# Cumulative learning through intrinsic reinforcements

Vieri G. Santucci Gianluca Baldassarre Marco Mirolli

**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 reinforcement learning 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 by both extrinsic and intrinsic reinforcements, would be suitable to improve the implementation of cumulative learning in artificial agents. To validate our hypothesis we performed experiments with a simulated robotic system that has to learn different skills to obtain extrinsic rewards. We compare different versions of the system varying the composition of the learning signal and we show that the only system able to reach high performance in the task is the one that implements the learning signal suggested by our hypothesis.

## 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 goal for robotics and machine learning. One of the features that allows

Vieri G. Santucci Gianluca Baldassarre Marco Mirolli
Istituto di Scienze e Tecnologie della Cognizione (ISTC), Consiglio Nazionale delle Ricerche (CNR), Laboratory of Computational Embodied Neuroscience (LOCEN)
Via San Martino della Battaglia 44, 00185, Roma, Italia
{vieri.santucci, gianluca.baldassarre, marco.mirolli}@istc.cnr.it
Vieri G. Santucci
School of Computing and Mathematics, University of Plymouth
Plymouth PL4 8AA, United Kingdom

agents to achieve autonomous development and high versatility [1] is *cumulative learning*: the ability to use previously acquired skills to learn new ones, to combine sequences of actions to interact in different and more complex ways with the environment. Implementing cumulative learning in artificial agents presents many difficulties: two of the main and more general problems [2] are (a) the generation of the learning signals that can drive cumulative learning and (b) the type of architecture that can support such a process. In this work we focused on problem (a), trying to suggest a novel way to solve it.

In the computational literature a way to tackle the problem of cumulative learning has been to replace tasks-specific learning signals with new non-tasks-specific learning signals inspired by what psychologists have been calling *intrinsic motivations* (IM) [3, 4, 5]. These were introduced in the 1950s in animal psychology to explain experimental data (e.g.[6, 7]), incompatible with the classic motivational theory (e.g. [8]), showing that stimuli not related to (extrinsic) primary drives present a reinforcing value capable of conditioning instrumental responses [9, 10, 11]. Some authors focused on learning signals determined by the acquisition of knowledge by the system (e.g. [12, 13, 14, 15], while other authors used learning signals based on what the system is doing, and in particular on the acquisition of new competences (e.g. [16, 17]). Although with different solutions, the intrinsically motivated approach influenced many works (for a review see [18]) focused on the development of more versatile and autonomous systems able to acquire repertoires of skills, possibly in a cumulative fashion [19].

Our idea (first presented in a preliminary version in [20]) is that if we want to solve the problem of which learning signal can be suitable for the implementation of cumulative learning, a good solution is to look at biological organisms: the characteristics that we are trying to implement in artificial systems are typical of biological agents, that are able to cumulatively (and autonomously) learn new skills and to combine them together to optimise their survival chances. What we suggest is to look at those data that can explain how these features are developed in biology, focusing on those signals that can support cumulative learning.

The neuromodulator dopamine (DA) has long been recognized to play a fundamental role in motivational control and reinforcement learning processes [21, 22, 23]. In particular, phasic DA activations have been related to the presentation of unpredicted rewards [24, 25, 26, 27] but also to other phasic, non reward-related, unexpected stimuli [28, 29, 30, 31]. These data led to the formulation of two main hypotheses on the functional role of the DA signal. One hypothesis [32, 33] looks at the similarities of DA activations with the temporal-difference (TD) error of computational reinforcement learning [34], 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 [35, 36] 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 [37], 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 IM: unpredicted events are intrinsic reinforcers that drive the same reinforcement learning processes as extrinsic reinforcers.

In this work we propose that providing artificial agents with a learning signal that resembles the characteristic of the phasic DA signal, determined both by extrinsic and intrinsic reinforcements, would be an advancement in the development of more autonomous and versatile systems. Moving from biology to artificial agents, we can identify extrinsic reinforcements with those determined by the achievement of the tasks decided by the researchers, whereas intrinsic reinforcements are identified with those determined by a category of more general events, such as the unexpected activations of the sensors of the robot, determined by its interactions with the environment. Similarly to what happens in biological systems [38], we believe that intrinsic reinforcements can play a key role in determining a proper signal for the implementation of the cumulative learning of skills and for the acquisition of complex behaviours that would not be learned simply with extrinsic reinforcements.

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 and the simulated robot*

The system is a simulated kinematic robot composed of a fixed head with a "mouth", a moving eye, and a two-degree-of-freedom kinematic arm with a hand that can "grasp" objects. 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). In the middle of the table we add a visual "distractor" of a different colour (blue) that can only be foveated while, for simplicity, it cannot be touched or grasped: interacting with this second object does not increase the chance for the system to achieve the final goal.

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, while others 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: for this reason we put a "distractor" that has no relations with the task, in order to provide a set up where not all the possible interactions with the environment are related to the main task of the experiment.

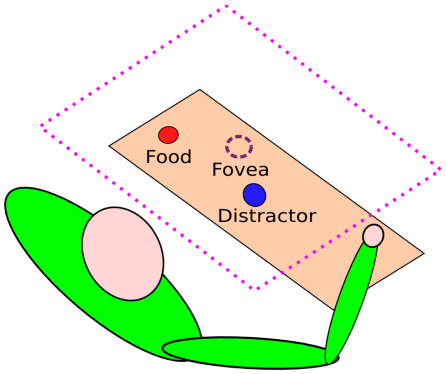


**Fig. 1** Set up of the experiment: the system composed by a two dimensional arm and a moving eye (dotted square with a fovea at the centre). Food and a fixed distractor are positioned on a table in front of the robot. The task consists in eating the food by bringing it to the mouth. See text for details.

Since we are focusing on cumulative learning, there is a 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 necessary for grasping it and bringing it to the mouth.

The sensory system of the robot is composed of: (a) an artificial retina (a square of 14 units per size; note that this implies that at the beginning of each trial the whole table is always within the eye image) 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: for simplicity collisions are not simulated). The eye moves along the x and y axes with a maximum step of 8 units. The two joints of the arm move within the interval [0, 180] degrees, with maximum step of 25 degrees.

## *2.2 Architecture and experimental conditions*

As we are proposing to look at biological organisms to improve the implementation of cumulative learning in artificial agents, 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 [39], where the acquisition of new motor skills and the selection of motor commands take place [40]. We implemented the system as an actor-critic reinforcement learning architecture based on TD-learning because there is evidence [41] that the dorsal regions of the basal ganglia reflect the characteristics of this structure and because this solution has also some appealing theoretical properties from the machine learning point of view [34, 42]. Moreover, the reinforcement learning signal is unique for both the sub-controllers, because the phasic DA signal is likely to be the same for all sensory-motor subsystems [43]: this simplifies the computation of the learning signal and allows to reinforce some actions also if they determine the activations of sensors not directly connected to the effectors that generated those effects.

As described in sec. 1, the reinforcement signal is determined by both the extrinsic rewards provided by eating the food and by the intrinsic reinforcements provided by the unpredicted activations of the fovea and the touch sensors. To implement the intrinsic reinforcements, the system includes also three predictors, two for the fovea sensor (one for each colour of the ob-

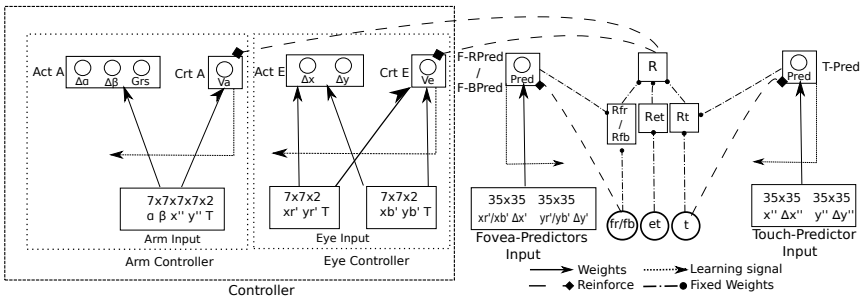


**Fig. 2** The controller formed by two components (arm and eye controllers), the two predictors of the fovea sensor (for simplicity, in this schema they are presented as a single structure), the predictor of the touch sensor, and the reinforcement system. $\alpha$ and $\beta$ are the angles of the two 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 the critic of the arm; xr', yr' and xb', yb' are the positions of food and distractor with respect to the fovea on the x and y axes; $\Delta$x and $\Delta$y 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 food; R is the total reinforcement. See text for details.

jects) 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 unexpected 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 sensor caused by the food ($R_{ff}$), or by the "distractor" ($R_{fd}$) and the unpredicted activations of the touch sensor ($R_t$) caused by the food. In particular, 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. In this way we use only the positive reinforcements generated when the activation of $A_S$ is not fully predicted by $P_S$.

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 reinforcements provided by eating the food ($R_e$). This condition is useful to test if extrinsic reinforcements by themselves are able to drive the cumulative learning of skills. In the *sub-tasks* 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. With this condition we want to test if the temporary nature of intrinsic reinforcement is necessary to facilitate learning.

### *2.3 Input coding*

All the inputs are encoded with population coding [44] through Gaussian radial basis functions (RBF) [45]:

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

where $a_i$ is the activation of unit $i$, $c_d$ is the input value on dimension $d$, $c_{id}$ is the preferred value of unit $i$ with respect to dimension $d$, and $\sigma_d^2$ 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. We add the status of the touch sensor because for computational limits the eye is not able to follow the food when it is moved by the hand: providing this information we can separate the two situation (object not grasped from object grasped) and prevent the controller of the eye from losing the ability of looking at the objects. 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 an input formed by two 7x7x2 grids, one for each of the two objects.

The dimensions of the input to the arm controller are the angles of the two joints ($\alpha$ and $\beta$), 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 two dimensions (7x7) ranging in [0; 180] whereas the preferred positions of the hand with respect to the fovea are uniformly distributed on other two dimensions (7x7) with ranges [-7; 7]. Hence, considering the binary activation of the touch sensor, the input is formed by a total 7x7x7x7x2 grid.

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

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

where $M$ is the total number of input units, $w_{ji}$ is the weight of the connection linking input unit $i$ to output unit $j$ and $b_j$ is the bias of output unit $j$. Each actual motor command $o_j^n$ is generated by adding some noise to the activation of the relative output unit:

$$o_j^n = o_j + n$$

where $n$ is a random value uniformly drawn in [-0.02; 0.02]. The resulting commands (in [0; 1]) are remapped in [-8, 8] and control the displacement of the eye along the two dimensions.

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 activation of the grasping output is slightly punished with a negative reinforcement of 0.0001 to avoid that the system performs grasping also when it is not on the target.

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 one 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 one encoding the position of the hand with respect to the food along one axis and the programmed displacement of the hand along the same axis. Preferred inputs 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 units of the predictor.

## *2.4 Learning*

Learning depends on the TD reinforcement learning algorithm [34] that was introduced to solve the temporal credit assignment problem, i.e. the problem of learning which of many actions contributed to the achievement of reward: the TD learning solves the problem with the use of predictions and in particular with the use of the TD-error as the learning signal reinforcing all those actions that lead the system closer to rewards. The TD-error $\delta_k$ 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 $\gamma_k$ is the discount factor, set to 0.9 for both the eye and the arm controllers.

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 $\eta_k^c$ 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 $\eta_k^a$ 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.

Predictors are trained through a TD-learning algorithm (for a generalization of TD-learning to general predictions, see [46]). We decided to use TD-learning neural networks to implement the predictors because it is difficult to built predictors able to perfectly anticipate the activations of the sensors: a TD neural network solves the problem because it starts to anticipate the activations earlier than a one-step predictor.

For each predictor $p$, the TD-error $\delta_p$ 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 $\gamma_p$ is the discount factor, set to 0.7 for each predictor. Finally, the weights of the predictors are updated as those of the critics of the two sub-controllers, with a learning rate set to 0.00008 for each predictor.

## 3 Results

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

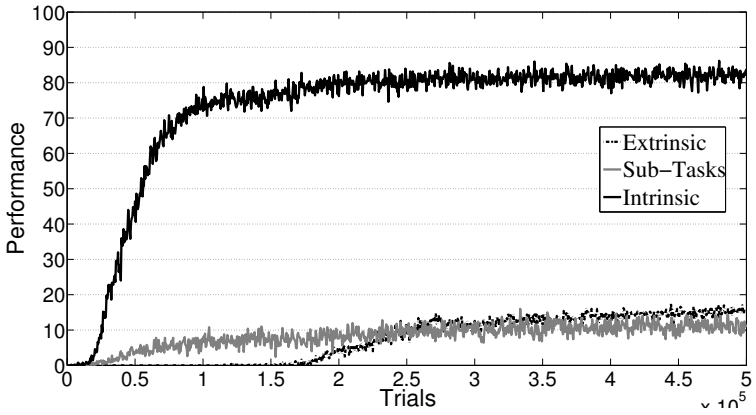


**Fig. 3** Performance (percentage of test trials in which the robot eats the food) in the three experimental conditions.

Fig. 3 shows the performance in the task of the three experimental conditions. In the *extrinsic* condition the robot is not able to learn to eat reliably. Adding permanent reinforcements for every possible interaction with the environment, as in the *sub-tasks* condition, 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 in the eating task (about 85%).

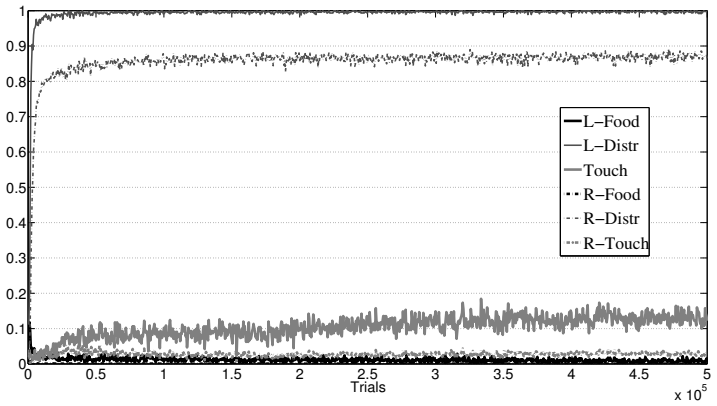


**Fig. 4** Behaviour of the eye and of the arm in the *sub-tasks* condition. Average percentage of test trials in which the eye foveates the food (L-Food) and the distractor (L-Distr) and in which the hand touches the food (Touch); average reinforcements per step generated by the unpredicted activations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

It is quite easy to understand why in the *extrinsic* condition the system is not able to achieve the final goal: the only reinforcement provided by the final reward is too distant and infrequent to drive the learning of the sub-tasks needed for bringing the food into the mouth. Although the TD algorythm is built to solve the credit assignment problem it is difficult to trace back few rewards provided by a complex sequence of different actions.

It is more interesting to analyse the results of the other two conditions where further reinforcements are given in addition to the final one. To understand the reason of these results we have to look at the behaviour of the eye. In the *sub-tasks* condition (fig. 4), the robot starts to look at the distractor, which is simpler to find within the table. Because of the permanent reinforcements provided by the activation of the fovea sensor the system is stuck on this activity, but looking at the distractor is not related to the other skills so the agent is not able to develop the capacity to look at the food, which is a prerequisite for the other abilities (reaching and grasping the food) and for the achievement of 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 sensory event 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, that is the prerequisite for the arm to learn to touch and eventually grasp the food and then to bring it to the mouth.

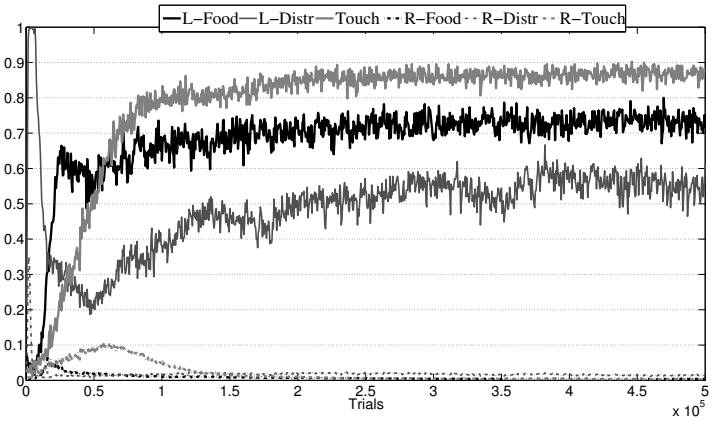


**Fig. 5** Behaviour of the eye and of the arm in the *intrinsic* condition. Same data as in fig. 4.

In the *intrinsic* condition the activations of the sensors determined by the interactions with the objects 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 sensors are not continuous as in the *sub-tasks* condition: they rapidly grow when the related ability is encountered and repeated, and they fade away when the motor skills are learned and their consequences become predictable. Although those skills do not directly generate more reinforcements, they are still performed when they constitute the prerequisites for successive actions that can provide new reinforcements and for the maximization of extrinsic rewards.

Notice (fig. 5) 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 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 at it: when skills are learned it is sufficient that the eye looks close to food to allow the arm to reach it.

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.

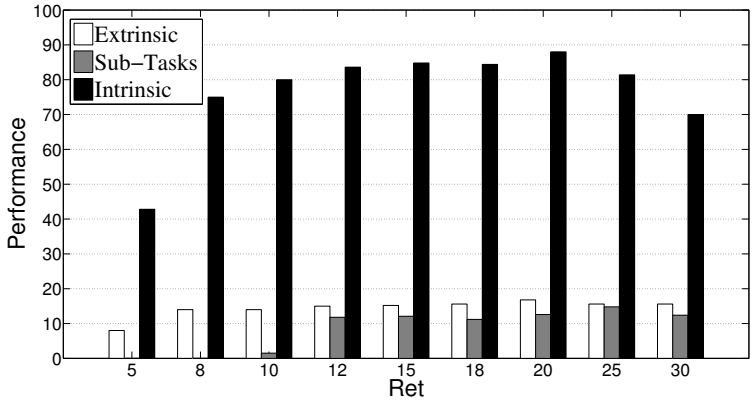


**Fig. 6** Average final performance of the three conditions as a function of the value of the extrinsic reinforcement (Re) provided by eating the food. See text for details.

## 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 and cumulative learning. Intrinsic reinforcements are present only when they are needed: when the system discovers a new possible way to interact with the environment, the consequences of its actions provide high reinforcement; once the system has learnt to systematically generate an effect (after some repetitions of the same actions), that effect can be predicted and for this reason it is no more reinforcing; the system then is not stuck on the repetition of the same actions and can move to different activities. In this way intrinsic reinforcement are able to guide agents in the discovery of novel interactions with the environment, increasing their repertoire of skills. Moreover, such a learning signal can be useful to develop more autonomous agents: IM are able to push systems to learn every possible interactions with the environment just because of the novelty of those interactions, also if those new skills are not immediately related to the fitness of the system [47, 38]. These skills can then be deployed in the appropriate situations exploiting the reinforcing value of extrinsic reinforcements.

Looking at the implementation of our hypothesis, the system still has some limits. Schmidhuber [12] underlined how using the prediction error as an intrinsic reinforcement can generate problems if the environment is unpredictable or the system has limited learning capabilities: in such cases, the reinforcement would never decrease and the system would get stuck, trying to reproduce outcomes with unpredictable consequences. To avoid this problem, he proposed the progress in predictions error as a better intrinsic reinforcement. However, we believe that this hypothesis does not reflect the biology underlying IM and we built our system using the simple prediction error to implement intrinsic reinforcements.

Another limit is connected to the second problem related to the implementation of cumulative learning (that we decided not to tackle in this work),

the architectural problem: 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 *catastrophic forgetting*, the phenomenon by which neural networks forget past experiences when exposed to new ones. A good solution to this problem is to develop hierarchical architectures (e.g. [48, 49]. See [16] for a review) 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. Moreover, we believe that within the framework of hierarchical organization of actions, we can provide a reinforcement signal that, without loosing the inspiration provided by biological organisms, can cope with the problem raised by Schmidhuber: intrinsic reinforcements can be determined by the learning progress in skills acquisition [50]. If nothing can be learnt, there will be no learning progress and the system will move away looking for new skills to acquire.

## Aknowledgements

This research was supported by the European Community 7th Framework Programme (FP7/2007-2013), "Challenge 2 - Cognitive Systems, Interaction, Robotics", grant agreement No. ICT-IP-231722, project "IM-CLeVeR - Intrinsically Motivated Cumulative Learning Versatile Robots".

## References

1. Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., Thelen, E.: Artificial intelligence. autonomous mental development by robots and animals. Science **291**(5504) (Jan 2001) 599–600
2. 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
3. White, R.: Motivation reconsidered: the concept of competence. Psychological Review **66** (1959) 297–333
4. Berlyne, D.: Conflict, Arousal and Curiosity. McGraw Hill, New York (1960)
5. Ryan, Deci: Intrinsic and extrinsic motivations: Classic definitions and new directions. Contemporary Educational Psychology **25**(1) (2000) 54–67
6. Montgomery, K.: The role of the exploratory drive in learning. Journal of Comparative Psychology **47**(1) (1954) 60–64
7. Butler, R.A., Harlow, H.F.: Discrimination learning and learning sets to visual exploration incentives. J Gen Psychol **57**(2) (1957) 257–264
8. Hull, C.L.: Principles of behavior. Appleton-century-crofts (1943)
9. 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

10. Glow, P., Winefield, A.: Response-contingent sensory change in a causally structured environment. Learning & Behavior **6** (1978) 1–18
11. 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
12. Schmidhuber, J.: Curious model-building control system. In: Proceedings of International Joint Conference on Neural Networks. Volume 2., IEEE, Singapore (1991b) 1458–1463
13. Huang, X., Weng, J.: Novelty and reinforcement learning in the value system of developmental robots. In Prince, C., ans Y. Marom, Y.D., Kozima, H., Balkenius, K., eds.: Proceedings of the Second International Workshop on Epigenetic Robotics. Volume 94., Lund University (2002) 47–55
14. Oudeyer, P., Kaplan, F., Hafner, V.: Intrinsic motivation system for autonomous mental development. In: IEEE Transactions on Evolutionary Computation. Volume 11. (2007) 703–713
15. Baranes, A., Oudeyer, P.Y.: Intrinsically motivated goal exloration for active motor learning in robots: a case study. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan (2010)
16. Barto, A., Singh, S., Chantanez, N.: Intrinsically motivated learning of hierarchical collections of skills. In: Proceedings of the Third International Conference on Developmental Learning (ICDL). (2004) 112–119
17. Schembri, M., Mirolli, M., Baldassarre, G.: 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 (2007) E1–6
18. Oudeyer, P.Y., Kaplan, F.: What is intrinsic motivation? a typology of computational approaches. Frontiers in Neurorobotics **1**(1) (2007) 1–14
19. Baldassarre, G., Mirolli, M.: Intrinsically Motivated Learning in Natural and Artificial Systems. Springer, Berlin (in press)
20. Santucci, V., Baldassarre, G., Mirolli, M.: Biological cumulative learning through intrinsic motivation: A simulated robotic study on the development of visually-guided reaching. In Johansson, B., Sahin, E., Balkenius, C., eds.: Proceedings of the Tenth International Conference on Epigenetic Robotics, Lund University Cognitive Studies, Lund (2010) 121–128
21. Wise, R.: Dopamine, learning and motivation. Nature Reviews Neuroscience **5**(6) (2004) 483–494
22. Schultz, W.: Behavioral theories and the neurophysiology of reward. Annual Reviews of Psychology **57** (2006) 87–115
23. Berridge, K.: The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology **191**(3) (2007) 391–431
24. 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
25. Ljungberg, T., Apicella, P., Schultz, W.: Responses of monkey midbrain dopamine neurons during delayed alternation performance. Brain Research **567**(2) (1991) 337–341
26. 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
27. Mirenowicz, J., Schultz, W.: Importance of unpredictability for reward responses in primate dopamine neurons. Journal of Neurophysiology **72**(2) (1994) 1024–1027
28. Ljungberg, T., Apicella, P., Schultz, W.: Responses of monkey dopamine neurons during learning of behavioral reactions. Journal of Neurophysiology **67**(1) (1992) 145–163

29. Schultz, W.: Predictive reward signal of dopamine neurons. Journal of Neurophysiology **80**(1) (1998) 1–27
30. Horvitz, J.C.: Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. Neuroscience **96**(4) (2000) 651–656
31. 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
32. Houk, J., Adams, J., Barto, A.: A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Models of Information Processing in the Basal Ganglia. MIT Press, Cambridge, MA (1995) 249–270
33. Schultz, W., Dayan, P., Montague, P.R.: A neural substrate of prediction and reward. Science **275**(5306) (1997) 1593–1599
34. Sutton, R., Barto, A.: Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA (1998)
35. Redgrave, P., Gurney, K.: The short-latency dopamine signal: a role in discovering novel actions? Nature Reviews Neuroscience **7**(12) (2006) 967–975
36. Redgrave, P., Vautrelle, N., Reynolds, J.N.J.: Functional properties of the basal ganglia's re-entrant loop architecture: selection and reinforcement. Neuroscience (2011)
37. 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)
38. Baldassarre, G.: What are intrinsic motivations? a biological perspective. In Cangelosi, A., Triesch, J., Fasel, I., Rohlfing, K., Nori, F., Oudeyer, P.Y., Schlesinger, M., Nagai, Y., eds.: Proceedings of the International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob-2011). IEEE, Piscataway, NJ (2011) E1–8
39. 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
40. Graybiel, A.M.: The basal ganglia: learning new tricks and loving it. Current Opinions in Neurobiology **15**(6) (2005) 638–644
41. 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
42. Sutton, R., McAllester, D., Singh, S., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. Advances in neural information processing systems **12**(22) (2000)
43. Schultz, W.: Getting formal with dopamine and reward. Neuron **36**(2) (2002) 241–263
44. Pouget, A., Snyder, L.H.: Computational approaches to sensorimotor transformations. Nature Neuroscience **3 Suppl** (2000) 1192–1198
45. Buhmann, M.: Radial Basis Functions. Cambridge University Press, New York, NY, USA (2003)
46. Sutton, R., Tanner, B.: Temporal-difference networks. Advances in neural information processing systems **17** (2005) 1377–1348
47. Singh, S., Lewis, R., Barto, A., Sorg, J.: Intrinsically motivated reinforcement learning: An evolutionary perspective. IEEE Transactions on Autonomous Mental Development **2**(2) (2010) 70–82
48. Doya, K., Samejima, K., Katagiri, K., Kawato, M.: Multiple model-based reinforcement learning. Neural Compututation **14**(6) (2002) 1347–1369
49. 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) 27–34
50. Schembri, M., Mirolli, M., Baldassarre, G.: Evolving childhood's length and learning parameters in an intrinsically motivated reinforcement learning robot. In Berthouze, L., Dhristiopher, G., Littman, M., Kozima, H., Balkenius, C., eds.: Proceedings of the Seventh International Conference on Epigenetic Robotics, Lund University Cognitive Studies, Lund (2007b) 141–148