into account all the empirical evidence on phasic DA activations. What we proposed in that work is that phasic DA represents a reinforcement prediction error learning signal analogous to the computational TD-error, but for a learning system that receives two different kinds of reinforcements: (1) temporary reinforcements provided by unexpected events, and (2) permanent reinforcements provided by biological rewards. In our hypothesis, the DA signal has the function of driving both the formation of a repertoire of actions and the maximisation of biological rewards through the deployment of the acquired skills.

Moreover, we suggest that phasic DA activations determined by unexpected events may constitute part of the neural substrate of what psychologists have been calling *intrinsic motivations* (IM) [22–24]. IM were introduced in the 1950s in animal psychology to explain experimental data (e.g.[25, 26]) incompatible with the classic motivational theory: what is crucial is that stimuli not related to (extrinsic) primary drives present a reinforcing value capable of conditioning instrumental responses [27–29].

What we propose in this paper is that providing artificial agents with a learning signal that resembles the characteristic of the phasic DA signal, determined both by intrinsic and extrinsic reinforcements, would be an advancement in the development of more autonomous and versatile systems. In particular, such a signal would be the proper one to improve the implementation of the cumulative learning of skills.

To test our hypothesis, we built a simulated robotic system that has to autonomously acquire a series of skills in order to maximise its rewards (sec. 2). We compare the performance of the system with different compositions of the learning signal and we show (sec. 3) that the system implementing our hypothesis is the only one that is able to learn the task. We then draw the conclusions (sec. 4) by analysing the results of the experiments and discussing the implications of our hypothesis.

2 Set up

2.1 The task

The system is a simulated kinematic robot composed of a fixed head with a "mouth" and a moving eye, and a two degrees of freedom kinematic arm with

a han that can "grasp object". The task consists in learning to eat food (i.e., bring a red object to the mouth) randomly placed on a rectangular table (with dimensions of 4 and 7 units, respectively) set in front of the robot (fig. 1). To implement some complexity in the task, we put a fixed visual target of a different colour (blue) in the middle of the table: this second object can only be foveated while, for simplicity, it cannot be touched or grasped with the hand. This "distractor" has no relations with the task: interacting with it does not increase the chance for the system to obtain rewards. In real environments the organisms are surrounded by many different objects with which they can interact in many different ways. However, not every interaction has the same importance: some actions could turn out to be the basis for more complex ones, others might not be related with new skills in the same environment, yet other ones may even result useless. Since we want to improve the versatility of artificial agents, we want to test our hypothesis in an environment that presents, although much simplified, some of the characteristics of the real world.

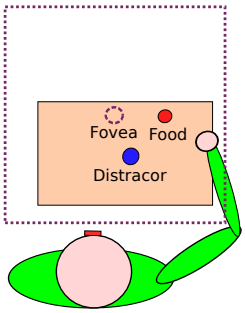

Fig. 1. Set up of the experiment: the system composed by a two dimensional arm and a moving eye (dotted square with a fovea at the centre). Food and a fixed distractor are positioned on a table in front of the robot. The task is to eat the food by bringing it to the mouth. See text for details.

The sensory system of the robot is composed of: (a) an artificial retina (a square of 14 units per size) sensible to the two different colours of the objects, encoding the position of the hand, of the food (a circle with 0.3 units diameter) and of the distractor (diameter 0.4) with respect to the centre of the visual field; (b) a "fovea", encoding whether the food or the distractor are perceived in the centre of the visual field; (c) the proprioception of the arm (composed of two segments of 4 units), encoding the angles of the two arm joints; (d) a touch sensor encoding whether the hand is in contact with the food (i.e, if the hand and the object are overlapping: collisions are not simulated). The eye moves on x and y axes with maximum step of 8 units. The two joints of the arm move within the interval [0, 180] degrees, with maximum step of 25 degrees.

Since we are focusing on cumulative learning, there is a sort of dependency between the skills that the robot can learn: the arm receives as input what the eye sees, so that learning to systematically look at the food is a prerequisite for learning to reach for it; at the same time, reaching for the food is the necessary condition for grasping it and bring it to the mouth.

2.2 Architecture and experimental conditions

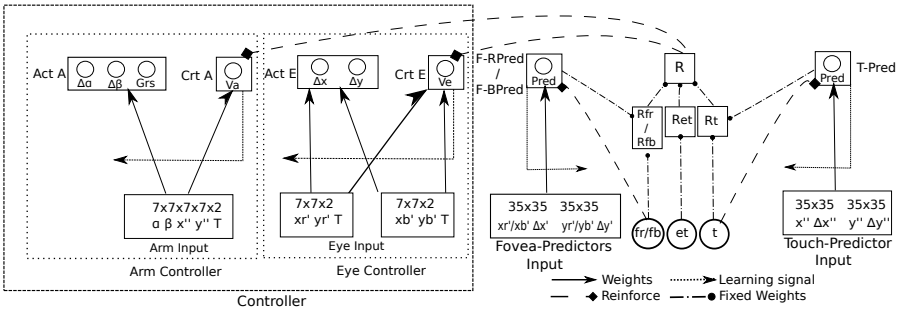

Fig. 2. The controller formed by two components (arm and eye controllers), the two fovea-predictors, the touch-predictor, and the reinforcement system. α and β are the angles of the two arm joints; x" and y" are the hand positions with respect to the fovea on the x and y axes; $\Delta\alpha$ and $\Delta\beta$ are the variations of angles as determined by the arms actor; Grs is the grasping output; Va is the evaluation of arms critic; xr', yr' and xb', yb' are the positions of food and distractor with respect to the fovea on the x and y axes; Δx and Δy are the displacements of the eye determined by the actor of the eye; Ve is the evaluation of the critic of the eye; F-RPred and F-BPred are the predictions of the fovea-predictors; T-Pred is the prediction of the touch-predictor; fr and fb are the activations of the fovea sensor for the two colours; t is the activation of the touch sensor; Rfr, Rfb and Rt are the reinforcements related to sensors activations; Ret is the reinforcement provided by eating the object; R is the total reinforcement. See text for details.

As we want to implement characteristics typical of biological organisms, we tried to build the architecture of the system (fig.2) following some constraints deriving from the known biology underlying reinforcement learning in real animals. The controller of the system reflects the modular organization of the basal-ganglia-thalamo-cortical loops [30], where the acquisition of new motor skills and the selection of motor commands take place [31]. We implemented the system as an actor-critic reinforcement learning architecture based on TD-learning because there is evidence [32] that the dorsal regions of the basal ganglia reflect the characteristics of this structure. Moreover, the reinforcement learning signal is unique for both the sub-controllers, because phasic DA signal is likely to be the same for all sensory-motor subsystems [33].

As described in sec. 1, the reinforcement signal is determined both by the extrinsic reward provided by eating the food and by the intrinsic reinforcement provided by the unpredicted activations of the fovea and the touch sensors. For this reason the system includes also three predictors, two for the fovea sensor (one for each colour of the objects) and one for the touch sensor. Each predictor is trained to predict the activation of the corresponding sensor and inhibits the part of the intrinsic reinforcement that depends on the activation of that sensor. Hence, the total reinforcement (R) driving TD-learning is:

$$R = R_e + R_{ff} + R_{fd} + R_t$$

where R_e is the extrinsic reinforcement provided by bringing the food to the mouth (with a value of 15), while R_{ff}, R_{fd} and R_t are the intrinsic reinforcements provided by the unpredicted activations of the fovea and touch sensors. For a generic sensor S, the reinforcement R_S provided by the activation of S is:

$$R_S = max[0; A_S - P_S]$$

where A_S is the binary activation $\{0; 1\}$ of sensor S and P_S is the prediction generated by the predictor of sensor S.

To test our hypothesis, we compare the described condition (called *"intrinsic"* condition), with two different conditions, where we vary the composition of the learning signal. In the *"extrinsic"* condition the reinforcement is given only by the extrinsic reinforcement of eating the food (R_e), while in the *"subtasks"* condition, the additional reinforcements provided by the activations of the sensors (R_{ff}, R_{fd} and R_t) are also "permanent", in the sense that they are not modulated by the activities of the predictors and hence do not change throughout training.

2.3 Input coding and learning

All the inputs were encoded with population codes through Gaussian radial basis functions (RBF) [34]:

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

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, their activation is 0.5).

The dimensions of the input to the two "retinas" of the eye controller are the position of the respective object (in x and y) with respect to the centre of the visual field and the activation of the touch sensor. The preferred object positions of input units are uniformly distributed on a 7x7 grid with ranges [-7; 7], which, multiplied by the binary activation of the touch sensor, form a total 7x7x2 grid. In total, the eye has two 7x7x2 grids input, one for each of the two objects.

The dimensions of the input to the arm controller are the angles of the two joints (α and β), the position of the hand (x and y) with respect to the fovea, and the activation of the touch sensor. The preferred joint angles of input units are uniformly distributed on a 7x7 grid ranging in [0; 180] whereas the preferred positions of the hand with respect to the fovea are uniformly distributed on a 7x7 grid with ranges [-7; 7]. Hence, considering the binary activation of the touch sensor, a total 7x7x7x7x2 grid input.

The input units of the eye controller are fully connected to two output units with sigmoidal activation:

$$o_j = \Phi(b_j + \sum_{i}^{N} a_i w_{ji}) \qquad \Phi(x) = \frac{1}{1 + e^{-x}}$$

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. Each output unit controls the displacement of the eye along one dimension. Each actual motor command o_j^n is generated by adding some noise to the activation of the relative output unit:

$$o_j^n = o_j + r$$

where r is a random value uniformly drawn in [-0.02; 0.02]. The resulting commands (in [0; 1]) are remapped in [-8, 8].

The arm controller has three output units. Two have sigmoidal activation, as those of the eye, with noise uniformly distributed in [-0.2; 0.2]. Each resulting motor command, remapped in [-25; 25] degrees, determines the change of one joint angle. The third output unit has binary activation $\{0; 1\}$, and controls the grasping action (the activation is determined by the sigmoidal activation of the output unit plus a random noise uniformly drawn in [-0.2; 0.2], with a threshold set to 0.5).

The evaluation of the critic of each sub-controller k (V_k) is a linear combination of the weighted sum of the respective input units.

The input units of the predictors of fovea activation are formed by two 35x35 grids, each encoding the position of the respective object with respect to the fovea along one axis and the programmed displacement of the eye along the same axis. Similarly, the input of the predictor of the touch sensor is formed by two 35x35 grids, each encoding the position of hand with respect to the food along one axis and the programmed displacement of the hand along the same axis. All preferred input are uniformly distributed in the range [-7; 7] for objects positions and [-25; 25] for displacements. The output of each predictor is a single sigmoidal unit receiving connections from all the predictor's input units.

Learning depends on the TD reinforcement learning algorithm, where the TD-error δ_k of each sub-controller k is computed as:

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

where R^t is the reinforcement at time step t, V_k^t is the evaluation 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 activation of the grasping output is slightly punished with a negative reinforcement of 0.0001.

The weight w_{ki} of input unit i of critic k is 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 0.02 for both the eye and the arm controllers.

The weights of actor k are updated as follows:

$$\Delta w_{kji} = \eta_k^a \delta_k (o_{kj}^n - o_{kj})(o_{kj}(1 - o_{kj}))a_{ki}$$

where η_k^a is the learning rate (set to 0.2 for both the eye and the arm controller), and $o_{kj}(1 - o_{kj})$ is the derivative of the sigmoid function.

Event predictors are trained through a TD-learning algorithm (for a generalization of TD-learning to general predictions, see [35]). For each predictor p, the TD-error δ_p is calculated as follows:

$$\delta_p = (A_p^t + \gamma_p O_p^t) - O_p^{t-1}$$

where A_p^t is the activation of the sensor related to predictor p at time step t, O_p^t is the output of predictor p at time step t, and γ_p is the discount factor, set to 0.7. Finally, the weights of predictor p, are updated as the ones of the critics of the two sub-controllers, with a learning rate set to 0.00008.

3 Results

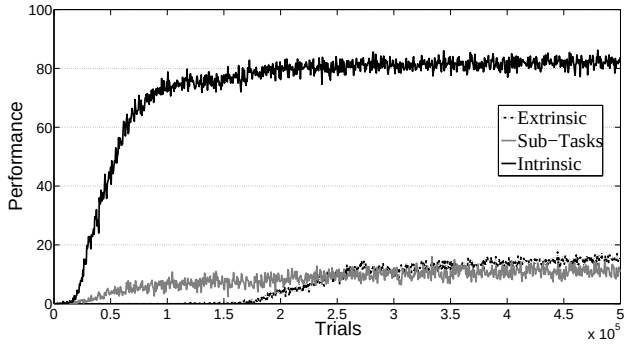

Fig. 3. Performance (percentage of test trials in which the robot eats the food) of the three experimental conditions in the task

We tested each condition on the experimental task for 500000 trials, each trials terminating when food was eaten or when it "fell" from the table (i.e. if the food is moved outside the table and not "grasped"), or after a time up of

40 steps. At the end of every trial the food, the eye centre and the hand were repositioned randomly without overlaps, with the first two always inside the table. Every 500 trials we performed 50 test trials (where learning was switched off). For each condition we ran ten replications of the experiment and here we present the average results of those replications.

Fig. 3 shows the performance on the eating task in the three experimental conditions. In the *extrinsic* condition the robot is not able to learn with satisfying results the task. This is because the final reward is too distant and infrequent to drive in a significant way the learning of the sub-tasks needed for the eating skill.

Adding permanent reinforcements for every possible interaction with the environment, as in the *sub-tasks* conditions, does not improve the performance of the system in the final task. Differently, in the *intrinsic* condition, where the activations of the sensors are reinforcing only when unpredicted, the system is able to reach high performance on the eating task (about 85%).

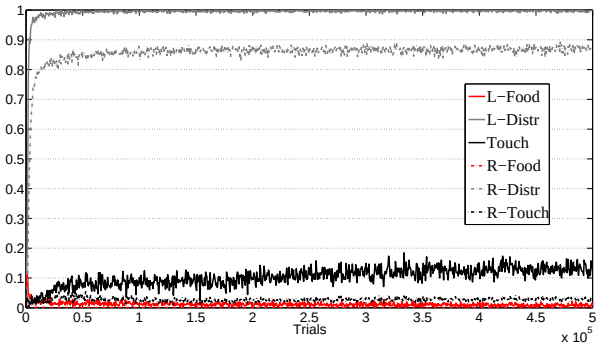

Fig. 4. Behaviour of the eye and of the arm in the *sub-tasks* condition. Average percentage of test trials in which the eye foveates the food (L Food) and the distractor (L Distr) and in which the hand touch the food (Touch); average reinforcements per step generated by the unpredicted activations of the sensors (R-Food, R-Distr and R-Touch)

To understand the reason of these results we have to look at the behaviour of the eye in the two conditions where further reinforcements are given in addition to the final one. In the *sub-tasks* condition (fig. 4), the robot starts to look at the distractor, that is simpler to find within the table. The system is stuck on this activity by the continuous reinforcements and because looking at the distractor is not related to the other skills the agent is not able to develop the capacity to look at the food, which is a prerequisite for the other skills of reaching and grasping it and in general for achieving the final goal.

On the contrary, in the *intrinsic* condition (fig. 5) the robot is able to learn the correct sequence of actions. Also in this case the system starts with looking at the fixed target, but after the predictor of the fovea sensor for the blue colour starts

to predict the perception of the distractor, that interaction is no more reinforcing. As a result, the robot can discover that also foveating the food can be reinforcing and so starts acquiring this second ability. This gives the prerequisite for the arm to learn to touch and eventually grasp the food and then to bring it to the mouth. Here the interactions with the objects are not simply reinforced, but they are implemented as IM: they are reinforcing only when they are unexpected. If we look at fig. 5, we can see that the reinforcements provided by the fovea and the touch sensor are not continuous as in the *sub-tasks* condition: they rapidly grow when the related ability is encountered and repeated, and they fade out when the motor skills are learned and their consequences became predictable. Although they turned to be no more reinforcing, the skills are still performed when they constitute the prerequisites for successive actions and for the maximization of extrinsic rewards.

Notice that as the robot learns to eat the food, the number of times it looks at the distractor increases again. Due to architectural limits, the eye is not able to track the food while the hand is grasping and moving it (the eye controller is not informed about the movements of the arm). As a result, the eye resorts to the behavior that it has previously learned, i.e. foveating the distractor. Moreover, the performance of the arm in touching the food is higher than the one of the eye in looking it: when skills are learned it is sufficient that the eye looks close to food to allow the arm to reach it.

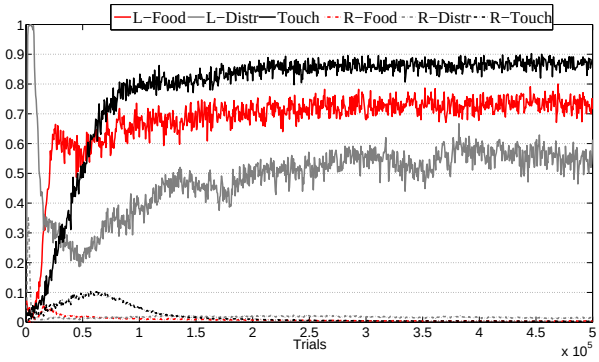

Fig. 5. Behaviour of the eye and of the arm in the *intrinsic* condition. Same data of fig. 4

We wondered if the results of the experiments are dependent on the values that we assigned to the different reinforcements: to verify this possibility, we tested the three conditions varying the value assigned to eating the food. The results (fig. 6) show that changing the value of the extrinsic reward in the learning signal does not modify the comparison between the different conditions: lowering or rising the reward for eating the food maintains the *intrinsic* condition as the best performer.

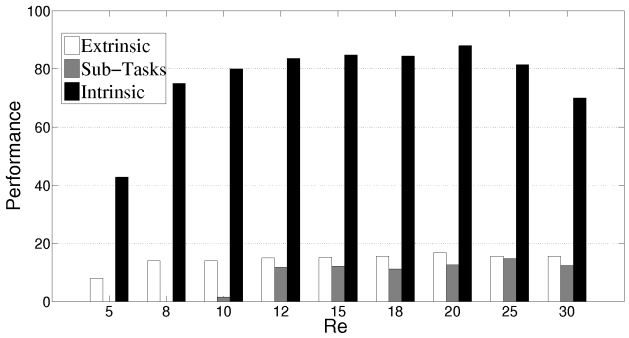

Fig. 6. Average final performance of the three conditions as a function of the value of the extrinsic reinforcement (Re) provided by eating the food. See text for details.

4 Discussion

This paper validates our hypothesis that implementing artificial agents with a learning signal that resembles the phasic activations of DA neurons of biological organism can support cumulative learning. We tested a simulated robotic agent in a simulated environment where not all the possible interactions with the world are useful for the achievement of the final goal. We varied the composition of the learning signal and we verified that only the one implementing our hypothesis was able to guide the simulated robot in the achievement of the task.

Extrinsic reinforcements by themselves are not sufficient to drive the acquisition of complex sequences of actions. Simply adding a further reinforcement for every interaction with the environment will lead the agents to get stuck in useless activities. Differently, a learning signal based both on the temporary reinforcements provided by unexpected events and by the permanent reinforcements of extrinsic rewards is able to guide the discovery of novel actions and the deployment of the acquired skills for the achievement of goals.

The nature of IM fits particularly well with the complexity of real environments. Intrinsic reinforcements are present only when they are needed: once the system has learnt to systematically generate an effect in the environment, that effect is easily predicted and for this reason it is no more reinforcing; so the agent is not stuck on the repetition of acquired actions and can move to discover novel interactions with the world so increasing its repertoire of skills.

Looking at the implementation of our hypothesis, the system still has some limits: building a complex repertoire of actions needs an architecture that is able to discover and retain different abilities. In fact, another problem related to cumulative learning is the so called catastrophic forgetting, the phenomenon by which some neural networks forget past memories when exposed to a set of new ones. A good solution to this problem is to develop hierarchical architectures [36, 37] that are able to store new skills without impairing the old ones. We designed our system in order to bypass some of the problems related to catastrophic

forgetting, but we will certainly need to move towards hierarchical structures in order to fully support cumulative learning processes.

Aknowledgements

This research was supported by the EU Project IM-CLeVeR (Intrinsically Motivated Cumulative Learning Versatile Robots), contract no. FP7-IST-IP-231722.

References

1. Baldassarre, G., Mirolli, M.: What are the key open challenges for understanding autonomous cumulative learning of skills? Autonomous Mental Development Newsletter **7**(2) (2010) 2–9
2. Wise, R.A., Rompre, P.P.: Brain dopamine and reward. Annu Rev Psychol **40** (1989) 191–225
3. Wise, R.: Dopamine, learning and motivation. Nature Reviews Neuroscience **5**(6) (2004) 483–494
4. Schultz, W.: Behavioral theories and the neurophysiology of reward. Annual Reviews of Psychology **57** (2006) 87–115
5. Berridge, K.: The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology **191**(3) (2007) 391–431
6. Romo, R., Schultz, W.: Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. Journal of Neurophysiology **63**(3) (1990) 592–606
7. Ljungberg, T., Apicella, P., Schultz, W.: Responses of monkey midbrain dopamine neurons during delayed alternation performance. Brain Research **567**(2) (1991) 337–341
8. Schultz, W., Apicella, P., Ljumberg, T.: Responses of monkey dopamine neurons to reaward and conditioned stimuli during successive steps of learning a delayed response task. Journal of Neuroscience **13** (1993) 900–913
9. Mirenowicz, J., Schultz, W.: Importance of unpredictability for reward responses in primate dopamine neurons. Journal of Neurophysiology **72**(2) (1994) 1024–1027
10. Ljungberg, T., Apicella, P., Schultz, W.: Responses of monkey dopamine neurons during learning of behavioral reactions. Journal of Neurophysiology **67**(1) (1992) 145–163
11. Schultz, W.: Predictive reward signal of dopamine neurons. Journal of Neurophysiology **80**(1) (1998) 1–27
12. Horvitz, J.C.: Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. Neuroscience **96**(4) (2000) 651–656
13. Dommett, E., Coizet, V., Blaha, C.D., Martindale, J., Lefebvre, V., Walton, N., Mayhew, J.E.W., Overton, P.G., Redgrave, P.: How visual stimuli activate dopaminergic neurons at short latency. Science **307**(5714) (2005) 1476–1479
14. Houk, J., Adams, J., Barto, A.: A model of how the basal ganglia generate and use neural signals that predict reinforcement. MIT Press, Cambridge, MA (1995)
15. Montague, P.R., Dayan, P., Sejnowski, T.J.: A framework for mesencephalic dopamine systems based on predictive hebbian learning. Journal of Neuroscience **16**(5) (1996) 1936–1947

16. Schultz, W., Dayan, P., Montague, P.R.: A neural substrate of prediction and reward. Science **275**(5306) (1997) 1593–1599
17. Sutton, R., Barto, A.: Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA (1998)
18. Redgrave, P., Prescott, T.J., Gurney, K.: Is the short-latency dopamine response too short to signal reward error? Trends in Neuroscience **22**(4) (1999) 146–151
19. Redgrave, P., Gurney, K.: The short-latency dopamine signal: a role in discovering novel actions? Nature Reviews Neuroscience **7**(12) (2006) 967–975
20. Redgrave, P., Vautrelle, N., Reynolds, J.N.J.: Functional properties of the basal ganglia's re-entrant loop architecture: selection and reinforcement. Neuroscience (2011)
21. Mirolli, M., Santucci, V., Baldassarre, G.: Phasic dopamine as a prediction error of intrinsic and extrinsic reinforcements driving both action acquisition and reward maximization: A simulated robotic study. Neural Networks (Submitted)
22. White, R.: Motivation reconsidered: the concept of competence. Psychological Review **66** (1959) 297–333
23. Berlyne, D.: Conflict, Arousal and Curiosity. McGraw Hill, New York (1960)
24. Ryan, Deci: Intrinsic and extrinsic motivations: Classic definitions and new directions. Contemporary Educational Psychology **25**(1) (2000) 54–67
25. Montgomery, K.: The role of the exploratory drive in learning. Journal of Comparative Psychology **47**(1) (1954) 60–64
26. Butler, R.A., Harlow, H.F.: Discrimination learning and learning sets to visual exploration incentives. J Gen Psychol **57**(2) (1957) 257–264
27. Kish, G.B.: Learning when the onset of illumination is used as reinforcing stimulus. Journal of Comparative and Physiological Psychology **48**(4) (1955) 261–264
28. Glow, P., Winefield, A.: Response-contingent sensory change in a causally structured environment. Learning & Behavior **6** (1978) 1–18 10.3758/BF03211996.
29. Reed, P., Mitchell, C., Nokes, T.: Intrinsic reinforcing properties of putatively neutral stimuli in an instrumental two-lever discrimination task. Animal Learning and Behavior **24** (1996) 38–45
30. Romanelli, P., Esposito, V., Schaal, D.W., Heit, G.: Somatotopy in the basal ganglia: experimental and clinical evidence for segregated sensorimotor channels. Brain Research Reviews **48**(1) (2005) 112–128
31. Graybiel, A.M.: The basal ganglia: learning new tricks and loving it. Current Opinions in Neurobiology **15**(6) (2005) 638–644
32. Joel, D., Niv, Y., Ruppin, E.: Actor-critic models of the basal ganglia: new anatomical and computational perspectives. Neural Networks **15**(4-6) (2002) 535–547
33. Schultz, W.: Getting formal with dopamine and reward. Neuron **36**(2) (2002) 241–263
34. Pouget, A., Snyder, L.H.: Computational approaches to sensorimotor transformations. Nature Neuroscience **3 Suppl** (2000) 1192–1198
35. Sutton, R., Tanner, B.: Temporal-difference networks. Advances in neural information processing systems **17** (2005) 1377–1348
36. Doya, K., Samejima, K., ichi Katagiri, K., Kawato, M.: Multiple model-based reinforcement learning. Neural Compututation **14**(6) (2002) 1347–1369
37. Caligiore, D., Mirolli, M., Parisi, D., Baldassarre, G.: A bioinspired hierarchical reinforcement learning architecture for modeling learning of multiple skills with continuous states and actions. In Johansson, B., Sahin, E., Balkenius, C., eds.: Proceedings of the Tenth International Conference on Epigenetic Robotics. (2010)