Intrinsic motivation signals for driving the acquisition of multiple tasks: A simulated robotic study

Vieri Giuliano Santucci (vieri.santucci@istc.cnr.it)
Ist. di Scienze e Tecnologie della Cognizione (ISTC), Consiglio Nazionale delle Ricerche (CNR)
Via San Martino della Battaglia 44, 00185 Roma, Italia
School of Computing and Mathematics, University of Plymouth
Plymouth PL4 8AA, United Kingdom

Gianluca Baldassarre (gianluca.baldassarre@istc.cnr.it)
Ist. di Scienze e Tecnologie della Cognizione (ISTC), Consiglio Nazionale delle Ricerche (CNR)
Via San Martino della Battaglia 44, 00185 Roma, Italia

Marco Mirolli (marco.mirolli@istc.cnr.it)
Ist. di Scienze e Tecnologie della Cognizione (ISTC), Consiglio Nazionale delle Ricerche (CNR)
Via San Martino della Battaglia 44, 00185 Roma, Italia

Abstract

Intrinsic Motivations (i.e motivations not connected to reward-related stimuli) drive humans and other biological agents to autonomously learn different skills in absence of any biological pressure or any assigned task. In this paper we investigate which is the best learning signal for driving the training of different tasks in a modular architecture controlling a simulated kinematic robotic arm that has to reach for different objects. We compare the performance of the system varying the Intrinsic Motivation signal and we show how a Task Predictor whose learning process is strictly connected to the competence of the system in the tasks is able to generate the most suitable signal for the autonomous learning of multiple skills.

Keywords: Intrinsic Motivations, Modular Architecture, Reinforcement Learning, Adaptive Behaviour, Simulated Robot, Non-Stationary Reward

Introduction

Biological agents are able to learn and cash multiple skills in order to use them whenever future situations will require those competences. More interestingly, humans and other mammals (e.g. rats and monkeys) are able to explore the environment discovering and learning new abilities not only on the basis of reward-related stimuli, but also on the basis of novel or unexpected neutral stimuli. The mechanisms underlying these kind of learning processes have been studied since 1950s both in human and animal psychology under the heading of "Intrinsic Motivations" (IMs) (e.g., White, 1959; Berlyne, 1960; Ryan & Deci, 2000). Recently, researchers have also begun to investigate the neural basis of such mechanisms both through experiments (e.g., Wittmann, Daw, Seymour, & Dolan, 2008; Duzel, Bunzeck, Guitart-Masip, & Duzel, 2010) and through computational models (e.g., Kakade & Dayan, 2002; Mirolli, Santucci, & Baldassarre, 2013) and nowadays IMs are becoming a central topic of research (Baldassarre & Mirolli, 2013).

Looking at the computational literature, different IM models have been proposed both as methods to improve the ability of artificial agents and as models of human and animal learning (Schmidhuber, 1991b; Kakade & Dayan, 2002; Barto, Singh, & Chantanez, 2004; Schembri, Mirolli, & Baldassarre, 2007b; Oudeyer, Kaplan, & Hafner, 2007; Santucci, Baldassarre, & Mirolli, 2010; Mirolli et al., 2013). IMs can be considered a useful tool to improve the implementation of more autonomous and more adaptive artificial agents. Most of the IM computational models are implemented within the framework of reinforcement learning (Sutton & Barto, 1998), In this framework, IMs are modelled as self-generated reward signals able to drive the learning of the agent without any assigned "extrinsic" reward for specific tasks. Following the work of Schmidhuber (1991a, 1991b), most of these models implement intrinsic reinforcements as learning signals based on the prediction error (or the improvement of the prediction error) of a predictor of future states.

However, it is still not clear which kind of IM signal is the most suitable for driving a system able to *acquire and cash different abilities to learn the largest number of skills in the shortest time.* To our knowledge, the only study dedicated to this important issue is our previous work (Santucci, Baldassarre, & Mirolli, 2012). In that work, we showed that when the prediction error (or the improvement of the prediction error) is related to the prediction of any possible future state (as proposed by Schmidhuber, 1991b, 2010) the system focuses on actions that simply maximise that error, thus improving the model of the world but without learning any particular ability. Differently, if the learning process is driven by the error in the prediction of a particular state (goal), the system focuses only on actions related to the specific target, building a task/goal-oriented model of the world able to foster competence acquisition. Furthermore, we analysed different mechanisms for the implementation of such a signal. The results showed that the best mechanism was not based on step-by-step predictions determined by the perceptive results of each movement, but on predictions about the achievement of the goal made at the beginning of every trial (i.e., when the system decides to pursue a particular goal). Indeed, only this mechanism guarantees a close coupling between the intrinsic

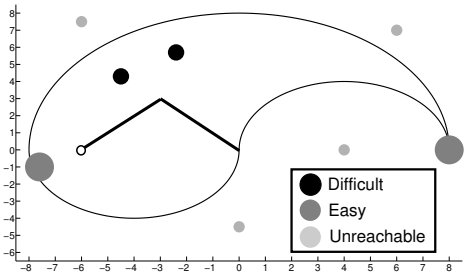

Figure 1: The two dimensional work space of the simulated kinematic robotic arm with the target objects. Small light-grey objects are unreachable by the arm

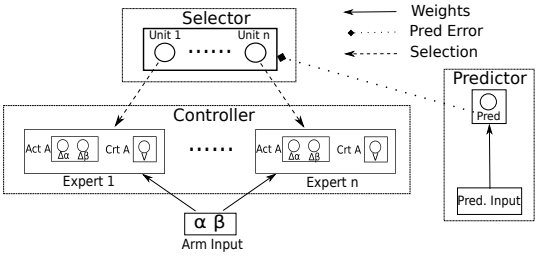

Figure 2: The modular architecture of the controller. The n actor-critic experts, the selector with n units and the predictor that generates the signal driving the selector, where n is the number of the tasks

reinforcement signal and the actual competence acquired by the system. Such a signal is present only when a skill is learnt while it fades away when the competence has been acquired and the predictor is able to anticipate the achievement of the goal.

However, that work presented two main limitations: 1) it analysed only the prediction error signal produced in the learning of a single task, not considering if such a signal could be useful to autonomously learn multiple skills; 2) the experiments took place in a simple grid-world scenario with discrete states and actions, while our interest is in animal, human and robotic learning which takes place in continuous states and actions.

In this paper, we cope with both these two limitations: 1) we investigate which is the best IM signal to drive the selection and acquisition of multiple skills in a hierarchical and modular system able to learn and cash different abilities in different modules; 2) we test the system in a robotic set-up with continuous states and actions.

Setup

The task and the simulated robot

The task (Fig. 1) consists in learning to reach circular objects placed within the work space of a simulated robotic arm. The system has to learn in the best way and possible shortest time the largest number of different skills, based on solely IM signals. There are 8 different objects: 2 are easy, 2 are difficult, and 4 are impossible to reach (we estimated the difficulty of different tasks by measuring the average time needed for an expert to reach 95% performance). This choice is due to the fact that in any moment an agent (be it an animal, a human or a robot) can try to learn a number of different abilities that typically vary considerably with respect to their learnability, including many (probably the vast majority) that are not learnable at all (consider, for example, trying to learn to reach the ceiling). For this reason, it is important for a learning system to avoid trying to learn unlearnable skills and to focus for the proper amount of time on those that can be learned.

The system is a simulated kinematic robot composed by a two degree-of-freedom arm with a "hand" that can reach for objects. The sensory system of the robot encodes the proprioception, i.e. the angles of the two joints of the arm. The output of the controller determines the displacement of the two joints in the next time step.

Architecture, input coding and learning

In order to avoid the well known problem of catastrophic forgetting (McCloskey & Cohen, 1989), for which previously learned abilities are disrupted by new ones, we needed an architecture where different abilities are cashed in different parts of the system. For this reason, the controller of the arm consists of a modular architecture (Fig. 2) composed by n experts (8 in this implementation, one for each task to be learned) and a selector that determines which task/expert will be trained. For simplicity, selecting a particular expert corresponds to selecting a task because each expert is rewarded only for reaching the associated object (this assumption is neutral with respect to the aim and results of the paper).

Each expert is a neural network implementation of the actor-critic architecture (Sutton & Barto, 1998) adapted to work with continuous state and action spaces (Doya, 2000). Both the critic and the actor of the experts receive as input the angles of the two joints of the arm, α and β (ranging in [0, 180]), coded through Gaussian radial basis functions (RBF) (Pouget & Snyder, 2000) in a two dimensional 10X10 grid. The evaluation of the critic of each expert (V) is a linear combination of the weighted sum of the respective input units. The actor of each expert has two output units fully connected with the input, with a logistic transfer function:

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

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

$$o_j^n = o_j + q$$

where q is a random value uniformly drawn in [-0.1; 0.1]. The resulting commands (ranging in [0; 1]) are remapped in [-25, 25] degrees and control the displacement of the related arm joint.

In each trial, the expert that controls the arm is trained through a TD reinforcement learning algorithm (Sutton, 1988). The TD-error δ is computed as:

$$\delta = (R_e^t + \gamma_k V^t) - V^{t-1}$$

where R_e^t is the reinforcement for the expert at time step t, V^t is the evaluation of the critic at time step t, and γ is a discount factor, set to 0.9. The reinforcement is set to 1 when the hand touches the object associated with the selected expert, 0 otherwise.

The connection weight w_{ni} of input unit i of critic n is updated in the standard way:

$$\Delta w_{ni} = \eta^c \delta_n a_i$$

where η^c is the learning rate of the critic, set to 0.08.

The weights of the actor are updated as follows (see (Schembri, Mirolli, & Baldassarre, 2007a)):

$$\Delta w_{ji} = \eta^a \delta_n (o_j^n - o_j)(o_j(1-o_j))a_i$$

where η^a is the learning rate of the actor, set to 0.8, $o_j^n - o_j$ is the discrepancy between the action executed by the system (with noise) and that produced by the controller, and $o_j(1-o_j)$ is the derivative of the sigmoid function.

The selector of the experts is composed by n units, one for each expert/task to be trained/learnt. At the beginning of every trial the selector determines, through a *softmax* selection rule (Sutton & Barto, 1998) with *temperature* set to 0.01 (we tested different values and selected the one that guarantees the best performance), which expert will control the arm during that trial. The activity of each unit of the selector is determined by a rule used to cope with n-armed bandit problems with non-stationary reward (Sutton & Barto, 1998)

$$K^{tr+1} = K^t r + \alpha[R^t r_s - K^t r]$$

where $K^t r$ is the activation of the unit corresponding to the selected expert during trial tr, α is a temporal parameter set to 0.35 (the value that we found to give the best performance) and $R^t r_s$ is the reinforcement signal obtained by the selector. More precisely, $R^t r_s$ is the intrinsic reinforcement that we want to analyse in order to find the most suitable signal for these kind of learning processes. It is determined by the error in predicting the achievement of the selected task: 0 when the arm is not able to reach the target object and 1-p when the object is reached, where p is the output of the predictor.

Since the focus of the present work is the comparison of different IM signals able to drive the selector of a hierarchical architecture in guiding the learning of multiple skills, we tested the system in different conditions varying for the implementation of the IM mechanism that generates the intrinsic reinforcement for the selector.

IM reinforcement signals

In this section we describe the different IM mechanisms implemented. Note that all the predictors also receive as input the information about which expert/task has been selected for controlling the robot during the current trial. The output of all the mechanisms is a prediction about the achievement of the target state related to the selected task.

State-Action Predictor (SAP) The input of this predictor is the same as the one provided to the majority of IM mechanisms implemented in literature (e.g., Schmidhuber, 1991b; Oudeyer et al., 2007; Santucci et al., 2010). It is composed by the present state (the two joints of the arm) and the planned action ($\Delta\alpha$ and $\Delta\beta$). Input is coded through RBF and training follows a standard delta rule.

State Predictor (SP) The SP is not widespread in the literature (a similar predictor can be found in Barto et al., 2004). In our previous work (Santucci et al., 2012) we found that this kind of mechanisms could be more closely coupled to the competence of the system than the SAP. Its input is composed only by the actual state of the arm and is coded through RBF. Training follows a standard delta rule.

SAP-TD SAP-TD has the same input of SAP but it is trained through a TD-learning algorithm with a discount factor set to 0.99. This type of predictor derives from the knowledge we acquired in previous works (Santucci et al., 2010; Mirolli et al., 2013), where we found that normal predictors have many problems in anticipating future state when working in continuous states and actions. Providing the predictors with a TD algorithm solves some of these problems (for a generalisation of TD-learning to general predictions, see Sutton & Tanner, 2005).

SP-TD As SAP-TD, this predictor is the TD version of SP.

Task Predictor (TP) This predictor is inspired by our previous work (Santucci et al., 2012) and it is similar to the mechanism present in Hart and Grupen (2013). It does not make step-by-step predictions as the previous mechanisms, but a single prediction on the achievement of the selected task at the beginning of the trial. The input of this predictor consists only in the information about the task/expert that has been selected (encoded in a n-long binary vector), and the predictor is trained through a standard delta rule. These characteristics should provide a complete coupling between the signal generated by the predictor and the competence of the system in achieving the tasks: the predictor has no further information and can learn to anticipate the achievement of the target state only when the agent has really acquired a high competence in the related skill.

Results

The experiments last 400,000 trials, each trial ending if the selected expert reaches its target object or after a time out of 20 time steps. Fig. 3 shows the number of trials needed by the different conditions to achieve a performance of 95% in

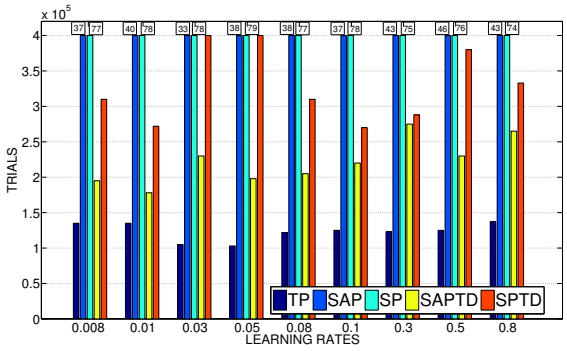

Figure 3: Number of trials needed by the different conditions to achieve a performance of 95% in the 4 learnable tasks (average results of 20 replications) with different learning rates for the predictors. If a condition has not reached the target performance, we report above the corresponding bar the average performance at the end of the simulation

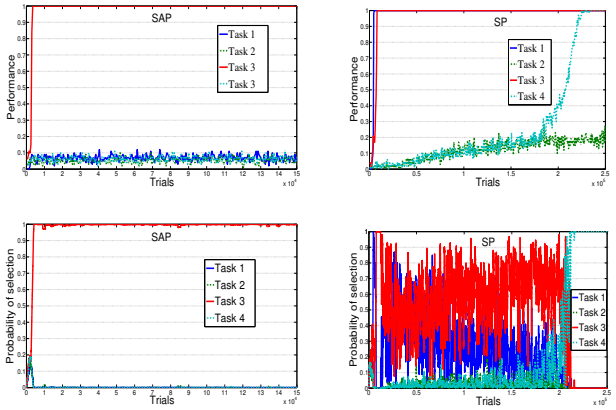

Figure 4: Performance on the 4 learnable tasks (top) and selection probability for the associated expert (bottom) in a representative replication of the SAP and SP conditions. SP figures have a different scale to better show the change in the behaviour during the simulation

the 4 learnable tasks (average results of 20 replications). For each condition we ran different experiments varying the value of the learning rate (LR) of the predictor (x-axis), because we wanted to be sure that results were not dependent on the use of a specific set of LRs.

As shown by the results, SAP and SP generate a signal that is not able to drive the system in acquiring a good performance, with SP achieving a better performance than SAP. The other conditions are able to reach the target performance within the time limit, but the efficiency of the learning process is different between them. SP-TD always takes the longest time in achieving the target performance and in two cases it reaches the 95% only at the end of the experiment. SAP-TD performs better than SP-TD, but it is sensitive to the value of the LR and often takes many trials to complete the tasks (especially with high LR values). Differently, TP is always the best performer, allowing the system to reach the target performance in less than 150,000 trials independently of the value of the LR.

To understand the causes of these results, for each condition we analysed the learning process of a representative replication (consider that all the replications have similar developments). In particular, we focused on data showing the selections of the different experts during time and the level of performance achieved on the tasks (for simplicity and clarity, we only show data related to the 4 learnable tasks within the first 150,000 trials). In this way we can check if the intrinsic reinforcement signal generated by the predictors is able to drive the selector in a proper way, following the actual competence acquired by the experts.

Fig. 4 shows data related to SAP and SP conditions. As described in the previous section, we knew that these kinds of predictors could have problems with continuous space and actions. In SAP the system starts focusing on Task 2 and because the predictor is not able to properly anticipate the achievement of the target object, the intrinsic reinforcement signal for that task is not cancelled and the selector is not able to switch to different experts and learn all the skills. In SP the system is able to learn the two easy tasks and, at a certain point, to shift to one of difficult tasks (Task 4): however, the predictor presents problems similar to those of SP condition and is not able to learn the remaining task.

Fig.5 shows data related to the other implemented conditions. While SAP and SP have the problem of not being able to properly anticipate the achievement of the target states, SAP-TD and SP-TD have the opposite problem: these mechanisms learn very fast to predict the reaching of the objects, even faster than the actual competence of the system in those tasks. The learning process of these predictors is not strictly coupled with the ability of the system to reach for the objects.

This is clear if we look at the figures related to SAP-TD: the selector reduces the selection of an expert before that expert has achieved an high performance in the related task. While this does not affect the learning of easier tasks (they need very few trails to be trained), this is a problem for the complex ones: because the predictor cancelled all the intrinsic reinforcement signals to the selector, experts start to be selected randomly, thus loosing time on previously learned (easy) tasks, which impairs the training of the difficult ones.

This is even more evident in data related to SP-TD, where the predictor drastically cancel the signals determining a random selections already from the early trials. The reason is that having only the actual state as input, the SP-TD mechanisms is able to generalise better than the SAP-TD: since from a single state there are many different actions that bring to the target object, SP-TD is able to generalise among all those actions, while SAP-TD has to learn to anticipate the achievement of the tasks using the different actions that it receives as input.

Differently, the TP mechanism is able to generate a signal

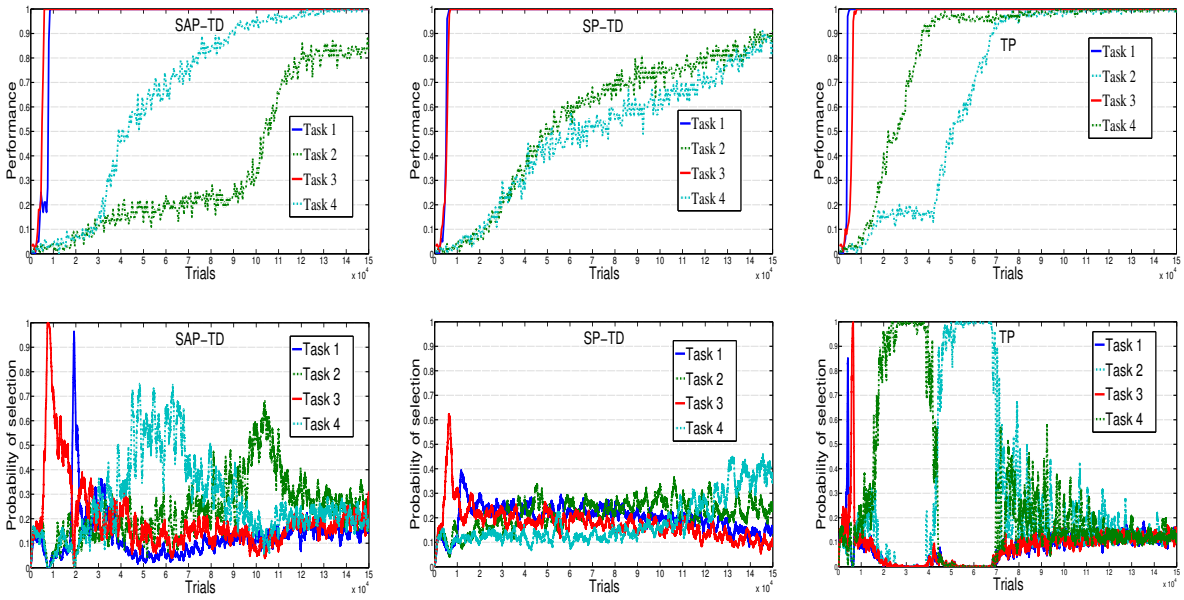

Figure 5: Data related to SAP-TD, SP-TD and TP conditions. See Fig.4 for description

that persists until the system has learnt the related task, providing a learning focus that moves from one skill to another. If we look at the selection of the experts and compare it to the performance in the related tasks we can see how the selector switches from the expert that is currently selecting only when its performance has achieved an high value. This guarantees the complete acquisition of competence and a reduction of the total amount of learning time, because no trials are spent on previously learnt tasks: obviously, when all the experts are trained, the system starts a random selection because no task produces reinforcements for the selector any more.

Conclusions and future works

In this paper we modelled the characteristic, typical of humans and other animals, of autonomously learning multiple skills on the basis of what have been called Intrinsic Motivations, focusing on which is the intrinsic reinforcement signal that is best suited for driving the acquisition of several skills in the shortest time. In particular, we implemented a modular architecture composed by different experts and a selector that determines which task/expert will be trained on the basis of an IM signal. We tested different IM mechanisms and compared the performance of the system in learning different tasks.

The experiments show that the best performance is achieved by the system whose selector is reinforced by a signal determined by the error in predicting the achievement of the target state only on the basis of the selected expert/task. The reason is that this is the best way to couple the intrinsic reinforcement signal driving the selection with the competence of the experts in achieving their goals. In particular, in this way the intrinsic reinforcement is present when the system is learning a new task, it is cancelled when the competence on that task has been learnt and reappears when a new, still-to-be-learnt task is encountered by the system.

Differently, the other implemented mechanisms (inspired by the current computational literature on IM) generate signals that determines lower performances. Some (SP and SAP) lack the complexity required to cope with continuous states and actions. In the other cases, the additional information provided to SP-TD (actual state of the system) and SAP-TD (actual state plus planned actions) made the resulting signals suitable for measuring the knowledge of the system in anticipating future states given the current information but less effective for driving the acquisition of several skills because those signals are less directly connected to the competence of the system and they can disappear before the agent has learnt the related task.

The system proposed in this paper may encounter problem in stochastic environments: if the achievement of a target state is probabilistic, the predictor will continue to make errors indefinitely. This means that the reinforcement will never be completely cancelled and the system may keep on trying to train a skill even if it cannot improve any more. In order to solve this problem, several systems (e.g. (Schmidhuber, 1991b; Oudeyer et al., 2007)) use the prediction improvement rather than the prediction error as IM signals. In future work we plan to merge the idea of prediction improvement with that of expert-based prediction proposed in this paper, as we did in Santucci et al. (2012) , in a robotic modular system as the one implemented in this work.

Another limit of the present work is that in the current experimental set up we decided that reaching for the objects was the task to learn for our system. But if we are interested in truly autonomous development, in future works we will need the agent to be able to self-determine its goals. This will prob-

ably require the introduction of other complementary intrinsic motivation signals that can make the agent autonomously select useful and achievable goals (for a discussion of how different IM mechanisms might serve different sub-functions even if the general function is driving the acquisition of skills, see Mirolli & Baldassarre, 2013).

Acknowledgments

This research was supported be the European Commission 7th FP, project IM-CLeVeR - Intrinsically Motivated Cumulative Learning Versatile Robots.

References

Baldassarre, G., & Mirolli, M. (Eds.). (2013). *Intrinsically motivated learning in natural and artificial systems*. Berlin: 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).

Berlyne, D. (1960). *Conflict, arousal and curiosity*. McGraw Hill, New York.

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

Duzel, E., Bunzeck, N., Guitart-Masip, M., & Duzel, S. (2010). Novelty-related motivation of anticipation and exploration by dopamine (nomad): implications for healthy aging. *Neuroscience Biobehavioural Review*, *34*(5), 660–669.

Hart, S., & Grupen, R. (2013). Intrinsically motivated affordance discovery and modeling. In G. Baldassarre & M. Mirolli (Eds.), *Intrinsically motivated learning in natural and artificial systems.* Berlin: Springer-Verlag.

Kakade, S., & Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks*, *15*(4-6), 549–559.

McCloskey, M., & Cohen, N. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 24, pp. 109–165). San Diego, CA: Academic Press.

Mirolli, M., & Baldassarre, G. (2013). Functions and mechanisms of intrinsic motivations: The knowledge vs. competence distinction. In G. Baldassarre & M. Mirolli (Eds.), *Intrinsically motivated learning in natural and artificial systems.* Berlin: Springer-Verlag.

Mirolli, M., Santucci, V. G., & Baldassarre, G. (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*(0), 40 - 51.

Oudeyer, P., Kaplan, F., & Hafner, V. (2007). Intrinsic motivation system for autonomous mental development. In *Ieee transactions on evolutionary computation* (Vol. 11, pp. 703–713).

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

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67.

Santucci, V., Baldassarre, G., & Mirolli, M. (2010). Biological cumulative learning through intrinsic motivations: A simulated robotic study on the development of visually-guided reaching. In B. Johansson, E. Sahin, & C. Balkenius (Eds.), *Proceedings of the tenth international conference on epigenetic robotics.* Lund University Cognitive Studies, Lund.

Santucci, V., Baldassarre, G., & Mirolli, M. (2012). Intrinsic motivation mechanisms for competence acquisition. In *Development and learning and epigenetic robotics (icdl), 2012 ieee international conference on* (p. 1-6).

Schembri, M., Mirolli, M., & Baldassarre, G. (2007a). Evolving childhood's length and learning parameters in an intrinsically motivated reinforcement learning robot. In L. Berthouze, G. Dhristiopher, M. Littman, H. Kozima, & C. Balkenius (Eds.), *Proceedings of the seventh international conference on epigenetic robotics* (pp. 141–148). Lund University Cognitive Studies, Lund.

Schembri, M., Mirolli, M., & Baldassarre, G. (2007b). Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot. In Y. Demiris, D. Mareschal, B. Scassellati, & J. Weng (Eds.), *Proceedings of the 6th international conference on development and learning* (p. E1-6). Imperial College, London.

Schmidhuber, J. (1991a). A possibility for implementing curiosity and boredom in model-building neural controllers. In J. Meyer & S. Wilson (Eds.), *Proceedings of the international conference on simulation of adaptive behavior: From animals to animats* (pp. 222–227). MIT Press/Bradford Books, Cambridge, Massachusetts/London, England.

Schmidhuber, J. (1991b). Curious model-building control system. In *Proceedings of international joint conference on neural networks* (Vol. 2, pp. 1458–1463). IEEE, Singapore.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *Autonomous Mental Development, IEEE Transactions on*, *2*(3), 230 -247.

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

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA.

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

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

Wittmann, B., Daw, N., Seymour, B., & Dolan, R. (2008). Striatal activity underlies novelty-based choice in humans. *Neuron*, *58*(6), 967–73.