Mirolli Marco, Santucci Vieri, Baldassarre Gianluca (2013): Phasic dopamine as a prediction error of intrinsic and extrinsic reinforcements driving both action acquisition and reward maximization: A simulated robotic study, Neural Networks, 39: 40-51
doi: dx.doi.org/10.1016/j.neunet.2012.12.012

Phasic dopamine as a prediction error of intrinsic and extrinsic reinforcements driving both action acquisition and reward maximization: A simulated robotic study

Marco Mirolli[a,*], Vieri G. Santucci[a,b], Gianluca Baldassarre[a]

[a] *Istituto di Scienze e Tecnologie della Cognizione (ISTC), CNR*
Via San Martino della Battaglia 44, 00185, Roma, Italia
[b] *School of Computing and Mathematics, University of Plymouth*
Plymouth PL4 8AA, United Kingdom

Abstract

An important issue of recent neuroscientific research is to understand the functional role of the phasic release of dopamine in the striatum, and in particular its relation to reinforcement learning. The literature is split between two alternative hypotheses: one considers phasic dopamine as a reward prediction error similar to the computational TD-error, whose function is to guide an animal to maximize future rewards; the other holds that phasic dopamine is a sensory prediction error signal that lets the animal discover and acquire novel actions. In this paper we propose an original hypothesis that integrates these two contrasting positions: according to our view phasic dopamine represents a TD-like reinforcement prediction error learning signal determined by both unexpected changes in the environment (temporary, intrinsic reinforcements) and biological rewards (permanent, extrinsic reinforcements). Accordingly, dopamine plays the functional role of driving both the discovery and acquisition of novel actions and the maximization of future rewards. To validate our hypothesis we perform a series of experiments with a simulated robotic system that has to learn different skills in order to get rewards. We compare different versions of the system in which we vary the composition of the learning signal. The results show that only the system

*Corresponding author. Tel.: +39-06-44595231; fax:+39-06-44595243
Email addresses: marco.mirolli@istc.cnr.it (Marco Mirolli), vieri.santucci@istc.cnr.it (Vieri G. Santucci), gianluca.baldassarre@istc.cnr.it (Gianluca Baldassarre)

reinforced by both extrinsic and intrinsic reinforcements is able to reach high performance in sufficiently complex conditions.

Keywords: Phasic dopamine, reinforcement learning, intrinsic motivation, TD learning, actor critic, computational model

---

1. Introduction

The neuromodulator dopamine (DA) has long been recognized to play a fundamental role in motivational control and reinforcement learning processes (Wise and Rompre, 1989; Robbins and Everitt, 1992; Wise, 2004; Schultz, 2006; Berridge, 2007). The main sources of dopamine in the brain are the dopaminergic neurons of the substantia nigra pars compacta (SNc) and the Ventral Tegmental Area (VTA), which release dopamine in a number of cortical and subcortical areas, including the pre-frontal cortex, the striatum, the hippocampus, and the amygdala (Bjorklund and Dunnett, 2007). Two modes of dopamine release have been identified: a tonic mode, in which dopaminergic neurons maintain a steady activation for prolonged periods of time, and a phasic mode, in which the firing rates of dopaminergic neurons sharply increase for 100-500 ms (Grace et al., 2007; Schultz, 2007). An important issue of recent neuroscientific research on dopamine is to understand the functional role of the phasic release of dopamine in the striatum, and in particular its relation to reinforcement learning.

1.1. Dopamine as a reward prediction error

Single neurons recording have clearly demonstrated that most dopamine neurons are activated by the rewarding characteristics of somatosensory, visual, and auditory stimuli (Schultz, 1998). In particular, most dopaminergic neurons show phasic activations in response to unpredicted rewards (Romo and Schultz, 1990). If the reward is preceded by a conditioned stimulus that reliably predict it, activations of dopaminergic neurons do not occur at the time of reward, but at the time of the (unpredicted) reward-predicting stimulus (Ljungberg et al., 1992; Schultz et al., 1993). Furthermore, dopamine neurones are phasically depressed when a predicted reward, or even a predicted reward-predicting stimulus, is omitted (Ljungberg et al., 1991; Schultz et al., 1993).

These characteristics of the phasic activation of dopamine neurons closely match the properties of the *Temporal-Difference* (TD) error postulated by

the computational theory of *Reinforcement Learning* (Barto et al., 1983; Sutton, 1988; Sutton and Barto, 1998). The TD-error (δ) is an error in the prediction of future rewards calculated on the basis of the reward itself (R) and the difference in two consecutive predictions (P):

$$\delta^t = R^t + \gamma P^t - P^{t-1}$$

where γ (ranging in $[0, 1]$) is a discount factor.

The TD error has been introduced as a learning signal that can drive an agent to learn to maximize the sum of acquired rewards. In particular, the TD learning algorithm is able to solve the problem of *temporal credit assignment*. An agent that receives rewards only as a result of a sequence of actions must learn which are the specific actions that contribute to the achievement of the reward. TD learning solves this problem through the use of predictions: using the TD error as the learning signal instead of the simple reward, all those actions that bring the agent closer to the reward (i.e. in states in which the prediction of discounted future rewards is higher) will be reinforced.

The recognition that phasic dopamine behaves like the TD error signal led to the hypothesis that phasic dopamine plays in real animals the same functional role that the TD error signal plays in artificial agents: according to this hypothesis dopamine is a reward prediction error learning signal that drives the agent in learning to deploy its actions in order to maximize rewards (Houk et al., 1995; Schultz et al., 1997). In accordance with this hypothesis, dopamine is known to modulate the plasticity of cortico-striatal synapses (Reynolds et al., 2001; Reynolds and Wickens, 2002; Calabresi et al., 2007; Wickens, 2009), and dopamine release in the striatum has been recently shown to be both necessary and sufficient for appetitive instrumental conditioning (Robinson et al., 2006; Zweifel et al., 2009). The reward prediction error hypothesis of phasic dopamine has so far received a large amount of empirical support (e.g. Hollerman and Schultz, 1998; Waelti et al., 2001; Tobler et al., 2005; Bayer and Glimcher, 2005; Daw et al., 2005; Fiorillo et al., 2008), and is currently a widely accepted tenet of contemporary neuroscience (e.g. Schultz, 2002; Suri, 2002; Montague et al., 2004; Ungless, 2004; Wise, 2004; Sugrue et al., 2005; Salzman et al., 2005; Frank, 2005; Doya, 2007; Graybiel, 2008; Glimcher, 2011).

However, the reward prediction error hypothesis has an important limit: it ignores the well known fact that phasic DA is triggered not only by reward-related stimuli, but also by other phasic, unexpected stimuli (Chiodo et al.,

1980; Steinfels et al., 1983; Strecker and Jacobs, 1985; Ljungberg et al., 1992; Horvitz et al., 1997; Schultz, 1998; Horvitz, 2000; Dommett et al., 2005). Since these activations occur in presence of stimuli that have never been associated with reward, it is not clear how the dopamine-as-TD-error hypothesis might account for them.

1.2. Novelty bonuses

A possible explanation of the dopaminergic responses to unexpected events within the computational reinforcement learning framework has been proposed by Kakade and Dayan (2002), who linked those responses to the problem of exploration (see also Fellous and Suri, 2003). A reinforcement learning agent must not focus on what it has already learned; rather, it needs to keep exploring its environment so to discover new, possibly more efficient, ways to get rewards. In the reinforcement learning literature a possible way to do this is by making reinforcements (Sutton, 1990; Dayan and Sejnowski, 1996), or reinforcement predictions (Ng et al., 1999), depend not only on bare rewards but also on other signals, called *bonuses*. Hence, according to Kakade and Dayan, the dopaminergic responses to unexpected events might be explained by assuming that animals are reinforced not only by biological rewards but also by the novelty of perceived states: such novelty bonuses would have the function of increasing the animal's tendency to explore, thus possibly improving the maximization of rewards.

The exploration bonuses hypothesis presents two problems: first, bonuses are given as a function of the novelty of the perceived *states*, whereas phasic dopamine activations have been recorded in response to *unexpected events* (i.e. unpredicted changes of state), like the switching on of a light (irrespective of whether the light is novel or familiar); second, according to this proposal, the adaptive function of novelty bonuses is a general increase in exploration, whereas there is ample evidence that unpredicted events can be used as reinforcers for learning new instrumental actions (see, e.g. Kish, 1955; Williams and Lowe, 1972; Glow and Winefield, 1978; Reed et al., 1996, see also Fiore et al., 2008 for a computational model. For a more detailed discussion on these points, see section 5 below).

1.3. Dopamine as a sensory prediction error

Redgrave and colleagues have long been criticizing the reward prediction error hypothesis of phasic dopamine (Redgrave et al., 1999) and have recently proposed an interesting alternative hypothesis (Redgrave and Gurney, 2006;

Redgrave et al., 2008, 2011, 2012). This hypothesis distinguishes between two separate sub-processes underlying instrumental conditioning: 1) action discovery and learning (i.e. learning which changes in the environment are caused by the animal and which are the sequences of movements that systematically produce those changes) and 2) learning which action to deploy in a given context so to maximize the acquisition of biological rewards. Most computational models of reinforcement learning, in particular those on which the dopamine as reward prediction error hypothesis is based, assume that the system has already a repertoire of actions (thus ignoring problem 1) and are focused on problem 2.

According to Redgrave and colleagues, the phasic dopaminergic signal is not suitable for solving problem 2 (reward maximization) for at least two reasons: first, it is triggered also by unexpected events not related to rewards; second, its latency is too short for the signal to encode the biological value of the detected event, as required by the reward-prediction error hypothesis (in particular, the latency is shorter than that of saccadic eye movements, meaning that dopamine is released before the animal has the time to turn and see the value of the appeared stimulus). On the contrary, they propose that the dopaminergic signal is ideal for solving problem 1, that is action discovery and acquisition: a pre-saccadic signal is what is needed for reinforcing those actions that have been produced *just before* the unexpected event and that might have contributed to cause it (Redgrave and Gurney, 2006; Redgrave et al., 2008). Hence, according to Redgrave and colleagues phasic dopamine is a *sensory prediction error* signal that drives action discovery and acquisition, *rather than* a reward prediction error driving reward maximization. According to this hypothesis, reward maximization is not due to the reinforcement of cortico-striatal connections in the basal ganglia, but to the reward-related modulation of stimuli representations in the sensory areas that send input to the striatum: it is this modulation of reward-related stimuli, due to yet-unknown dopamine-independent mechanisms, that can favor the selection of reward-maximizing behaviors (Redgrave et al., 2011, 2012).

We consider the distinction between the two sub-problems of instrumental conditioning very useful, and the arguments according to which phasic dopamine is particularly well suited for solving the problem of action discovery and acquisition as compelling. However, the arguments related to the second problem, according to which learning how to deploy actions for maximizing rewards does not depend on dopamine but on stimulus modu-

lation, suffer of two important flaws. First, the main argument against the reward-prediction-error hypothesis, according to which dopamine is too fast for encoding stimulus value, is just contradicted by facts: phasic dopamine has been repeatedly and consistently shown to behave like a reward prediction error, encoding both the value and the probability of predicted rewards (e.g. Morris et al., 2004; Tobler et al., 2005; Bayer and Glimcher, 2005; Fiorillo et al., 2008). By fostering new empirical research, the argumentations of Redgrave and colleagues can help in discovering *how* this is possible (e.g. May et al., 2009), but cannot disprove *that* it is true. Second, the mechanism proposed by Redgrave and colleagues for driving reward maximization, i.e. the modulation of stimulus representation by reward, is neither sufficient nor necessary to do the job. It is not sufficient because stimulus modulation may at most help the animal to focus its attention to the stimuli that are related to reward, but it cannot, by itself, tell the animal which action to perform on those stimuli: in order to maximize reward in instrumental tasks changing representations of stimuli is not enough; you need to change the probability of performing a specific action given a specific stimulus (hence, if action selection is performed in the striato-cortical loops, as Redgrave and colleagues suggest, to change cortico-striatal synapses). Moreover, the modulation of stimulus representation is not even logically necessary to maximize future rewards since the mechanism suggested for action discovery and learning, i.e. dopamine-dependent synaptic plasticity in cortico-striatal synapses, is all that is needed also for reward maximization. If one accepts, as empirical research suggests and as Redgrave and colleagues do, that (a) dopamine do reinforces actions, and (b) dopamine never habituates when rewards are involved, rewards maximization follows. Indeed, the large amount of evidence regarding the similarity between phasic dopamine and the TD-error signal demonstrates just this: phasic dopamine is the ideal learning signal for learning to maximize future rewards.

1.4. Summary and overview

In summary, the neuroscientific literature on the functional role of phasic dopamine is split between two main hypotheses. According to the predominant view, phasic dopamine is a reward prediction error learning signal whose function is to train an animal to maximize future rewards. On this view, the triggering of phasic dopamine by unexpected events is either ignored or treated as novelty bonuses with the function of fostering exploration. According to the second view, phasic dopamine is a sensory prediction error learning

signal whose function is to let an animal discover which events it can cause and how (i.e. to drive action acquisition). On this view, learning how to deploy acquired actions in order to maximize rewards depends on processes that do not happen in the striatum and do not depend on dopamine.

In this paper we propose a new hypothesis on the adaptive function of phasic dopamine which integrates these two opposing positions (section 2). We also validate our hypothesis through a series of experiments performed on a simulated robotic system that have to autonomously acquire a series of skills in order to maximize its rewards (sections 3 and 4). In particular, we compare the performance of the system with different compositions of the learning signal and show that the system that implements our hypothesis is the only one that is able to learn to maximize rewards in sufficiently complex conditions. We conclude (section 5) by discussing our hypothesis with respect to both the neuroscientific and the computational literature on reinforcement learning.

2. Dopamine reconciled: reinforcement prediction error for action acquisition and reward maximization

Our hypothesis is that phasic dopamine represents a *reinforcement* prediction error learning signal analogous to the computational TD error, in a system where both biological rewards and unexpected changes in the environment act as reinforcers. The function of such a signal is to drive both the discovery and acquisition of novel actions and the learning of how to deploy actions in order to maximize future rewards. Phasic dopamine is able to play both roles just because it is triggered by the aforementioned two different kinds of reinforcers. In particular, unexpected events constitute "temporary" reinforcers whose function is driving action discovery and acquisition, whereas biological rewards are "permanent" reinforcers whose principal function is to drive reward maximization.

The reinforcements provided by unexpected events are "temporary" in the sense that they change during an organism's lifetime: as events become predictable, they fade away. This is the reason they are particularly well suited to drive action acquisition. As an unpredicted event is detected, phasic dopamine is released, reinforcing (through dopamine-dependent learning in the striatum) the behaviours produced just before the detection of the event. As the organism repeats those behaviours with some modification (e.g. due to noise), sometimes the event will re-occur (thus reinforcing behaviours) while

other times it will not (thus suppressing them). This mechanism should make the animal converge on just those components of its motor output that are required for systematically producing the event. As this happens, the event becomes predictable for the animal, and thus stops to trigger dopamine. In this way, the agent has acquired a new action, i.e. a sequence of motor commands that systematically produce a specific change in the environment. Since the production of that action ceases to be reinforced, the animal will stop to trigger it, unless the outcome of the action becomes valuable because it turns out to be part of a chain of actions that leads to reward.

The reinforcements produced by biological rewards are "permanent" in the sense that they do not change during an organism's lifetime: e.g. eating (when the organism is hungry) is innately rewarding, from birth to death. Hence, when the animal has learned how to systematically get to the reward in a given context, the reinforcement signal will not fade away. This is the reason why, with serial conditioned stimuli, the (unpredicted) appearance of the earliest reward-predicting stimulus keeps on triggering phasic dopamine (Schultz et al., 1993; Schultz, 1998). And this is why the same mechanisms that allow the discovery and acquisition of novel actions can also drive the learning of how to deploy acquired actions so to maximize rewards: since biological rewards prevent that phasic dopamine fades away, the actions that bring to them keep on being reinforced indefinitely, thus leading to reward maximization. Note that the processes that make rewards "permanently rewarding" and that prevent dopamine habituation do not need to involve dopamine itself. Indeed, they might depend on the influences that rewards have on physiological variables (like water or glucose concentrations in the body) and work through the release of hormones or other neuromodulators. Hence, what in experiments are considered as "unconditioned" rewards (e.g. the sight or the taste of a food) may in fact be stimuli that have been conditioned during the animal's pre-experimental experience (as Schultz, 1998 suggests). What is important is that phasic stimuli that are predictive of biological rewards constitute permanent reinforcers, which can drive the maximization of reward through a TD-like learning process.

Our hypothesis is related to what psychologists have been calling *intrinsic motivations* (White, 1959; Berlyne, 1960; Ryan and Deci, 2000; Baldassarre and Mirolli, 2012). The concept of IM was introduced in the 1950s in animal psychology to explain experimental phenomena (e.g. Harlow, 1950; Butler, 1953; Montgomery, 1954; Butler and Harlow, 1957) that were incompatible with the Hullian theory of motivations as drives (Hull, 1943). In particular,

what is most relevant here is that phasic stimuli not related to biological rewards can be used to condition instrumental responses (Kish, 1955). Our hypothesis explains this well documented phenomenon (Williams and Lowe, 1972; Glow and Winefield, 1978; Reed et al., 1996) by assuming that unpredicted events represent intrinsic reinforcers that drive the same reinforcement learning processes as extrinsic rewards.

3. Testing the hypothesis through a simulated robotic model

To sum up, our hypothesis states that phasic dopamine is a TD-like learning signal dependent on two kinds of reinforcers: (1) temporary, intrinsic reinforcers, which drive the acquisition of a repertoire of actions; and (2) permanent, extrinsic reinforcers, which drive the learning of when to deploy acquired actions in order to maximize future rewards. The reason why animals need both kinds of reinforcers is that in real life the path that leads from basic movements to the acquisition of biological rewards is often too long for extrinsic reinforcers to suffice (Baldassarre, 2011). By helping the system to acquire a repertoire of actions, intrinsic reinforcers dramatically simplify the "search space" for the agent, and thus significantly facilitate the discovery of the path that leads to biological rewards (see the "intrinsically motivated reinforcement learning" framework proposed by Barto and collegues: Barto et al., 2004; Singh et al., 2010; Barto, 2012 and developed also by ourselves: Schembri et al., 2007c,b,a).

In order to test the computational soundness of our hypothesis we developed a simulated robotic set-up in which the acquisition of extrinsic rewards depends on the deployment of a sequence of "actions" that must themselves be learned. In such a set-up we show that extrinsic rewards alone are not sufficient to drive the reinforcement learning system, while adding intrinsic reinforcers dramatically facilitate reward acquisition. In order to ease reading, in what follows we describe only the most relevant features of the experiments whereas the details needed to replicate the simulations can be found in the Appendix.

3.1. 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 "hand" that can "grasp objects". The task consists in learning to eat food (i.e., bring a red object to the mouth) which is randomly placed on a

rectangular table in front of the robot (fig.1). The task and the perceptual system of the robot have been developed so that in order to eat the food the robot must learn and deploy a sequence of actions that depend the one on the other: since the arm controller is informed about food location through what the eye sees, learning to systematically look at the food is a prerequisite for learning to reach for it; similarly, reaching the food with the hand and "grasping" it are necessary pre-conditions for bringing it to the mouth and receiving the extrinsic reward.

The sensory system of the robot is composed by an artificial "retina", encoding the position of the hand and of the food with respect to the centre of the visual field, a "fovea", encoding whether the food is perceived in the centre of the visual field (i.e. if the food and the position of the fovea sensor are overlapping), the proprioception of the arm, encoding the angles of the two arm joints, and a touch sensor encoding whether the hand is in contact with the food (i.e, if the hand and the food are overlapping: collisions are not simulated).

The motor system of the robot is composed by two outputs encoding the displacements of the eye along the x and y axes, two outputs encoding the changes in the angles of the two arm joints, and a single output encoding whether grasping is performed or not (if the hand touches the food and the grasping output is activated the food moves together with the hand).

3.2. The control architecture

The control system of the robot (figure 2) has been developed by following general constraints that come both from the task and from the known biology behind reinforcement learning in real animals. In particular, the controller is composed of two sub-controllers, one for the eye and one for the arm, reflecting the modular organization of the striato-cortical loops that are known to subserve action selection and reinforcement learning (Doya, 2000; Graybiel, 2005; Grahn et al., 2009; Redgrave et al., 2011), for which different pathways subserve different effectors (Romanelli et al., 2005).

Each subcontroller is implemented as an actor-critic reinforcement learning model (Barto et al., 1983; Sutton and Barto, 1998), as this architecture can be considered as a good model of reinforcement learning in the basal ganglia (Barto, 1995; Suri, 2002; Joel et al., 2002; Khamassi et al., 2005). Both subcontrollers are trained through standard TD learning, reflecting the hypothesis that the phasic dopaminergic signal represents the biological substrate of the TD error (Houk et al., 1995; Schultz et al., 1997). Furthermore,

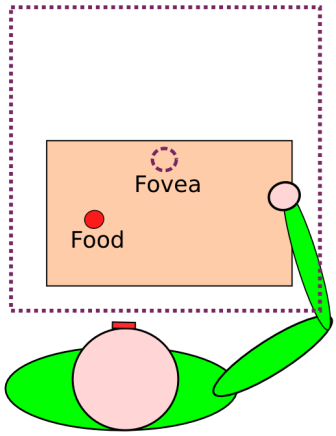


Figure 1: Set up of the experiment: the system is composed by a two dimensional arm and a moving eye (dotted square with a fovea at the centre). The task is to eat the food that is randomly positioned on a table (grey rectangle), by bringing it to the mouth (small rectangle in front of the robot's face). See text and Appendix for details.

while there are different controllers for different effectors, the reinforcement learning signal is unique for all the controllers, in accordance with the fact that the phasic DA signal is likely to be the same for all sensory-motor subsystems (Schultz, 2002).

Finally, the reinforcement depends not only on the extrinsic reward provided by eating the food, but also on intrinsic reinforcements provided by the unexpected activations of the fovea and the touch sensors, in accordance with the fact that unexpected events are able to trigger phasic dopamine (Ljungberg et al., 1992; Horvitz et al., 1997; Schultz, 1998; Horvitz, 2000). For this purpose, the robot controller includes also two predictors, one for the fovea sensor 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 composed by both extrinsic and intrinsic reinforcements:

$$R = R_e + R_f + R_t$$

where R_e is the extrinsic reinforcement provided by eating the food (bringing it to the mouth), and R_f and R_t are the intrinsic reinforcements provided by

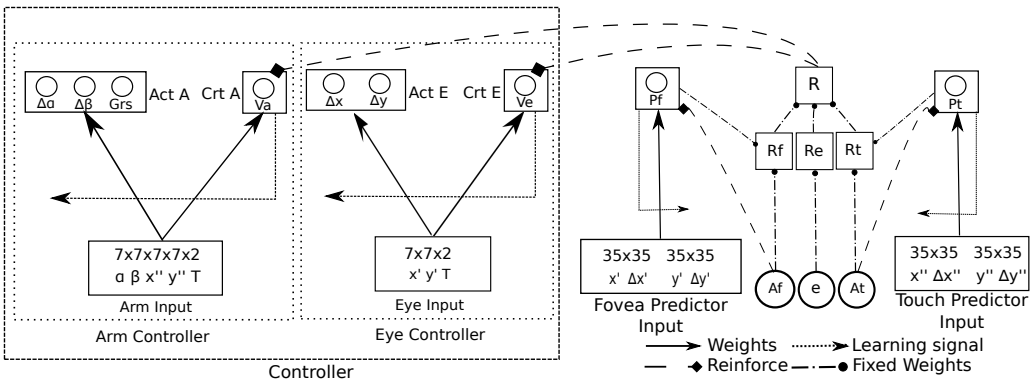


Figure 2: The controller, composed by the two sub-controllers (one for the arm and one for the eye), and the reinforcement system, which includes two predictors, one for the fovea sensor and one for the touch sensor. α and β are the angles of the two arm joints; x' and y' are the distances of the hand with respect to the centre 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 actor of the arm; Grs is the grasping output; V_a is the evaluation of the critic of the arm; x'' and y'' are the distances of the food 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 the actor of the eye; V_e is the evaluation of the critic of the eye; P_f and P_t are the predictions of the fovea and touch sensor predictors, respectively; A_f and A_t are the activations of the fovea and touch sensors, respectively; R_f and R_t are the reinforcements related to foveating and touching the food, respectively; R_e is the reinforcement provided by eating the food; R is the total reinforcement. See text and Appendix for details.

the unpredicted activations of, respectively, the fovea and the touch sensor:

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

where A_S is the binary activation of sensor S (A_f and A_t for the fovea and the touch sensor, respectively) and P_S is the prediction generated by the predictor of sensor S (P_f and P_t, both in [0, 1], see appendix).

3.3. Experimental conditions

In order to test our hypothesis, we compare the condition just described (which we call *"intrinsic"* condition) with two other conditions, in which we vary the composition of the reinforcement signal. In the *"extrinsic"* condition the reinforcement is given only by the extrinsic reward for eating the food (R_e). The *extrinsic* condition serves to test whether in a situation that requires the cumulative acquisition of different skills extrinsic reinforcements

alone are sufficient to drive learning. In the *"sub-tasks"* condition, the additional reinforcements provided by the activations of the two sensors (R_f 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 (i.e the prediction P_S of previous equation is always 0). This condition serves to investigate whether the temporary nature of intrinsic reinforcement is important for facilitating learning.

3.4. Results

Each experiment lasts 500000 trials. At the beginning of each trial the food is positioned randomly on the table, the joint angles of the arm are randomly initialized so that the hand is also on the table but does not touch the food, and the eye center is randomly positioned inside the table so that it does not look at the food. A trial terminates if food is eaten, if it falls off the table (i.e. if the food is outside the table and not "grasped"), or after a time-out of 40 time-steps. Every 500 trials we perform 50 test trials (where learning is switched off) during which we record useful statistics of the system's behavior. All reported data represent the average results of ten replications of each experiment with random initial conditions.

Figure 3 shows the percentage of test trials in which the robot eats the food in the three experimental conditions as a function of learning time. After 500000 trails the performance in the *extrinsic* condition is still below 20% (see figure 4a): as predicted, the extrinsic reinforcement is so difficult and unlikely to be reached that it is not able to drive the learning of the system, and, in particular, the learning of the sub-skills that are required to get to the reward (consider that the system is not guaranteed to learn the task even with infinite time since due to their partial sensory systems the problem for the two sub-controllers is non-markovian: see appendix for details).

In the *sub-tasks* condition, at the end of learning the robot eats the food in 80% of the test trials. Adding reinforcements for foveating and touching the food highly improves performance because it greatly facilitates the acquisition of the necessary sub-skills (fig. 4b): first, the eye learns to look at the food, and then the arm learns to touch and grasp it, which is a prerequisite for learning to eat. Notice that when the system has learned to reach for the food and grasp it, the time spent by the eye on the target diminishes, as indicated by the lowering of the reinforcements provided by the activation of the fovea sensor: the reason is that for architectural limits the eye is not

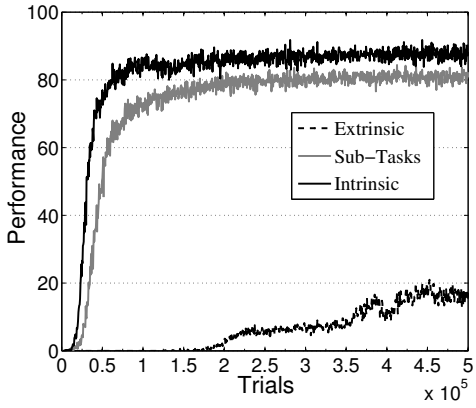


Figure 3: Percentage of test trials in which the robot eats the food throughout learning in the three experimental conditions (*Extrinsic, Sub-Tasks, Intrinsic*).

able to follow the food while the hand is grasping and moving it (the eye controller is not informed about the movements of the arm).

The *intrinsic* condition is the one in which performance increases most fastly and reaches the highest level (about 90%). The reason is that the reinforcements provided by the unpredicted activations of the sensors are ideally suited for driving the cumulative acquisition of a sequence of skills thanks to their temporal character. In this condition the reinforcements provided by the activations of the fovea and of the touch sensors rapidly grow as the related abilities (of foveating and reaching the food, respectively) are being acquired. But as the system learns to systematically foveate and touch the food, the related predictors also learn to predict the activations of the sensors, thus making the intrinsic reinforcements fade away (figure 4c). In this way, once a skill has been acquired the related reinforcement does not influence learning any more, and the system can focus on learning the next skill on the basis of its relative reinforcement.

At this point it is important to understand whether the results we have found really depend on the different experimental conditions or just on the particular quantitative relation between the value of intrinsic and the extrinsic reinforcements that we used. In order to check this we run again the experiment for all the three conditions varying the value of the extrinsic reinforcement provided by eating the food (from 5 to 30: much lower or higher values did not permit to any condition to learn the task). Figure 5 shows the average final performance (after 500000 trials) of ten repetitions

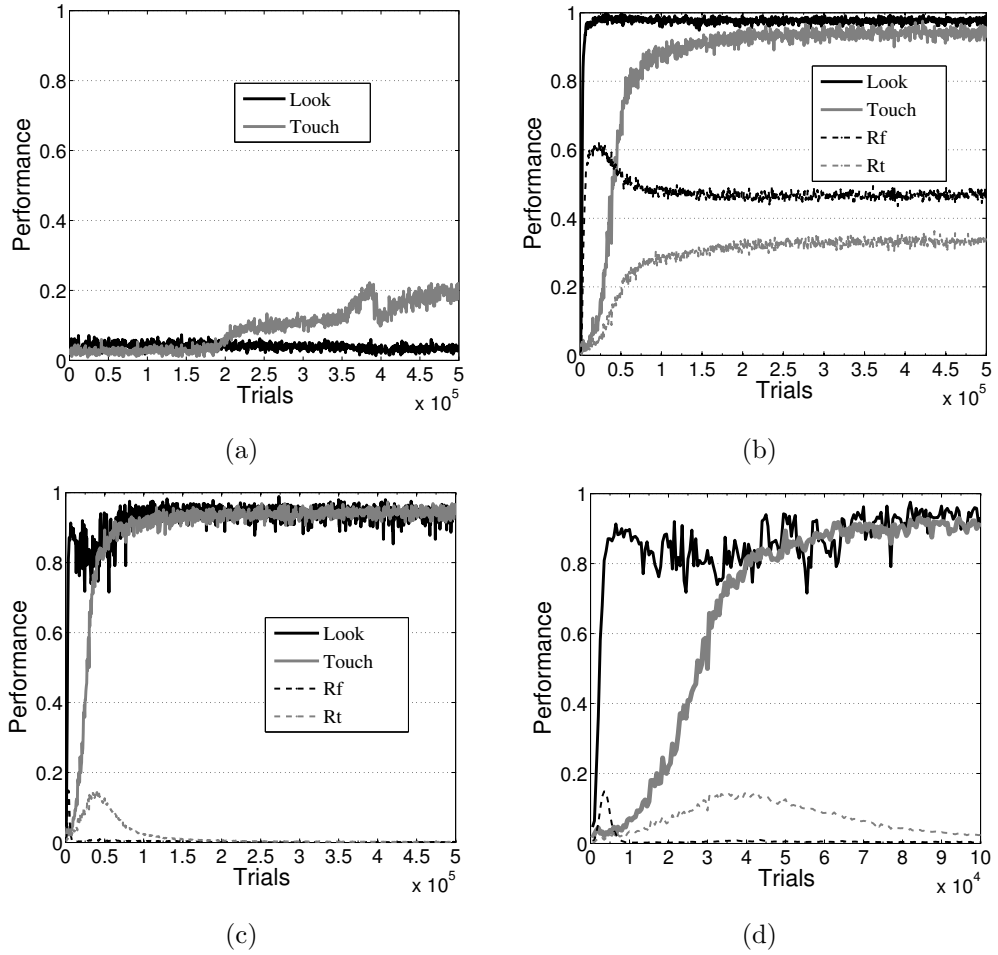


Figure 4: Average percentage of test trials in which the robot performs the sub-tasks (Look, Touch) in the three conditions: Extrinsic (a), Sub-Tasks (b) and Intrinsic (c). (d) zooms-in the first 100000 trials of the intrinsic condition. (b), (c) and (d) also show the average reinforcements related to the activation of the fovea sensor (Rf) and the touch sensor (Rt). Note that since the maximum reinforcements for each time step for foveation and touch are 1, in the Sub-Tasks condition at the end of learning the system foveates food about 50% of time steps and touches it about 35%.

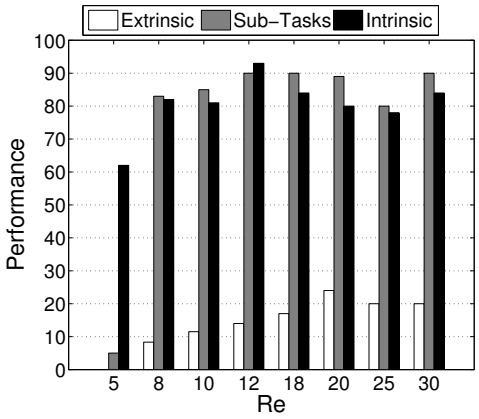


Figure 5: Average final performance (percentage of trials in which the system eats the food) of the three conditions (Extrinsic, Sub-Tasks and Intrinsic) as a function of the value of the extrinsic reinforcement (Re).

for each condition. The figure demonstrates that the results do not depend on the quantitative relation between the value of the intrinsic and extrinsic reinforcements: apart for $R_e = 5$, where only the *intrinsic* condition reaches good performance, for any value of the extrinsic reinforcement the *extrinsic* condition never learns to solve the task, whereas the *intrinsic* and the *sub-task* conditions reach comparable high performance.

Hence, while the results of the *extrinsic* condition clearly show that extrinsic reinforcements are not sufficient by themselves to drive the maximization of rewards in this set-up, the comparable results of the *sub-tasks* and the *intrinsic* conditions do not support our hypothesis regarding the importance of the temporal character of additional reinforcements. However, this may be due to a peculiar and un-realistic characteristic of the present set-up: additional reinforcements are given only for reaching those states that are required for getting to the final reward. In sharp contrast with this, for real organisms it is not possible to know a priori which are the actions needed for getting closer to biological rewards and which are not. Importantly, if all the changes that an organism can make in the environment would be permanently reinforcing, then the animal would easily get stuck in producing irrelevant events without passing on and eventually discover how to maximize biological rewards. It is the temporary nature of intrinsic reinforcements given by unexpected events that let organisms acquire a repertoire of actions, while freeing the animal from compulsively deploying those actions in case they do

not directly lead to rewards. In order to show this, we need a slightly more realistic set-up in which not all reinforced events are relevant for obtaining reward.

4. A more realistic test: adding a distractor

4.1. Modifications of the set-up, architecture and learning signal

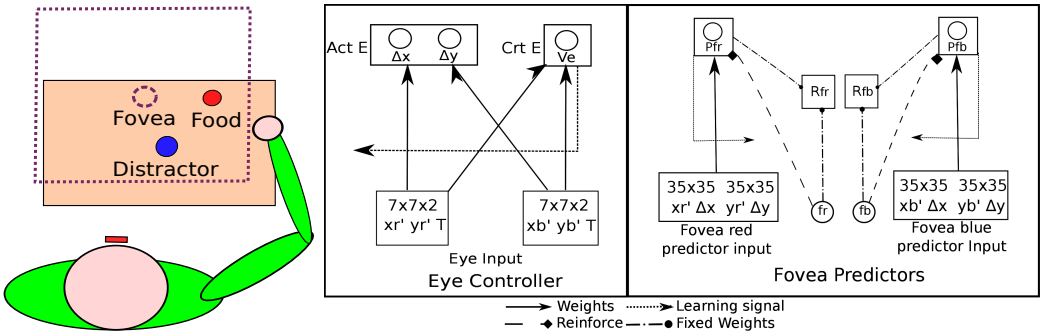
In order to test our idea that the temporary character of intrinsic reinforcements is necessary for preventing the system to get stuck in producing actions that are not relevant for the acquisition of rewards, we modified the set-up by simply adding another object on the table, which can be seen but not touched nor grasped, and which is not related to the final reward (figure 6a). The new object, which can be considered as a "distractor" with respect to the goal of eating the food, has a different "colour" with respect to the food (i.e. the two objects are visually detected by different sensors) and is always positioned in the middle of the table (which make the task more difficult because the distractor is more easily encountered than the food; we run also simulations with the distractor randomly positioned on the table and the results are almost identical).

With respect to the control system, the only modification that had to be done with respect to the previous scenario was to duplicate the visual system (of both the eye controller and of the fovea predictor) so that it can detect, with different sensors, the two objects: the food (red) and the distractor (blue) (figure 6b).

Finally, also the component of the reinforcement signal that depends on the activation of the fovea is duplicated as foveating the distractor (blue object) is reinforcing just as foveating the food (red object). As in the previous set-up, in the *intrinsic* condition intrinsic reinforcements are temporary as they depend on the *unpredicted* activations of the fovea and touch sensors, while in the *sub-task* condition additional reinforcements are permanent as they are not inhibited by predictors. The reinforcement of the *extrinsic* condition does not change, as it depends only on bringing the food to the mouth.

4.2. Results

Figure 7 shows the performance of the three experimental conditions in the new scenario. In the *extrinsic* condition the distractor does not influence the reinforcement learning system. As a consequence, the results are



(a) (b)

Figure 6: Differences of the second set-up with respect to the first one. (a) A "distractor", which can be seen but not touched, is added in the middle of the table. (b) Differences in the control system. Both the eye controller (left) and the fovea predictor (right) have been duplicated: they have two sets of receptors, each sensible to one of the objects (red food and blue distractor). Furthermore, also the fovea sensor and the relative component of the reinforcement signal have been duplicated, one for the unpredicted activation caused by the food and one for that caused by the distractor.

substantially similar to those obtained in the previous experiment, with a final performance of about 15%. This confirms the conclusion that extrinsic rewards alone are not sufficient to drive the learning of the skills that are necessary for eating food.

The comparison between the *sub-tasks* and the *intrinsic* conditions is more interesting. Whereas in the first experimental set-up the performance of the two conditions were comparable, the addition of the second object disrupts the performance of the *sub-tasks* condition (10%) while leaving substantially unchanged that of the *intrinsic* condition (about 85%).

To understand why this is so we have to look at data regarding the behavior of the eye in the two conditions (figure 8). In the *sub-tasks* condition (figure 8a), the robot rapidly learns to foveate the distractor, because it is always in the same position in the middle of the table and so it is easier to learn to look at it than to look at the food. The problem is that, since foveating the distractor is permanently reinforcing, the robot keeps on looking at it indefinitely, and never learns to look at the food. As a consequence, the robot does not learn to reach and grasp the food, which is a prerequisite for learning to bring it to the mouth.

Also in the *intrinsic* condition (fig.8b) the robot starts by looking at the distractor, but after the ability to foveate it has been learned, the activation

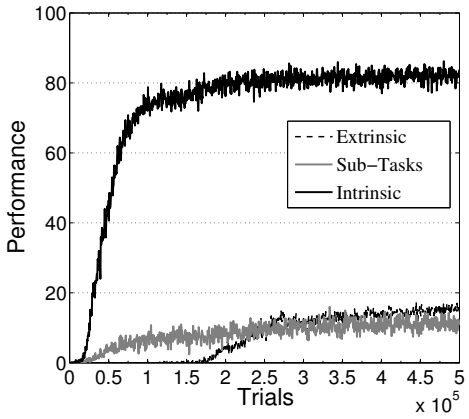
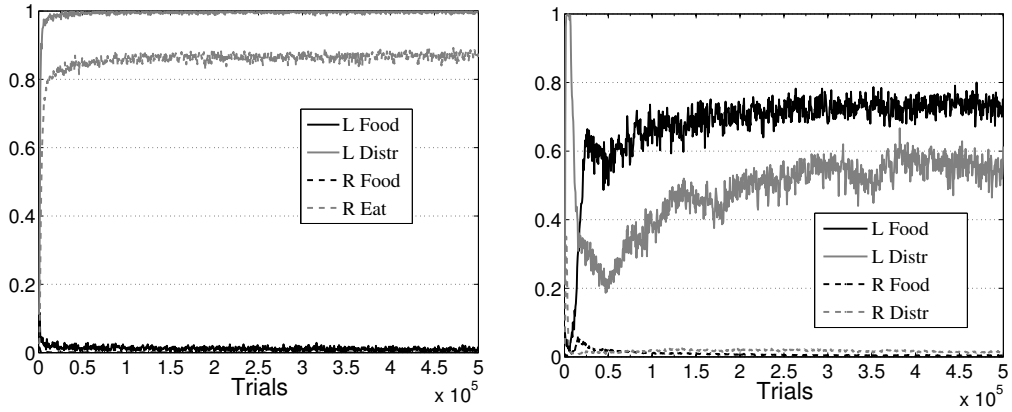


Figure 7: Performance (percentage of test trials in which the robot eats the food) of the three experimental conditions in the second set-up, containing the distractor

of the fovea sensor that is sensitive to the blue object (the distractor) starts to be predicted by the corresponding predictor, which rapidly makes this event 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. Even the reinforcement given by foveating the food fades aways as soon as the skill is acquired and the activation of the fovea is predictable, but the robot never stops producing this behaviour because it leads to the acquisition of other reinforcements: first the temporary ones that depend on touching the food, and then the permanent, extrinsic ones provided by bringing the food to the mouth. Notice that as the robot learns to eat the food, the number of times the robot looks at the distractor increases again. This is due to the same architectural constraints that decreased the percentage of time spent by the eye on the food in the first experimental scenario: as the food is grasped and moved towards the mouth, the lack of information about the arm movement of the eye controller does not allow it to follow the food. As a result, the eye resorts to the behavior that it had previously learned, i.e. foveating the distractor.

These results seem to confirm our hypothesis regarding the necessity of the temporal character of intrinsic reinforcements, but we need to check whether the results depend on the quantitative relation between value of the intrinsic and extrinsic reinforcements, as done for the first set-up. Figure 9 shows the average final performance of ten repetitions for each condition as a function of the value of the extrinsic reinforcement (R_e). The results



(a) (b)

Figure 8: Behavior of the eye in the second set-up (with distractor) for the Sub-Tasks (a) condition and the Intrinsic condition (b). Average percentage of test trials in which the eye foveates the food (L Food) and the other object (L Other) and average reinforcements per step generated by the activations of the two sensors (R Food and R Other)

clearly show that irrespective of the value of the extrinsic reward, the *intrinsic* condition is the only one that reaches high performance. Indeed, in all cases the *sub-task* condition reaches a performance that is even lower than that of the *extrinsic* condition, demonstrating that if one cannot know which events will lead closer to biological rewards (which is what happens for real organisms), permanently reinforcing all events is not only useless, but can even be deleterious. Only intrinsic reinforcements given by unexpected events are able to drive the cumulative acquisition of all the skills that are necessary for learning to maximize extrinsic rewards.

5. Discussion

The current debate over the role of phasic dopamine is split in two opposing views: the received wisdom, supported by a great number of empirical findings, holds that dopamine is a reward prediction error that drives animals to learn how to deploy actions in order to maximize biological rewards (e.g. Schultz, 2002; Ungless, 2004; Wise, 2004; Doya, 2007; Graybiel, 2008; Glimcher, 2011); an alternative position, based on different empirical evidences, holds that dopamine is a sensory prediction error that drives action discovery and acquisition (Redgrave and Gurney, 2006; Redgrave et al.,

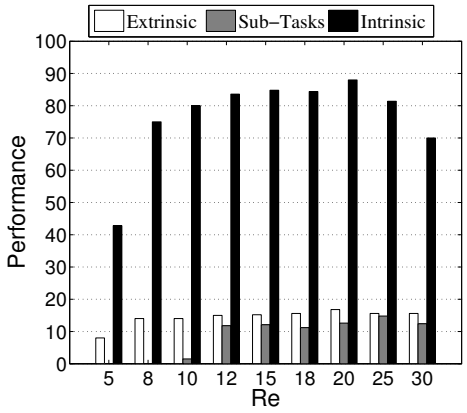


Figure 9: Average final performance on the eating task in the second experimental scenario of the three conditions (Extrinsic, Sub-Tasks and Intrinsic) as a function of the value of the extrinsic reinforcement (Ret) provided by eating the food. See text for details.

2008, 2011, 2012). Each hypothesis is insufficient in that it is not able to account for the data on which the other hypothesis is based: the reward prediction error hypothesis does not explain why dopamine is triggered also by unexpected events not related to rewards; the sensory prediction error hypothesis does not explain why dopamine corresponds so strictly to the TD reward prediction error postulated by computational reinforcement learning. In this paper we have proposed an original hypothesis that reconciles these two opposing views and thus is able to account for both kinds of empirical evidence on which the two views are based. According to our proposal phasic dopamine is a TD-like reinforcement prediction error signal in a learning system that is driven by two kinds of reinforcements: intrinsic, temporary reinforcements provided by unexpected events, and extrinsic, permanent reinforcements provided by biological rewards. As such, dopamine plays both functions: temporary reinforcements drive the discovery and acquisition of new actions, whereas permanent reinforcements drive the maximization of rewards. We have tested this hypothesis with a series of experiments involving a simulated robotic system that in order to get rewards has to cumulatively acquire different skills. The results showed that, if not all the possible skills that can be acquired directly lead to reward, only a system that receives intrinsic temporary reinforcements in addiction to the extrinsic ones is able to learn the task, thus supporting our hypothesis.

Kakade and Dayan (2002) had tried to reconcile the reward prediction error hypothesis with the fact that dopamine is also triggered by stimuli not related to reward by assuming that such dopamine activations constitute novelty bonuses whose function is to increase animal exploration. Our proposal differs from that of Kakade and Dayan with respect to both the function and the mechanism of reward-unrelated dopamine activations. With respect to the function, our proposal holds that reward-unrelated dopamine activations have the function of driving action discovery and acquisition (as in Redgrave and colleagues' theory) and not of enhancing exploration (as suggested by Kakade and Dayane: see also Fellous and Suri, 2003). Our view is supported by the long accepted evidence that unpredicted events can be used as reinforcers for training instrumental actions (e.g. Kish, 1955; Williams and Lowe, 1972; Glow and Winefield, 1978; Reed et al., 1996). With respect to the mechanism, according to our view dopamine is triggered by unexpected events (i.e. unpredicted changes of state), and not by the novelty of the presented stimuli (i.e. states). Even though in the literature dopamine activations triggered by stimuli not associated with rewards have been described as *novelty responses* (see, e.g. Schultz, 1998), in all experiments that we know about phasic dopamine is triggered by *events* (i.e. onset and/or offset of stimuli), and not by stimuli alone, be they novel or familiar. The same is true for the behavioral experiments showing that sensory *events* (and not novel stimuli) are able to drive the acquisition of instrumental responses. Indeed, given that most of these experiments use the simple switching on of a light as the dopamine triggering stimulus (or as the reinforcer), it seems plausible that what really matters is the unpredictability of the event, rather than the novelty of the stimulus itself, since it is difficult to consider the light as a novel stimulus. In fact, at least in behavioral experiment with rats, it has been shown that prior exposure to the light that is used to condition operant responses does not significantly affect the reinforcing effect of the light (Russell and Glow, 1974), as would be predicted by our hypothesis that it is the unpredictability of the event, rather than its novelty, that triggers phasic dopamine and reinforces behavior.

According to our hypothesis the phasic dopaminergic bursts determined by reward-unrelated unpredicted events constitute (part of) the neural substrate of intrinsic motivations: in particular, they represent the TD-error generated by intrinsic reinforcers. This hypothesis predicts also that, whereas the responses conditioned through biological rewards will be maintained because of the permanent nature of extrinsic reinforcers, the responses condi-

tioned through phasic events alone in the long term will tend to fade away because of the temporary nature of intrinsic reinforcers: once the action of "light switching" has been learned, the appearance of the light becomes predictable, hence the light switching stops to activate phasic dopamine and to reinforce the action, which makes the behavior extinguish. This prediction is confirmed by behavioral experiments with rats: it is well documented that the number of responses conditioned through intrinsic reinforcements do decline with prolonged training (e.g. Roberts et al., 1958; Carlton, 1966; Russell and Glow, 1974).

The task and experimental set-up that we used for validating our hypothesis were rather simple and and somewhat artificial. It is important to consider that our experiments were not intended to model how humans learn to foveate, reach, and bring objects to the mouth, nor to demonstrate the computational power of the model that we used. Rather, the model has to be considered just as a *proof of concept* that the hypothesis we have proposed on the mechanisms and functional roles of phasic dopamine in real brains is computationally sound. In particular, that a system that is reinforced by both permanent extrinsic reinforcements and temporary intrinsic ones provided by unexpected events is able to cumulatively acquire complex skills that are very difficult to acquire on the basis of only the final reinforcements, and that the temporary nature of intrinsic reinforcements, that is the fact that they fade away as the system's learning proceeds, is pivotal for letting the system stop performing a skill once it has been acquired and if it doesn't lead to reward. One aspect of our set-up that is particularly critical is that we had to use predictors that were specifically designed to learn just the events that we planned to be significant: foveating objects and touching them. The main reason for this is that we had to keep the set-up simple enough that simulations could be computationally feasible in a reasonable period of time. Real organisms have much more computational resources and can take days, months and even years to learn a skill, so they can afford much less specific predictors as our own. However, it is important to note that the second set-up, with the distractor, was specifically intended to demonstrate that, contrary to what happens if the system does not receive temporary intrinsic reinforcements, for it to work it is not necessary that only the events that lead to reward are reinforcing. Hence, we contend that a process of cumulative acquisition of skills as the one demonstrated by our model (but much more powerful) should be present even in organisms, for which, we assume, any kind of unpredicted event is reinforcing. From the computational point

of view, we still do not know how to design powerful, general purpose event predictors, nor powerful controllers that may permit real open-ended skill learning. In this respect, the development of more sophisticated models is an important challenge for future computational research.

While the empirical evidence clearly shows that the phasic dopamine that is triggered by neutral events is temporary, the presence of predictors that learn to anticipate these events and thus inhibit dopamine release is an assumption of our model (a similar hypothesis has been made also by Redgrave et al., 2011). However, strictly speaking it is not even necessary that the temporary character of intrinsic reinforcement depend on event predictors for our general hypothesis on the functional roles of dopamine to hold. It may be that other processes, like for example sensory habituation, are involved. What is crucial for our hypothesis is that the events that at the beginning trigger phasic dopamine stop to do so after a while, which has been consistently reported in the literature. Independently from which is the reason for this, our hypothesis states (and our model confirms) that this temporary nature of intrinsic reinforcements serves the critical function of letting the system learn new actions and then pass to learn other things.

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; Uchibe and Doya, 2008; Lee et al., 2009; Baldassarre and Mirolli, 2012). The idea of using a sensory prediction error as an intrinsic reinforcement has been firstly proposed by Schmidhuber (1991a) and used in various subsequent models (e.g. Huang and Weng, 2002). In particular, the model probably more similar to ours is that proposed by Barto et al. (2004), where intrinsic reinforcements are given by the error in the prediction of *salient events*. The most important difference between Barto and colleagues' work and our own lies in the aims of the research: while Barto et al. took inspiration from biology to develop more efficient artificial systems, the goal of the present work is purely scientific, that is to propose a new hypothesis that reconciles the two opponent theories on phasic dopamine and accounts for all the available empirical evidence. This fundamental difference in perspective led to several differences in the details of the model's architecture: while Barto et al. use options (Sutton et al., 1999), a very powerful hierarchical reinforcement learning framework, we use plain reinforcement learning; while they use intra-option learning methods in which each skill has its own learning signal, in our system the reinforcement learning signal

is unique for all the controllers, as in the brain phasic DA is likely to be the same for all sensory-motor subsystems (Schultz, 2002); while they use option probabilistic models, we use simple event predictors; while they use Q-Learning (Sutton and Barto, 1998), we use the actor-critic architecture, which be considered as a good model of reinforcement learning in the basal ganglia (Barto, 1995; Suri, 2002; Joel et al., 2002; Khamassi et al., 2005).

Although using sensory prediction errors as intrinsic reinforcements has a relatively long history in computational work, Schmidhuber (1991b) pointed out that a pure sensory prediction error might not be a good reinforcement signal as it would create problems when the environment is unpredictable: in such cases, 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 predictions* was proposed as a better intrinsic reinforcement, a solution that has been adopted also in developmental robotic systems (e.g. Oudeyer et al., 2007). In contrast to this, the experimental data on phasic dopamine, and the hypothesis that we propose for explaining those data, seem to show that the intrinsic reinforcement signals that drive action learning depend on unpredicted events, not on progress in predictions. How could the problem of getting stuck on unpredictable events be solved? We think that a possible solution might depend on the presence of other motivational mechanisms working at an higher level of the hierarchical organization of behavior. In particular, if we assume that there is a level at which organisms decide what to learn and when (which skill to train in each context), intrinsic reinforcements given to this part of the learning system and based on the *learning progress in skill acquisition* (as the ones used in Schembri et al., 2007a,b,c; see also Stout and Barto, 2010) would solve the problem of unpredictability: if there are no skills to acquire due to the unpredictability of events, the reinforcement provided by competence progress will be zero, and the system will move forward and try to learn something else. While the presence of competence-based intrinsic motivations has been variously postulated in the psychological literature (e.g. White, 1959; De Charms, 1968; Glow and Winefield, 1978; Csikszentmihalyi, 1991), the identification of their possible biological implementation remains a fundamental open issue for future research (for a more detailed discussion of this point, see Mirolli and Baldassarre, 2012).

Appendix A. Computational details of the experiments

Here we provide all the details that are necessary to reproduce the simulations described in the paper.

The visual field of the robot is a square of 14 units per size. The arm of the robot is composed of two segments which are 4 units long. The food is a circle with 0.3 units diameter. In the second set of experiments, the "distractor" is a circle with a diameter of 0.4. The table is a rectangle measuring 4 and 7 units.

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

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

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 eye controller are the position of the object (x and y) relative to the centre of the visual field (the fovea) 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, forms a total grid of 7x7x2. In the second experiment, the input to the eye controller is formed by two 7x7x2 grids, one for the red object (food) and one for the blue object (distractor).

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 th 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, the total grid of the input to the arm is formed by 7x7x7x7x2 units.

The two sub-controllers (of the eye and of the arm) are neural network implementations of the actor-critic architecture (Sutton and Barto, 1998) adapted to work with continuous states and actions spaces (Doya, 2000; Schembri et al., 2007a), in discrete time.

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 command (in [0; 1]) is remapped in [-8, 8] and determines the displacement of the eye (Δx and Δy).

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 ($\Delta\alpha$ and $\Delta\beta$, respectively). The third output unit has binary activation $\{0; 1\}$, and controls the grasping action (The binary activation of the third output 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:

$$V_k = \sum_i^{N_k} a_{ki} w_{ki}$$

Learning depends on the TD reinforcement learning algorithm (Sutton and Barto, 1998), where the TD error δ_k of each sub-controller k is calculated 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 extrinsic reinforcement provided

by bringing the food to the mouth is 15 in all the conditions of the first experiment. In order to avoid that the system tries to perform grasping even when the hand is not close to the food, the activation of the grasping output (for each time step) 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:

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

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), $o_{kj}^n - o_{kj}$ is the 'error signal' (the produced noisy action minus the action chosen by the network before adding noise), and $o_{kj}(1 - o_{kj})$ is the derivative of the sigmoid function.

Also the input of the predictors is composed of RBF units. The input of the fovea sensor predictor is formed by two 35x35 grids, each encoding the position of the object with respect to the fovea along one axis (x and y, respectively), and the programmed displacement of the eye along the same axis (δx and δy, respectively). Similarly, the input of the touch sensor predictor is formed by two 35x35 grids, each encoding the position of hand with respect to the object 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 object positions and [-25; 25] for displacements. The output of each predictor is a single sigmoidal unit with activation in [0; 1] receiving connections from all the predictor's input units.

Event predictors are trained through a TD learning algorithm (for a generalization of TD learning to general predictions, see Sutton and Tanner, 2005). For each predictor p, the TD error δ_p is calculated as follows:

$$\delta_p = (A_S^t + \gamma_p P_S^t) - P_S^{t-1}$$

where A_S^t is the activation of sensor S (fovea or touch sensor) at time step t, P_S^t is the prediction relative to sensor S at time step t, and γ_p is the predictors' discount factor, set to 0.7.

Finally, the weights of predictor p, are updated as follow:

$$\Delta w_{pi} = \eta_p^c \delta_p a_p i$$

where η_p^c is the learning rate, set to 0.00008. Low values for predictors' gammas and learning rates prevent that predictors inhibit the intrinsic reinforcement too early, in particular before the system has acquired the relative skills. We have discussed more principled solutions to this potential problem in Santucci et al. (2012).

Aknowledgements

This research was supported by the EU Project IM-CLeVeR, contract no. FP7-IST-IP-231722. The authors thank Peter Redgrave and Kevin Gurney for our prolonged discussions on the reward prediction error and the sensory prediction error theories of phasic dopamine.

Baldassarre, G., 2011. 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, New York. pp. E1–8.

Baldassarre, G., Mirolli, M. (Eds.), 2012. Intrinsically Motivated Learning in Natural and Artificial Systems. Springer-Verlag, Berlin.

Barto, A., 1995. Adaptive critics and the basal ganglia, in: Houk, J., Davis, J., Beiser, J. (Eds.), Models of Information Processing in the Basal Ganglia. MIT Press, Cambridge, MA, pp. 215–232.

Barto, A., 2012. Intrinsic motivation and reinforcement learning, in: Baldassarre, G., Mirolli, M. (Eds.), Intrinsically Motivated Learning in Natural and Artificial Systems. Springer-Verlag.

Barto, A., Singh, S., Chantanez, N., 2004. Intrinsically motivated learning of hierarchical collections of skills, in: Proceedings of the Third International Conference on Developmental Learning (ICDL), pp. 112–119.

Barto, A., Sutton, R., Anderson, C., 1983. Neuron-like adaptive elements that can solve difficult learning control problems. IEEE Transactions on Systems, Man, and Cybernetics 13, 834–846.

Bayer, H.M., Glimcher, P.W., 2005. Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47, 129–141.

Berlyne, D., 1960. Conflict, Arousal and Curiosity. McGraw Hill, New York.

Berridge, K., 2007. The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology 191, 391–431.

Bjorklund, J., Dunnett, S., 2007. Dopamine neuron systems in the brain: An update. Trends in Neurosciences 30, 194–202.

Butler, R.A., 1953. Discrimination learning by rhesus monkeys to visual-exploration motivation. Journal of Comparative Physiology and Psychology 46, 95–98.

Butler, R.A., Harlow, H.F., 1957. Discrimination learning and learning sets to visual exploration incentives. J Gen Psychol 57, 257–264.

Calabresi, P., Picconi, B., Tozzi, A., Filippo, M.D., 2007. Dopamine-mediated regulation of corticostriatal synaptic plasticity. Trends in Neuroscience 30, 211–219.

Carlton, P.L., 1966. Scopolamine, amphetamine and light-reinforced responding. Psychonomic Science 5, 347–348.

Chiodo, L.A., Antelman, S.M., Caggiula, A.R., Lineberry, C.G., 1980. Sensory stimuli alter the discharge rate of dopamine (da) neurons: evidence for two functional types of da cells in the substantia nigra. Brain Research 189, 544–549.

Csikszentmihalyi, M., 1991. Flow: The Psychology of Optimal Experience. Harper Perennial, New York, NY.

Daw, N.D., Niv, Y., Dayan, P., 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neuroscience 8, 1704–1711.

Dayan, P., Sejnowski, T., 1996. Exploration bonuses and dual control. Machine Learning 25, 5–22.

De Charms, R., 1968. Personal Causation: The Internal Affective Determinants of Behavior. Academic Press, New York, NY.

Dommett, E., Coizet, V., Blaha, C.D., Martindale, J., Lefebvre, V., Walton, N., Mayhew, J.E.W., Overton, P.G., Redgrave, P., 2005. How visual stimuli activate dopaminergic neurons at short latency. Science 307, 1476–1479.

Doya, K., 2000. Reinforcement learning in continuous time and space. Neural Computation 12, 219–245.

Doya, K., 2007. Reinforcement learning: Computational theory and biological mechanisms. HFSP Journal of Comparative and Physiological Psychology 1, 30–40.

Fellous, J., Suri, R., 2003. Roles of dopamine, in: Arbib, M. (Ed.), The Handbook of Brain Theory and Neural Networks. MIT Press, Cambridge, MA, pp. 361–365.

Fiore, V., Mannella, F., Gurney, K., Baldassarre, G., 2008. Instrumental conditioning driven by neutral stimuli: a model tested with simulated robotic rat, in: Schlesinger, M., Balkenius, L.B.C. (Eds.), Proceedings of the Eight International Conference on Epigenetic Robotics, pp. 13–20.

Fiorillo, C.D., Newsome, W.T., Schultz, W., 2008. The temporal precision of reward prediction in dopamine neurons. Nature Neuroscience 11, 966–973.

Frank, M.J., 2005. Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated parkinsonism. Journal of Cognitive Neuroscience 17, 51–72.

Glimcher, P., 2011. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc Natl Acad Sci U S A 108 Suppl 3, 15647–15654.

Glow, P., Winefield, A., 1978. Response-contingent sensory change in a causally structured environment. Learning & Behavior 6, 1–18. 10.3758/BF03211996.

Grace, A., Floresco, S., Goto, Y., Lodge, D., 2007. Regulation of firing of dopaminergic neuron and control of goal-directed behavior. Trends in Neurosciences 30, 220–227.

Grahn, J.A., Parkinson, J.A., Owen, A.M., 2009. The role of the basal ganglia in learning and memory: neuropsychological studies. Behavioural Brain Research 199, 53–60.

Graybiel, A., 2008. Habits, rituals, and the evaluative brain. Annual Reviews Neuroscience 31, 359–387.

Graybiel, A.M., 2005. The basal ganglia: learning new tricks and loving it. Current Opinions in Neurobiology 15, 638–644.

Harlow, H.F., 1950. Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys. Journal of Comparative and Physiological Psychology 43, 289–294.

Hollerman, J.R., Schultz, W., 1998. Dopamine neurons report an error in the temporal prediction of reward during learning. Nature Neuroscience 1, 304–309.

Horvitz, J.C., 2000. Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. Neuroscience 96, 651–656.

Horvitz, J.C., Stewart, T., Jacobs, B.L., 1997. Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. Brain Research 759, 251–258.

Houk, J., Adams, J., Barto, A., 1995. A model of how the basal ganglia generate and use neural signals that predict reinforcement, in: Houk, J., Davis, J., Beiser, D. (Eds.), Models of Information Processing in the Basal Ganglia. MIT Press, Cambridge, MA, pp. 249–270.

Huang, X., Weng, J., 2002. Novelty and reinforcement learning in the value system of developmental robots, in: Prince, C., Demiris, Y., Marom, Y., Kozima, H., Balkenius, C. (Eds.), Proceedings of the Second International Workshop Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, Lund University Cognitive Studies, Lund. pp. 47–55.

Hull, C.L., 1943. Principles of behavior. Appleton-century-crofts.

Joel, D., Niv, Y., Ruppin, E., 2002. Actor-critic models of the basal ganglia: new anatomical and computational perspectives. Neural Networks 15, 535–547.

Kakade, S., Dayan, P., 2002. Dopamine: generalization and bonuses. Neural Networks 15, 549–559.

Kaplan, F., Oudeyer, P., 2003. Motivational principles for visual know-how development, in: Prince, C., 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.

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 13, 131–148.

Kish, G.B., 1955. Learning when the onset of illumination is used as reinforcing stimulus. Journal of Comparative and Physiological Psychology 48, 261–264.

Lee, R., Walker, R., Meeden, L., Marshall, J., 2009. Category-based intrinsic motivations, in: Canamero, L., Oudeyer, P., Balkenius, C. (Eds.), Proceedings og the Ninth International Conference on Epigenetic Robotics, Lund University Cognitive Studies, Lund. pp. 81–88.

Ljungberg, T., Apicella, P., Schultz, W., 1991. Responses of monkey midbrain dopamine neurons during delayed alternation performance. Brain Research 567, 337–341.

Ljungberg, T., Apicella, P., Schultz, W., 1992. Responses of monkey dopamine neurons during learning of behavioral reactions. Journal of Neurophysiology 67, 145–163.

May, P.J., McHaffie, J.G., Stanford, T.R., Jiang, H., Costello, M.G., Coizet, V., Hayes, L.M., Haber, S.N., Redgrave, P., 2009. Tectonigral projections in the primate: a pathway for pre-attentive sensory input to midbrain dopaminergic neurons. European Journal of Neuroscience 29, 575–587.

Mirolli, M., Baldassarre, G., 2012. Functions and mechanisms of intrinsic motivations: The knowledge vs. competence distinction, in: Baldassarre, G., Mirolli, M. (Eds.), Intrinsically Motivated Learning in Natural and Artificial Systems. Springer-Verlag, Berlin.

Montague, P.R., Hyman, S.E., Cohen, J.D., 2004. Computational roles for dopamine in behavioural control. Nature 431, 760–767.

Montgomery, K., 1954. The role of the exploratory drive in learning. Journal of Comparative Psychology 47, 60–64.

Morris, G., Arkadir, D., Nevet, A., Vaadia, E., Bergman, H., 2004. Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. Neuron 43, 133–143.

Ng, A.Y., Harada, D., Russell, S., 1999. Policy invariance under reward transformations: Theory and application to reward shaping., in: Proceedings of the 16th International Conference of Machine Learning, Morgan Kaufmann Publisher Inc., San Francisco, CA, USA. pp. 278–287.

Oudeyer, P., Kaplan, F., Hafner, V., 2007. Intrinsic motivation system for autonomous mental development, in: IEEE Transactions on Evolutionary Computation, pp. 703–713.

Pouget, A., Snyder, L.H., 2000. Computational approaches to sensorimotor transformations. Nature Neuroscience 3 Suppl, 1192–1198.

Redgrave, P., Gurney, K., 2006. The short-latency dopamine signal: a role in discovering novel actions? Nature Reviews Neuroscience 7, 967–975.

Redgrave, P., Gurney, K., Reynolds, J., 2008. What is reinforced by phasic dopamine signals? Brain Research Reviews 58, 322–339.

Redgrave, P., Gurney, K., Stafford, T., Thirkettle, M., Lewis, J., 2012. The role of the basal ganglia in discovering novel actions, in: Baldassarre, G., Mirolli, M. (Eds.), Intrinsically Motivated Learning in Natural and Artificial Systems. Springer-Verlag.

Redgrave, P., Prescott, T.J., Gurney, K., 1999. Is the short-latency dopamine response too short to signal reward error? Trends in Neuroscience 22, 146–151.

Redgrave, P., Vautrelle, N., Reynolds, J.N.J., 2011. Functional properties of the basal ganglia's re-entrant loop architecture: selection and reinforcement. Neuroscience .

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

Reynolds, J.N., Hyland, B.I., Wickens, J.R., 2001. A cellular mechanism of reward-related learning. Nature 413, 67–70.

Reynolds, J.N.J., Wickens, J.R., 2002. Dopamine-dependent plasticity of corticostriatal synapses. Neural Netw 15, 507–521.

Robbins, T., Everitt, B., 1992. Functions of dopamine in the dorsal and ventral striatum. Semin. Neurosci. 4, 119–128.

Roberts, C.L., Marx, M.H., Collier, G., 1958. Light onset and light offset as reinforcers for the albino rat. Journal of Comparative and Physiological Psychology 51, 575–579.

Robinson, S., Sotak, B., During, M., Palmiter, R., 2006. Local dopamine production in the dorsal striatum restores goal-directed behavior in dopamine-deficient mice. Behav Neurosci 120, 196–200.

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 48, 112–128.

Romo, R., Schultz, W., 1990. Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. Journal of Neurophysiology 63, 592–606.

Russell, A., Glow, P., 1974. Some effects of short-term immediate prior exposure to light change on responding for light change. Learning & Behavior 2, 262–266. 10.3758/BF03199191.

Ryan, Deci, 2000. Intrinsic and extrinsic motivations: Classic definitions and new directions. Contemporary Educational Psychology 25, 54–67.

Salzman, C.D., Belova, M.A., Paton, J.J., 2005. Beetles, boxes and brain cells: neural mechanisms underlying valuation and learning. Current Opinions in Neurobiology 15, 721–729.

Santucci, V., Baldassarre, G., Mirolli, M., 2012. Intrinsic motivation mechanisms for competence acquisition, in: Proceedings of ICDL-Epirob 2012, San Diego.

Schembri, M., Mirolli, M., Baldassarre, G., 2007a. Evolution and learning in an intrinsically motivated reinforcement learning robot, in: y Costa, F.A., Rocha, L., Costa, E., Harvey, I., Coutinho, A. (Eds.), Advances in Artificial Life, Springer, Berlin. 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., 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., Wilson, S. (Eds.), Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats, MIT Press/Bradford Books, Cambridge, Massachusetts/London, England. pp. 222–227.

Schmidhuber, J., 1991b. Curious model-building control system, in: Proceedings of International Joint Conference on Neural Networks, IEEE, Singapore. pp. 1458–1463.

Schultz, W., 1998. Predictive reward signal of dopamine neurons. Journal of Neurophysiology 80, 1–27.

Schultz, W., 2002. Getting formal with dopamine and reward. Neuron 36, 241–263.

Schultz, W., 2006. Behavioral theories and the neurophysiology of reward. Annual Reviews of Psychology 57, 87–115.

Schultz, W., 2007. Multiple dopamine functions at different time scales. Annual Reviews of Neuroscience 30, 259–288.

Schultz, W., Apicella, P., Ljumberg, T., 1993. Responses of monkey dopamine neurons to reaward and conditioned stimuli during successive steps of learning a delayed response task. Journal of Neuroscience 13, 900–913.

Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275, 1593–1599.

Singh, S., Barto, R.L.A., Sorg, J., 2010. Instrinsically motivated reinforcement learning: An evolutionary perspective. IEEE Transactions on Autonomous Mental Development 2, 70–82.

Steinfels, G.F., Heym, J., Strecker, R.E., Jacobs, B.L., 1983. Response of dopaminergic neurons in cat to auditory stimuli presented across the sleep-waking cycle. Brain Research 277, 150–154.

Stout, A., Barto, A.G., 2010. Competence progress intrinsic motivation, in: Proceedings of the International Conference on Development and Learning (ICDL), pp. 257–262.

Strecker, R.E., Jacobs, B.L., 1985. Substantia nigra dopaminergic unit activity in behaving cats: effect of arousal on spontaneous discharge and sensory evoked activity. Brain Research 361, 339–350.

Sugrue, L.P., Corrado, G.S., Newsome, W.T., 2005. Choosing the greater of two goods: neural currencies for valuation and decision making. Nature Reviews Neuroscience 6, 363–375.

Suri, R.E., 2002. Td models of reward predictive responses in dopamine neurons. Neural Networks 15, 523–533.

Sutton, R., 1988. Learning to predict by the methods of temporal differences. Machine Learning 3, 9–44.

Sutton, R., 1990. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming, in: Proceedings of the Seventh International Conference on Machine Learning, Morgan Kaufmann. pp. 216–224.

Sutton, R., Barto, A., 1998. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.

Sutton, R., Precup, D., Singh, S., 1999. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. Artificial Intelligence 112, 181–211.

Sutton, R., Tanner, B., 2005. Temporal-difference networks. Advances in neural information processing systems 17, 1377–1348.

Tobler, P.N., Fiorillo, C.D., Schultz, W., 2005. Adaptive coding of reward value by dopamine neurons. Science 307, 1642–1645.

Uchibe, E., Doya, K., 2008. Finding intrinsic rewards by embodied evolution and constrained reinforcement learning. Neural Networks 21, 1447–1455.

Ungless, M., 2004. Dopamine: the salient issue. Trends in Neuroscience 27, 702–706.

Waelti, P., Dickinson, A., Schultz, W., 2001. Dopamine responses comply with basic assumptions of formal learning theory. Nature 412, 43–48.

White, R., 1959. Motivation reconsidered: the concept of competence. Psychological Review 66, 297–333.

Wickens, J.R., 2009. Synaptic plasticity in the basal ganglia. Behavioural Brain Research 199, 119–128.

Williams, D., Lowe, G., 1972. Response contingent illumination change as a reinforcer in the rat. Animal Behaviour 20, 259 – 262.

Wise, R., 2004. Dopamine, learning and motivation. Nature Reviews Neuroscience 5, 483–494.

Wise, R., Rompre, P., 1989. Brain dopamine and reward. Annual Reviews of Psychology 40, 191–225.

Zweifel, L.S., Parker, J.G., Lobb, C.J., Rainwater, A., Wall, V.Z., Fadok, J.P., Darvas, M., Kim, M.J., Mizumori, S.J.Y., Paladini, C.A., Phillips, P.E.M., Palmiter, R.D., 2009. Disruption of nmdar-dependent burst firing by dopamine neurons provides selective assessment of phasic dopamine-dependent behavior. Proc Natl Acad Sci U S A 106, 7281–7288.