Deciding Which Skill to Learn When: *Temporal-Difference Competence-Based Intrinsic Motivation* (TD-CB-IM)

Gianluca Baldassarre and Marco Mirolli

Laboratory of Computational Embodied Neuroscience, Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche
{gianluca.baldassarre,marco.mirolli}@istc.cnr.it

Abstract. Intrinsic motivations can be defined by contrasting them to extrinsic motivations. Extrinsic motivations are directed to drive the learning of behavior directed to satisfy basic needs related to the organisms' survival and reproduction. Intrinsic motivations, instead, are motivations that serve the evolutionary function of acquiring knowledge (e.g., the capacity to predict) and competence (i.e. the capacity to do) in the absence of extrinsic motivations: this knowledge and competence can be later exploited for producing behaviours that enhance biological fitness. Knowledge-based intrinsic motivation mechanisms (KB-IM), usable for guiding learning on the basis of the level or change of knowledge, have been widely modeled and studied. Instead, competence-based intrinsic motivation mechanisms (CB-IM), usable for guiding learning on the basis of the level or improvement of competence, have been much less investigated. The goal of this paper is twofold. First, it aims to clarify the nature and possible roles of CB-IM mechanisms for learning, in particular in relation to the cumulative acquisition of a repertoire of skills. Second, it aims to review a specific CB-IM mechanism, the *Temporal Difference Competence-Based Intrinsic Motivation* (TD-CB-IM). TD-CB-IM measures the improvement rate of skill acquisition on the basis of the Temporal-Difference learning signal (TD-error) that is used in several reinforcement learning (RL) models. The effectiveness of the mechanism is supported by reporting the results of experiments in which the TD-CB-IM mechanism is successfully exploited by a hierarchical RL model controlling a simulated navigating robot to decide when to train different skills in different environmental conditions.

1 Introduction

Intrinsic motivations (IM) are receiving an increasing attention for their potential to allow organisms and robots to acquire knowledge and skills cumulatively and in full autonomy (Baldassarre, 2011; Baldassarre and Mirolli, 2010; Barto et al., 2004b; Deci et al., 2001; Oudeyer and Kaplan, 2007; Schmidhuber, 2010).

As further explained in Sec. 2.1, intrinsic motivations allow organisms and robots to learn skills and knowledge in the absence of a guidance from extrinsic

motivations, that is motivations related to homeostatic drives such as hunger and thirst, or, in the case of robots, related to the tasks dictated by the user. An important class of IM, called "Competence-Based IM" (CB-IM) are related to measurements of the capacity to solve given tasks (Sec. 2.2). CB-IM can play various sub-functions within a cognitive system. Here we focus on a specific important computational challenge which can be briefly described as *deciding what to learn when* (Sec. 2.3). This challenge stems from the fact that an agent has to learn multiple skills so as to be capable of accomplishing several different goals, as it is often the case in animals and as it is becoming increasingly requested in robots. Specifically, the challenge resides in the fact that when an agent has to learn several different skills it has to decide to which skill dedicate its learning resources at each moment.

Here we review a specific CB-IM mechanism, called *Temporal Difference Competence-Based Intrinsic Motivation* (TD-CB-IM), that can solve this problem. The key idea is to focus learning on those skills that exhibit the maximum learning rate of competence. In particular, TD-CB-IM is based on the idea of using the TD-error learning signal that is at the heart of several reinforcement learning (RL) models (Sutton and Barto, 1998) as a reward signal for a higher-level RL component within a hierarchical system. [1]

The goal of the paper is twofold. First, it aims to clarify a specific computational problem, "deciding to when to learn what", that CB-IM can solve within autonomous learning agents. To this purpose, the paper will briefly review the overall function played by IM (Sec. 2.1), will present an analysis that contributes to clarify the difference existing between CB-IM and Knowledge-Based IM (KB-IM) (Sec. 2.2), and will illustrate the nature and importance of the deciding-when-to-learn-what problem. Second, it aims to review the TD-CB-IM. This algorithm was first presented and exploited in Schembri et al. (2007a,b,c). With respect to these papers here we will present only the basic results (Sec. 4), and we will instead present a deeper analysis of the nature and properties of TD-CB-IM (Sec. 4,5).

[1] Reinforcement learning models mimic the trial-and-error learning processes of animals directed to achieve an extrinsic reward, in particular those studied by behaviourist psychology with instrumental learning paradigms (Lieberman, 1993) (but the models are also used to capture some mechanisms of Pavlovian learning). One of the most biologically plausible RL models, the actor-critic RL model (Houk et al., 1995; Sutton and Barto, 1998), is formed by (a) an *actor*, which progressively learns to select actions so to maximize rewards, and (b) a *critic*, which progressively learns to assign an evaluation (an estimate of future rewards) to each state on the basis of the received rewards (the actor is trained to act so as to move the agent towards states with higher evaluations).

2 Competence-based intrinsic motivations

2.1 Functions of intrinsic motivations

Intrinsic motivations are usually defined by contrasting them to extrinsic motivations. Extrinsic motivations (EM) refer to homeostatic drives and other mechanisms that lead organisms to engage in an activity because it will eventually lead to a valuable outcome, such as food or water. Instead, intrinsic motivations (IM), an expression first used in psychology (Harlow, 1950; Ryan and Deci, 2000; White, 1959), refer to actions performed "for their own sake" rather than as a means to obtain a useful outcome.

From an evolutionary/adaptive perspective, extrinsic and intrinsic motivations have been proposed to have different functions (Baldassarre, 2011; note that in this paper we will focus only on the capacity of IM to produce learning signals, and not on their capacity to trigger/energize behavior). EM drive the performance and learning of behaviors directed to aid homeostatic regulations, and hence to improve the chances of survival and reproduction (i.e., improve biological fitness). Instead, IM have the function of leading organisms to learn complex behaviors, e.g. based on long chains of actions, that would never be acquired on the basis of extrinsic rewards alone. More precisely, IM generate learning signals (and motivate the execution of behaviors) that drive the learning of new knowledge and skills that only later are exploited to get extrinsic rewards, that is to improve biological fitness (cf. also Singh et al., 2010). The paradigmatic example of this is represented by children at play. Children spend their first years of life acquiring in a cumulative fashion a flexible repertoire of skills, and a wide knowledge of the world, mostly guided by intrinsic motivations (von Hofsten, 2007). Only later, in adult life, such skills are re-used to readily assemble complex behaviors directed to increase fitness. The importance of IM for children development is also manifested by the fact that most experiments of developmental psychology successfully leverage on IM to drive the behaviors they study as it is not possible to give direct instructions to young children and babies or to motivate them with EM.

2.2 Knowledge-based and competence-based intrinsic motivation mechanisms

In terms of mechanisms, IM learning signals are generated on the basis of "measurements" of the level of increase of knowledge and skills done directly in the brain (Baldassarre, 2011), or in the controller in the case of robots. Due to their function, the mechanisms producing IM learning signals usually cease to produce them once the knowledge or skill generating them have been acquired (Mirolli et al., subm; Santucci et al., 2010). This is different from EM learning signals that come back again and again with the homeostatic needs they are directed to satisfy.

Two main types of IM mechanisms can be identified: knowledge-based intrinsic motivations (KB-IM) and competence-based intrinsic motivations (CB-IM)

(Oudeyer and Kaplan, 2007, see also Mirolli and Baldassarre, 2012). Here with *knowledge* we will intend the "capacity to predict" (i.e., in the terminology of control theory, the capacity of *forward models* to anticipate future states based on current states and planned actions; note that in reality knowledge also includes other capabilities, e.g. to abstract and classify perceived stimuli, but we will not consider this here). With *competence* we will intend here the "capacity to do", that is to change the world in a certain way when the agent intends to do so (i.e., in the terminology of control theory, the capacity of *controllers* or *inverse models* to produce suitable actions on the basis of the pursued goal-state and the current state). Importantly, competence is dependent on *goals*: these are states, among all possible states, that the agent might consider as *desired states*, and hence will want work to achieve them. Knowledge-based IM mechanisms (here "KB-IM" for short) generate learning signals on the basis of measures of the level, or improvement, of the agent's capacity to predict. Instead, competence-based IM mechanisms (here "CB-IM" for short) generate learning signals based on measures of the level, or improvement, of the agent's capacity to achieve its goal-states.

Most computational research on IM has focussed on KB-IM. In a pioneering work, Schmidhuber (1991b) (see also Schmidhuber, 2012, for a review) proposed an agent endowed with a predictor, learning to predict the next state based on the current state and planned action, and a reinforcement-learning (RL) component, learning to produce actions based on a reward equal to the (absolute) value of the predictor's prediction error. The agent was capable of selecting actions that led the agent to explore new regions of the problem space and that led to a high error of the predictor, so fostering the improvement of the capabiltiy of prediction of the predictor itself. This system, however, was limited by the problem of getting stuck in regions of space and in activities that led to a high prediction error that could not be decreased due to the limitations of the predictor or the world intrinsic noise. To solve this problem, Schmidhuber (1991a) proposed another system that used a learning signal measuring the *improvement* of the prediction error to train the RL component. With various developments, in particular to make it applicable to real robots (e.g., Oudeyer et al., 2007; Oudeyer et al., 2012), this approach has become the IM system most used within the autonomous learning literature.

CB-IM have received much less attention than KB-IM. A possible reason is that, as we shall see, CB-IM seems to require complex hierarchical systems to be suitably investigated: this makes the investigations more challenging. Another possible reason is that measuring the improvement of competence is not easy: this is one of the main problems targeted in this paper and the TD-CB-IM is a possible solution to this problem.

To our knowledge, Barto's group (Barto et al., 2004a; Singh et al., 2005) has been the first to propose a model involving CB-IM. This model was based on the RL *option framework* (Sutton et al., 1999) and grid-world tests where the learning of the policy to achieve an extrinsic reward was supported by IM learning signals. In the model, the IM learning signals are generated as follows:

(a) the system is assigned different "salient states" to accomplish, and it creates a new skill (option) for each of them; (b) when pursuing one of these salient states, the system generates a reward in proportion to the probability of not achieving the target salient state from a state from which it was possible to achieve it. Although interesting, this approach suffers of a limitation analogous to the one of the KB-IM system of Schmidhuber (1991b) mentioned above: since the learning signal depends on the *level* of the skill, (probability of achieving the goal) rather than on its *improvement*, it can possibly lead the system to focus learning resources on goals that cannot be accomplished (but see Sec. 5 for a discussion of the possible conditions where this approach might work well).

2.3 Deciding when to learn what

Within the overall function of IM (guiding the acquisition of knowledge and skills in the absence of EM), CB-IM can play a specific important sub-function within systems that have a hierarchical architecture. Before looking closely to this sub-function, we want to stress the importance of hierarchical architectures for natural and artificial intelligent systems.

The key insight is that both animals and intelligent machines and robots can benefit of an organization of their behavior based on hierarchical architectures, which are usually either structurally or functionally modular. This is particularly important when they have to learn a multiple set of skills, as it is always the case in animals and in some cases for machines/robots.[2] The importance of hierarchical modularity is due to at least three reasons. First, the acquisition of multiple skills requires data structures and storing mechanisms that avoid the problem of catastrophic forgetting (McCloskey and Cohen, 1989). Catastrophic forgetting is relevant here as the acquisition of new skills can interfere and cause the forgetting of already acquired skills. Second, hierarchical modular architectures allow transferring skills and knowledge from one task to another when this is possible, thus enhancing the learning speed of new skills (see Taylor and Stone, 2009, for a review). Last, hierarchical modular architectures facilitate the re-use of the repertoire of previously acquired skills in the exploitation phase guided by extrinsic motivations, as they facilitate the composition of such skills to produce the needed action sequences (e.g., Hart and Grupen, 2011, 2012; Barto and Mahadevan, 2003, for a review; see also the example presented here).

For these reasons, hierarchy and modularity represent fundamental organizational principles of animal brains (see Meunier et al., 2010 for a review), and

[2] By "hierarchical" we mean that some components of the system, usually processing information at a more abstract level, exert an influence on other components, usually processing information at more detailed level. By "modular" we mean that different chunks of behavior are encoded in different portions of the system. Modularity can be either "structural", i.e. related to strong connections within groups of neurons and looser connections between groups, or "functional", e.g. leading to encode different chunks of behavior within different portions of a rather homogeneous system on the basis of specific self-organizing mechanisms.

they are exploited with success in several computational models (e.g., Baldassarre, 2001, 2002b; Caligiore et al., 2010; Hart and Grupen, 2011; Jacobs et al., 1991; Yao, 1999; see Barto and Mahadevan, 2003 for a review).

The development of hierarchical modular systems learning multiple skills presents at least three important challenges. The first challenge, discussed in Sec. 5 but not further expanded here, is related to the development of hierarchical modular architectures actually capable of storing multiple skills. The second challenge concerns the autonomous identification of tasks (goals) by the agent, again expanded in Sec. 5.

The third and last challenge, on which we focus in this paper, is related to the agent's decision on which task, among those autonomously identified, it should allocate its attention and learning resources in each period of its life. The idea proposed here is that *the agent should train the skills that have the highest rate of improvement*. There are at least to reason for this. First, the agent should not continue to focus learning processes on already acquired skills as this would lead it to waste time and learning resources. Second, learning of new skills can require the execution of previously acquired skills in order to create the conditions of the success of the new skills (e.g., a child has first to learn to look and reach/grasp a single block before having the possibility of learning to build towers formed by several blocks). This means that the agent should not focus on acquiring skills that, to be successful, require the execution of skills that have not yet been acquired. Behaviorally, an agent that follows these principles appears to focus learning resources on the *zone of proximal development* (Vygotsky, 1978) that lays between skills already acquired and skills too difficult to be acquired. The skills in the zone of proximal development are marked by a high acquisition rate which is instead low for already acquired and for too difficult to be acquired skills.

3 Mechanisms: Measuring competence improvement based on the TD-error learning signal

This section explains the TD-CB-IM mechanism, which allows measuring the competence improvement of a RL agent. The mechanism (presented for the first time in Schembri et al., 2007a,b,c) is based on the exploitation of the TD-error of a reinforcement learning component to infer the rate of improvement of the skills of the component. TD-CB-IM is explained here on the basis of a hierarchical architecture, guiding a simulated robot, which exploits such mechanism to decide which skill to train among a set of skills to be acquired.

Figure 1 shows the environment used for the task that the robot has to solve. The environment is a closed arena with a colored Red/Green/Blue pattern on the floor. The robot is a two-wheel simulated kinematic robot. The controller of the robot has to decide the translation speed and the rotation speed at each simulation step. The robot is endowed with a simplified RGB camera looking down towards the floor pattern and the input to the controller is formed by 2×6 abstract pixels for each RGB colour (in a real robot, each of these 12 abstract

pixels would be the average of the activation of a portion of the camera image pixels divided in 2×6 regular parts).

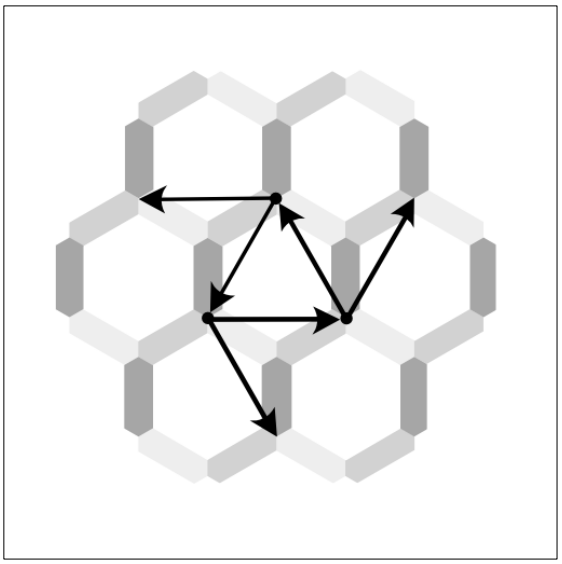


Fig. 1. The walled arena used to test the robot. The sides of the hexagons were coloured with blue (dark gray in the figure), red (gray) and green (light gray). The arrows represent the six sub-tasks: for each task, the tail and the head of the arrow indicate the start and the target positions, respectively. Reprinted with permission from Schembri et al. (2007c)

The task that the robot has to solve captures a typical situation that can be faced with IM. The life of the robot is composed of two phases called "childhood" and "adulthood". The robot performance is evaluated on the basis of how fast it learns to solve a given task during adulthood. The adult task is composed by six sub-tasks, and the robot's overall performance is an average of the performance in these sub-tasks (see Figure 1). In each sub-task the robot is set in a specific initial location of the arena and has to reach a specific target location (both the start and target locations are on a non-black portion of the environment). The robot has to learn a solution to each sub-task by suitably composing the skills acquired during childhood (see below). The robot gets an *extrinsic reward* of 1 when it reaches the target location of the pursued sub-task and 0 otherwise.

During childhood, the robot is not informed on the final task but it can freely explore the environment to acquire skills that can be used during adulthood to solve the final task. In particular, during childhood the robot has to acquire a repertoire of navigation skills (e.g., following a red or blue color, turning at a

certain color junction, etc.) based on: (a) the guidance of *reinforcers* components; (b) the TD-CB-IM mechanism explained below.

The reinforcers that guide learning during childhood are a set of two-layer feed-forward neural networks taking as input the input of the robot camera and giving as output a reward signal ranging in $[-1, +1]$. Importantly, the connection weights of the reinforcers are evolved with a genetic algorithm (GA) that uses the robot performance in the final task as the fitness function. The genetic algorithm generating the reinforcers mimics natural evolution, which generates the brain machinery of organisms that lead them to autonomously set goals. The overall (circular) evolutionary and learning cycle involving the system can hence be summarised as follows:

$$Evolution : GA \rightarrow Reinforcers$$
$$Childhood : Reinforcers + TD\text{–}CB\text{–}IM \rightarrow Skill\ learning$$
$$Adulthood : Extrinsic\ rewards \rightarrow Skill\ composition \rightarrow Fitness \rightarrow GA$$

Figure 2 shows the architecture of the model. The architecture is here described at a qualitative level, presenting only the critical formulas needed to explain TD-CB-IM (see Schembri et al., 2007b for other computational details on the model). The model is a two-level hierarchical reinforcement learning architecture. The higher level of the architecture is formed by a *selector* that learns to select the *experts* (3 in the experiments reported below) that form the lower-level of the architecture. Each expert can control the behavior of the robot and learns though reinforcement learning. During both adulthood and childhood, at each simulation cycle the selector selects one expert and the expert decides the robot's action. During childhood the selected expert also learns on the basis of the reinforcement signal produced by its reinforcer. During adulthood the experts do not learn but are only exploited by the selector. The selector, instead, learns in both phases of life: during adulthood it learns on the basis of the extrinsic reward related to the accomplishment of the final sub-tasks, whereas during childhood it learns on the basis of the TD-CB-IM mechanism. Now these processes are explained in more detail.

The selector and the experts are each constituted by an actor-critic reinforcement-learning model (Sutton and Barto, 1998), whose actor and critic are implemented as linear approximators (two-layer feed-forward neural networks). In addition each expert has also a reinforcer associated with it. The actor of each expert is a two-layer neural network that gets the camera image as input and has two sigmoidal output units with which it controls the rotation and translation speed of the robot (each output unit sets the center of a Gaussian function on the basis of which the actual command is randomly drawn to ensure exploration).

The critic of each expert is a two-layer neural network that gets the camera image as input and has one linear output unit with which it encodes the state evaluation depending on the expert's policy. During childhood (during adulthood the experts do not learn), two successive evaluations of the expert critic, together with the expert reinforcer's reward, are used to compute the *Temporal Difference*

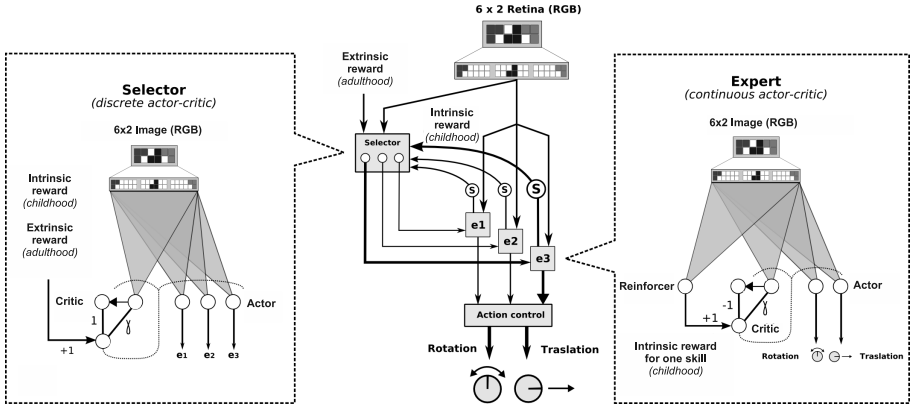


Fig. 2. The architecture of the model (centre) with a zoom on the components of the selector (left) and of one expert (right). Reprinted with permission from Schembri et al. (2007c)

error (TD-error) of the expert, as in the standard RL actor-critic model (Sutton and Barto, 1998):

$$TD_t^e = (r_t^e + \gamma v_t^e) - v_{t-1}^e \tag{1}$$

where TD_t^e is the TD-error, v_{t-1}^e and v_t^e are the two successive expert critic's evaluations, and r_t^e is the expert reinforcer's reward. The selected expert uses the learning signal TD^e to train its own critic component on the basis of the standard TD learning algorithm (Sutton and Barto, 1998). Moreover, the selected expert trains its actor with a delta rule that moves the output units activation towards values corresponding to the executed action (which includes exploratory noise) when $TD_t^e > 0$, and away from it when $TD_t^e < 0$ (the change is done in proportion to $|TD_t^e|$).

The selector actor is a two-layer neural network that gets the camera image as input and has a number of Sigmoidal output units equal to the number of experts. At each time-step, the activations of the output units are normalised so to sum to one, and are used as probabilities to select, with a winner-take-all competition, the expert that controls the motor system.

Even the selector critic is a two-layer neural network. The selector critic uses two different reinforcement signals during the agent's life to compute its TD-error. During adulthood, it computes the TD-error, TD_t^k, on the basis of the *extrinsic reward* r^k related to the pursued sub-task, k, and two succeeding evaluations related to it, v_{t-1}^k and v_t^k:

$$TD_t^k = (r_t^k + \gamma v_t^k) - v_{t-1}^k \tag{2}$$

This signal is then used to train the selector critic with the standard TD-learning rule, and to train the selector actor with a delta rule, to increase the probability

of selecting the expert that selects actions producing the highest TD_t^k at each state. In adulthood, at the beginning of each sub-task the selector actor and critic are reset to random weights as the policy and evaluation gradients they learn are different for each sub-task and for childhood.

Most importantly, during childhood the selector uses the TD-CB-IM mechanism to compute its TD-error, TD_t, and to learn. This mechanism is based on the use of the *TD-error of the selected expert*, namely TD_t^e, as a reinforcement:

$$TD_t = (TD_t^e + \gamma v_t) - v_{t-1} \tag{3}$$

The fact that during childhood the selector receives the TD-error of the selected expert as its reinforcement makes the selector learn, for each state, to *give control to the expert which has the maximum expected learning rate in such a state*. The reason of this is that the TD-error of an expert in a given state gives a measure of how much such expert learns in such a state. In fact, a positive TD-error means that the expert executed an action that is on average better than the actions previously performed in such state (viceversa if the TD-error is negative). If various experts are selected in a given state in successive experiences, the selector learns to produce an estimate of the average TD-error that can be obtained from that state on the basis of that selection: as in standard RL, this allows it to learn to improve its decision policy, in this case with respect to the selection of the experts that have to act (and learn) in the world.

An important feature of TD-CB-IM shows that it is indeed based on a measure of competence and not of knowledge. The expert's TD-error used as reward to train the selector (equation 3) is *not* used in absolute value ($|TD_t^e|$) as it is done in the case of the prediction errors used by KB-IM (see Schmidhuber, 2012). Indeed, even if TD-CB-IM is based on (the expert critics') prediction errors, it radically departs from KB-IM for two key reasons: (a) the prediction (of the expert critic) is about the *reward* (as produced by the *reward function*), not about *world states* (as produced by the *transition function*), hence the reward is related to the success of the expert in achieving a specific goal state, not to any state that the agent might experience; (b) the *sign* of the prediction error signal (the TD-error) indicates if and how-much the *competence of the expert actor* in achieving the goal is actually increasing (indeed, only if the TD-error is positive there is an improvement). Instead, in the case of KB-IM (e.g., Schmidhuber, 2012): (a) the prediction (of the predictor) is not about the rewards but about world states, which are neutral with respect to the skills to be acquired; (b) the sign of the prediction is not relevant as both positive and negative error signals indicate that the predictor has made an error, meaning that the *knowledge stored in such predictor* can improve.

The TD-CB-IM mechanism has other two important properties related to the fact that a second TD-learning processes, the one used by the selector, is used to estimate the TD-error of the experts. First, this implies that the TD-CB-IM is *prospective*. Indeed, the selector critic, for the properties of TD learning, learns to produce an estimation of the true evaluations, denoted here as v_t^{sa}, that is equal to the expected sum of the future discounted TD^e signals that can be

obtained from the current state by following the expert selection policy of the (current) selector actor (sa):

$$v_t^{sa} = E\left[TD_t^e + \gamma TD_{t+1}^e + \gamma^2 TD_{t+2}^e + ...\right] \quad (4)$$

where $E[.]$ is the expectation function. This implies that the selector learns to select an expert that not only has a high expected learning rate in the current state, but that also leads to states where such expert, or other experts, can learn the most.

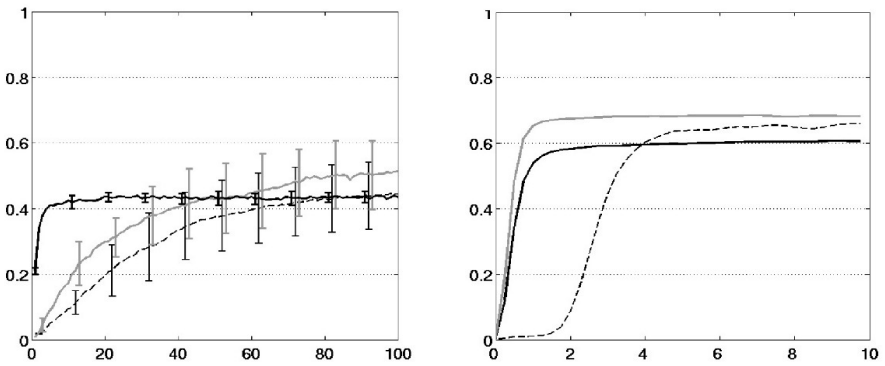
The second implication of the fact that the selector uses a TD-learning process to predict the expert TD-error is that the TD-CB-IM captures the idea of focusing learning within the zone of proximal development of the agent, i.e. on the experts that can learn the most in a certain developmental phase. Indeed, as the selector critic tends to learn the v_t associated to a give state in an incremental fashion, it tends to build up an estimation of the actual v^{sa} which averages out the strong noise usually affecting TD^e, and to capture its trend value for the visited states: this value is close to zero if the expert does not learn in such state, it tends to be positive if the expert is improving its competence, and it gets again close to zero when the expert has fully acquired the skill.

4 Results

This section reviews the basic results obtained in previous work (Schembri et al., 2007a,b,c), so to show how the system behaves.

In the experiments, the genetic algorithm rapidly reaches a high performance (Figure 3a). In this respect, the system showed to have a high degree of evolvability compared to other systems where other aspects of the architecture different from the reinforcers were evolved. For example, the architecture considered here took about 5 generations of the genetic algorithm to achieve high performance during adulthood whereas a system where the genetic algorithm directly searched the connection weights of the actors instead of the reinforcers took about 40 generations to achieve a comparable performance, and a systems where both the actors and the selector were evolved took about 80 generations (Figure 3a). This result is likely due to the fact that it is easier to evolve the criteria (reinforcers) for guiding learning of the behaviour needed to accomplish a certain task rather than directly evolving this behavior (see Schembri et al., 2007a for further details).

Figure 4a,b show the behavior of the robot during childhood. At the beginning of childhood (Figure 4a) the robot moves randomly in the arena as the selector selects random experts and the selected experts select random actions. With the progression of learning (Figure 4b), the robot acquires a very structured exploratory behavior. Indeed, the robot explores the environment following a regular pattern that allows it to experience states that are important for accomplishing the adulthood sub-tasks. In particular, during learning different experts specialize and acquire different skills. With different repetitions of the simulation with different random seeds the behaviors produced by the experts



(a) (b)

Fig. 3. (a) Evolution of the fitness of the best individuals (averaged over 10 runs) along 100 generations, for three conditions involving these versions of the system: evolved reinforcers, learning experts and learning selector (bold line); evolved experts and learning selector (gray line); evolved experts and evolved selector (dashed line). The graph also reports standard deviations. (b) Average performance during learning tests lasting 1,000,000 cycles for three conditions: evolved reinforcers, learning experts and learning selector (bold line); evolved experts and learning selector (gray line); simple learning expert reset before each adulthood task (dashed line). Curves refer to the average performance (normalized number of received rewards) over the 10 best individuals of 10 replications measured in 10 tests for each of the 6 tasks (i.e., average of 10x10x6 tests). Reprinted with permission from Schembri et al. (2007a)

can differ, although they are all very effective in adulthood. For example some experts make the robot follow a particular color while others make the robot follow two colors and to avoid the third one.

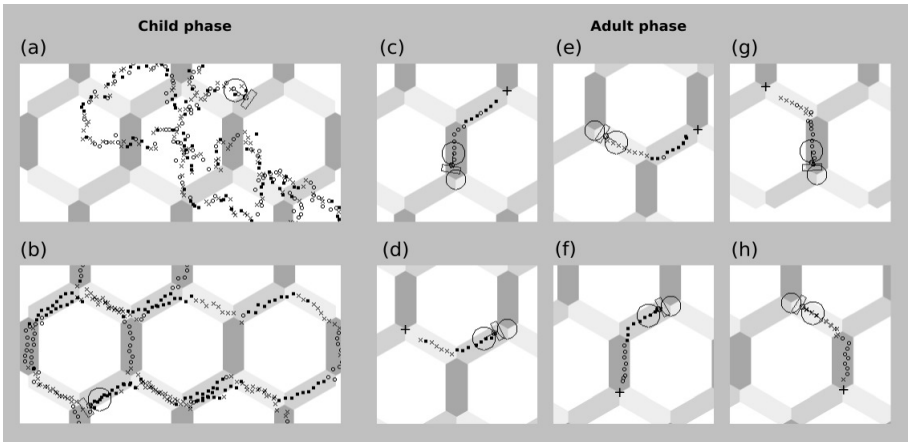


Fig. 4. Behavior of the robot: in all graphs crosses, empty circles, and full circles indicate the expert selected in that state (the marks are drawn every 10 steps). (a) Childhood: behavior exhibited by the robot at the beginning of development. (b) Childhood: behavior exhibited by the robot after the childhood learning. (c-h) Adulthood: behavior of the robot exhibited after learning the six adulthood tasks. Reprinted with permission from Schembri et al. (2007c)

Most notably, an important fact apparent from Figure 4b is that during childhood, thanks to the TD-CB-IM, the selector has learned to select a different expert for each different colour of the track. The selector acquires this capability as its reward is the TD-error signal of the experts, and this leads it to select, at each state, the expert that has the highest positive TD-error signal, that is the highest learning rate of the competence that allows the expert to accomplish the skill established by its reinforcer. Following this strategy, the selector allows the different experts to acquire different skills.

A second interesting fact apparent from Figure 4b is that the sequence with which the selector selects the experts creates a repetitive behavior based on the cyclic recall of sequences of skills. This is important as the agent is not artificially reset during childhood, so it has to procure the necessary experiences to learn. This behavior might be either the result of the structure of the environment, that favors cyclic behaviors, or the effect of an intended policy of the selector. In this respect, the prospective nature of the RL selector discussed in Sec. 2.1, for which the selector aims to maximize not only the next expert's TD-error but rather the sum of all future TD-errors (suitably discounted), might play an

important role in the acquisition of such capability. Although further research is needed to understand if and how TD-CB-IM can lead to find policies that solve the "problem of the reset" autonomously, this problem is clearly an important one to be solved to obtain a fully autonomous acquisition of skills.

The childhood learning process allows the robot to acquire a repertoire of skills suitable to solve the adulthood sub-tasks. The behavior of the robot involved in solving the six adulthood sub-tasks is shown in Figure 4c-h. During adulthood the selector learns quite rapidly to solve each of the six sub-tasks by assembling the skills acquired in childhood under the guidance of TD-CB-IM. If compared to a RL system that has to learn an adulthood task from scratch, the hierarchical system takes on average 4 times less (compare the bold and dashed lines of Figure 3b); for further details, see Schembri et al., 2007a). Indeed, based on the sub-task external reward, during adulthood the robot needs only to learn to select the suitable expert(s) for each colour and change it at colour junctions. Importantly, different sub-tasks can be learned by composing different sequences of skills: these skills represent readily available building blocks that do not need to be acquired from scratch. In other evolutionary runs the specialisation of the experts and their selection is more fuzzy/difficult to be described than the one shown in Figure 4 but the learning speed remains comparable.

5 Discussion and open challenges

5.1 Relevance of TD-CB-IM

This paper has first introduced the concept of intrinsic motivations (IM), and then it has focused on a particular class of them, namely competence-based IM (CB-IM). A first contribution of the paper has been the analysis of the relation existing between CB-IM and the problem of the cumulative acquisition of a repertoire of skills, in particular in relation to hierarchical architectures. A fundamental function that CB-IM can play in this context is to support the *autonomous* decision about which skill to train at each moment.

The second contribution of the paper has been to clarify the nature and functioning of a CB-IM mechanism, called "TD-CB-IM" and initially proposed in (Schembri et al., 2007a,b,c). TD-CB-IM exploits the TD-error learning signal used in most reinforcement learning (RL) models to measure the competence improvement of a RL agent (a mechanism similar to the one presented here, but based on the RL option framework and tested in a grid world, has been recently proposed by Stout and Barto, 2010). The effectiveness of TD-CB-IM has been shown by reviewing the core aspects of a hierarchical RL model where a higher-level RL selector has to learn to give control to a number of lower-level experts, themselves based on RL models, engaged in learning different tasks. Within this architecture, the key idea of the TD-CB-IM mechanism involves the use of the experts' TD-error as an index of the improvement of their competence. In particular, the TD-error of experts is used as an intrinsic reinforcement for the selector. This leads the selector to learn to select, at each state, the expert with the highest competence acquisition rate. Note that this mechanism of attribution

of responsibility (i.e., control and learning) is rather different from what done in other hierarchical RL models, for example in Doya et al. (2002), where the responsibility used to train RL expert modules is based on the capacity of the predictors of each module to predict the dynamics of the experienced portion of environment.

An important feature of the architecture made the mechanism prospective in the selection of experts. Indeed, as the TD-error of experts is given as reward to the selector which is itself a RL model, the selector learns to selects experts so as to maximize not only the immediate competence acquisition rate but also the future acquisition rate. For this reason, the selector learns to select the experts by taking in due consideration not only their (expected) learning rate, but also the possibilities of learning in the states that will be visited after their actions are executed. This gives a prospective nature to the selection of the experts, and leads to select sequences of skills that maximise the overall competence acquisition of the system.

The experiments reviewed here, related to a simulated robot endowed with a simplified camera, have shown how a hierarchical RL autonomous system, if endowed with a suitably system for self-identification of tasks (in this case a genetic algorithm), can benefit of the TD-CB-IM mechanism to acquire skills without the guidance of EM (i.e., in the case of robots, in complete autonomy from human intervention). Later the acquired skills can be readily composed to accomplish extrinsic tasks that would have required a long training to be accomplished, or that would have never been discovered (Vigorito and Barto, 2010). Although the test of the model presented here goes beyond the simple grid words usually used in in RL, it is nevertheless simplified: future investigations should aim to ascertain if and how the TD-CB-IM scales up to scenarios involving robots endowed with more complex sensory and motor systems.

5.2 Intrinsic motivations based on competence improvement versus competence level

An important issue concerns the advantages and disadvantage of CB-IM mechanisms based on measures of *levels of competence*, as the one proposed by Barto et al. (2004a); Singh et al. (2005), and those based on *competence improvement*, as the TD-CB-IM presented here. In Sec. 2.2 we have said that the former is potentially affected by the problem of focusing on tasks that cannot be learned because too difficult for the system or completely unsolvable. CB-IM mechanisms based on competence improvement such as TD-CB-IM allow overcoming this problem as they focus on a task only if the system can have a competence improvement on it: if there is no improvement due to the task difficulty or the limited potentiality of the learner, the engagement with the task ceases.

However, the learning signal used by TD-CB-IM, namely the TD-error of experts, is considerably affected by noise. The reason is that when the system has not fully acquired the capability of achieving its goals its stochastic policy continuously selects actions that can be either better or worse than the average, so the TD-error continuously shifts between positive and negative values. Hence,

the actual improvement of the competence of the system has to be captured as positive *average* TD-error. This fact might make the CB-IM mechanisms based on competence level preferable to those based on competence improvement in case the conditions for applying them are favorable, that is when: (a) the maximum level of achievable competence is known a-priori (this is needed to compute how far the actual competence is from the maximum one); (b) we are certain that the system has the necessary capability of acquiring a full competence in the task.

5.3 Intrinsic and extrinsic motivations

The model reviewed here has also clearly highlighted two issues important for IM: (a) the relation existing between extrinsic motivations (EM) and IM, and (b) their adaptive function for the survival and reproduction of organisms. In this respect, the model, by dividing the life of the agent in two distinct periods involving respectively IM and EM, has stressed how the primary adaptive function of IM is to guide the acquisition of skills and knowledge in the absence of EM learning signals. These skills and knowledge can then be exploited in succeeding phases of life to rapidly assemble behaviours that contribute to adaptation under the direct guidance of EM. Beyond the organisms' life, the success of this adaptation then guides the evolutionary process to improve the IM machinery that leads to acquire certain skills instead of others (e.g., specific reinforcers in the model), and that implements the TD-CB-IM (here hardwired). In this respect, the model captures some essential features of the evolutionary relation between IM and EM expanded at a theoretical level in Baldassarre (2011) and only briefly tackled in the original papers where TD-CB-IM was initially proposed (Schembri et al., 2007a,b,c). Another work (Singh et al., 2010) has presented an analysis and a model on these issues. This work agrees on the function of EM and IM discussed here (e.g., the model uses two distinct IM/EM learning phases to show the potential function of IM), but argues for a continuity existing between the two from evolutionary and computational perspectives.

Linked to the latter issue, we observe that the two distinct IM/EM phases of the model reviewed here are important for theoretical analysis, but in organisms IM and EM work at the same time (e.g., children are driven by both EM, such as those related to hunger and thirst, and IM, such as those related to curiosity and play). So, future work will have to investigate how the learning signals generated by EM and IM can be usefully arbitrated in situations where they tend to drive behaviour in different directions (see Kakade and Dayan, 2002 for a discussion about if an how novelty-related reinforcement signals might work together with long-term EM rewards). For example, how does a hungry child engaged in playing decide if continuing to play or looking for food?

The latter issues are relevant not only for the study of organisms, but also for robotics and machine learning. Indeed, also *autonomous* robots and machines have mechanisms equivalent to the EM mechanisms of organisms. In robots EM are first related to "survival", i.e. to physical integrity, energy maintenance, etc., which are an essential precondition for the robot correct functioning. In both

robots and machines, EM are also related to the user's requirements. Indeed, the "reproduction" of the robot/machine in several copies (eventually with variants) depends on the success of the robot/machine in accomplishing the users' requests. For this reason, the reward given to a RL robot by a user on the basis of a task useful for him/her can be considered as related to the EM of the robot. In this context, IM are can be used as means acquiring knowledge and skills before the user provides indications (i.e., EM learning signals). These previously-acquired knowledge and skills allow the machine/robot to learn to solve the users' tasks much faster (see, for example, Luciw et al., 2011).

5.4 Possible biological correspondents of TD-CB-IM

A last important problem, not mentioned so far, concerns the investigation of the possible biological correspondents of the TD-CB-IM. The investigation of the biological mechanisms possibly underlying IM is new but growing. A first proposal comes from Kakade and Dayan (2002). The authors propose that dopamine, one of the main neuromodulators used by brain for driving trial-and-error learning, carries information not only about primary rewards but also about *exploration bonuses*. These bonuses are quantities added to rewards or values to ensure appropriate exploration in new or changing environments. Another hypothesis is presented by Kaplan and Oudeyer (2007), who propose that the KB-IM signals generated by their model (Oudeyer et al., 2007) might correspond to tonic dopamine.

A elaborated theory has been proposed within the neuroscientific literature by Redgrave and Gurney (2006) (see also Redgrave et al., 2012). The idea is that bursts of dopamine signal the detection of sudden unexpected events, e.g. the unexpected onset of a light caused by the accidental pressure of a lever. These bursts of dopamine (DA), produced by the substantia nigra pars compacta (SNpc), are caused by the capacity of the superior colliculus (SC) to respond to unexpected luminance changes and, on this basis, to activate SNpc. In turn, the DA signal leads to increase the probability of execution of the actions that caused the event (putatively on the basis of trial-and-error learning processes implemented by the sensorimotor basal ganglia-cortical loops): this leads the agent to learn the experienced action-outcome contingencies and, on this basis, to later recall an action if the corresponding outcome becomes desirable (goal-based action recall; see Baldassarre et al. (subm) for a model of these processes).

A central aspect of this theory is that the dopaminergic burst is caused by apparently neutral events such as the light onset. A second important aspect is that the DA signal tends to disappear after a prolonged experience, putatively under the effect of a predictor of the phasic event that learns to progressively inhibits the sensory response (not to be confused with the inhibition inherent to the TD-error learning rule) (Mirolli et al., subm). These two aspects characterize the signal as an intrinsic rather than as an extrinsic signal. This theory has been proposed by contrasting it with the standard theories on phasic dopamine that claim that this signal corresponds to a *reward prediction error* equivalent to the TD-error of RL algorithms (e.g., Schultz, 2002). However, the two positions are

not necessarily in contrast, as suggested by Mirolli et al. (subm) on the basis of a computational model: dopaminergic signals might indeed correspond to TD-error signals that depend on both extrinsic rewards and to (intrinsic) reinforcements triggered by unexpected event, and thus might have the function of driving both reward maximization and action discovery and acquisition.

Paralleling the fact that they have been less investigated with computational models, we still lack a hypotheses on the possible brain correspondents of CB-IM. Based on its features, however, we can here try to speculate a possible biological implementation of the TD-CB-IM and propose an hypothesis that builds on the theory of Redgrave and Gurney (2006) reviewed above (cf. also Mirolli and Baldassarre (2012)). In this respect, an appealing feature of TD-CB-IM is that it relies upon the standard TD-error signal used by most RL models which, as mentioned above, putatively corresponds to phasic dopamine signals in brain (encoding the TD-error related to either a primary reward or to a sensory prediction error). This paves the way to search a higher-level RL system in the brain that: (a) has the capacity to select lower level actor-critic components capable of implementing actions; and (b) receives the TD-error from such lower level actor-critic components and uses it in place of the primary reward. We put forward the hypothesis that such two components of the system might be implemented respectively by: (a) the loops formed by medial and ventral BG (mvBG) and dorsolateral prefrontal cortex (dlPFC) which have been proposed to implement higher-level RL processes captured by hierarchical RL models (Baldassarre, 2002a; Botvinick et al., 2008); (b) the striosomes of ventral BG (vBG) and their connections to the ventral tegmental area (VAT), which form a second important dopaminergic system beyond SNpc. The idea would be that mvBG-dlPFC implement the actor of the selector of the model presented here: this would play the function of learning to select large chunks of behaviour (experts) implemented at a lower level within the sensorimotor BG-cortical loops. Instead, the vBG-VTA system, reached directly or indirectly by the DA caused by the lower level sensorimotor loops, would implement the critic of the selector: this would play the function of learning to predict the learning rate of the lower level experts as in the model presented here.

5.5 Open problems related to CB-IM

We see at least three open problems related to the use of CB-IM. The first is the importance of developing more powerful *hierarchical RL models* capable of storing multiple skills while avoiding catastrophic forgetting, exploiting information transfer between different tasks, and composing skills to build more complex behaviors. Various proposals already exist to face this problem: see Barto and Mahadevan (2003), for a review on hierarchical RL systems; Taylor and Stone (2009), for a review on the issue of transfer of information between tasks; Vigorito and Barto (2010), Hart and Grupen (2011), and Schembri et al. (2007a,b,c) for examples of models involving hierarchical/modular architectures and IM; Elfwing et al. (2007) for a model that has some resemblance to the one presented here and that uses evolutionary techniques to search a hierarchy

that minimizes the number of primitive subtasks that are needed for each type of problem. However, further advancements in this field are required to further develop CB-IM mechanisms.

The second open problem is related to the *autonomous identification of tasks/goals*, an important prerequisite for the functioning of the TD-CB-IM mechanism. The idea here is that a skill aims to accomplish a given *task*, i.e. to accomplish a given final state (*goal*) in the environment or in the body-environment relation (e.g., "reach and get the hand in contact with a visible object" or "grasp and carry the visible object from point A to point B in space"). When tasks and goals cannot be derived from EM, they must be found autonomously through IM. CB-IM mechanisms can play the sub-function of "deciding when to learn what" only if there are several different "whats" to be learned. The pioneering work of Singh et al. (2005) circumvented the problem: the tasks were hardwired by defining a limited sub-set of "salient states", among all states that the agent could experience, that defined the termination states of the options to be created. The work of Schembri et al. (2007c) reviewed here solved the problem by having a genetic algorithm find reinforcers that defined the skills to be acquired by the experts on the basis of the performance of the system on the adult's task. Another solution to the problem is based on the identification of critical states having a high frequency of visits during the solution of several different tasks (e.g., the door passage connecting two environments; McGovern and Barto, 2001; Pickett and Barto, 2002; Thrun and Schwartz, 1995). Yet another possible solution, which has however never been applied to autonomous skills discovery, might be based on the idea that relevant states are those where the agent is maximally "empowered", that is states where its actions can lead to explore the maximum number of future states (Jung et al., 2011; Klyubin et al., 2005). An important aspect of this problem is the relation between the need to seek the tasks to learn and the capacity of the agent to understand what it can learn and what is beyond its learning possibilities. For this reason, we think that mechanisms such as the TD-CB-IM might play an important role in the solution of the problem related to the autonomous search of tasks. Even KB-IM might play an important role in the autonomous finding of goals, as they allow the isolation of states of the world that are interesting and potentially relevant for the agent. In any case, the autonomous definition of useful goal-states remains one of the most important open problems for designing systems that are capable of undergoing a prolonged autonomous accumulation of competence with the guidance of IM.

A last problem related to TD-CB-IM is the issue of scalability. Here the mechanism was tested with only few experts/goals (e.g. three). In reality, organisms have to learn a large number of tasks. A similar condition might be encountered by future robots. The open problem is hence: would the TD-CB-IM work if used to learn a multitude of tasks/goals? The TD-CB-IM mechanism might have problems as its decision on when to learn what is based on a sampling of the competence improvement for each task: it does not seem efficient to continuously sample all tasks before knowing in which task the learning rate is highest. However, consider that *any* CB-IM or KB-IM mechanism used to ac-

quire several skills would encounter this problem. This suggests to look at more general solutions. This problem is closely related to the problem regarding the autonomous identification of goals. Indeed, these two problems might actually be the same: what is the goal the system should try engage with in each moment? In this respect, a possible solution might be that only *new goals discovered by chance*, e.g. while performing previously acquired skills in new conditions, are considered for learning: this would greatly restrict the attention to goals which are new but that are within the agent's capabilities. In this condition, CB-IM mechanisms such as TD-CB-IM might be used to continue to engage with the discovered goal only if the competence acquisition rate is above a certain level, rather than to select goals among the (possibly numerous) ones that are given externally as it happens in the artificial conditions considered here.

Acknowledgements

This research has received funds from the European Commission 7th Framework Programme (FP7/2007-2013), "Challenge 2 - Cognitive Systems, Interaction, Robotics", Grant Agreement No. ICT-IP-231722, Project "IM-CLeVeR - Intrinsically Motivated Cumulative Learning Versatile Robots".

Bibliography

Baldassarre, G. (2001). A modular neural-network model of the basal ganglia's role in learning and selecting motor behaviours. In Altmann, E. M., Cleermans, A., Schunn, C. D., and Gray, W. D., editors, *Proceedings of the Fourth International Conference on Cognitive Modeling (ICCM2001)*, pages 37–42. Lawrence Erlbaum, Mahwah, NJ.

Baldassarre, G. (2002a). A modular neural-network model of the basal ganglia's role in learning and selecting motor behaviours. *Journal of Cognitive Systems Research*, 3(2):5–13. Special Issue Dynamic and Recurrent Neural Networks.

Baldassarre, G. (2002b). *Planning with neural networks and reinforcement learning*. Phd thesis, Computer Science Department, University of Essex, Colchester, UK.

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., and Nagai, Y., editors, *Proceedings of the International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob-2011)*, pages E1–8. IEEE, Piscataway, NJ.

Baldassarre, G., Mannella, F., Fiore, V. G., Redgrave, P., Gurney, K., and Mirolli, M. (subm). Intrinsically motivated action-outcome learning and goal-based action recall: A system-level bio-constrained computational model. *Neural Networks*.

Baldassarre, G. and Mirolli, M. (2010). What are the key open challenges for understanding the autonomous cumulative learning of skills? The Newsletters of the Autonomous Mental Development Technical committee (IEEE CIS AMD Newsletters), volume 7 (1), page 11.

Barto, A., Singh, S., and Chentanez, N. (2004a). Intrinsically motivated learning of hierarchical collections of skills. In *International Conference on Developmental Learning (ICDL2004)*, Piscataway, NJ. IEEE.

Barto, A. G. and Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 13(4):341–379.

Barto, A. G., Singh, S., and Chentanez, N. (2004b). Intrinsically motivated learning of hierarchical collections of skills. In *International Conference on Developmental Learning (ICDL2004)*, Piscataway, NJ. IEEE.

Botvinick, M. M., Niv, Y., and Barto, A. (2008). Hierarchically organized behavior and its neural foundations: A reinforcement-learning perspective. *Cognition*.

Caligiore, D., Mirolli, M., Parisi, D., and Baldassarre, G. (2010). A bio-inspired hierarchical reinforcement learning architecture for modeling learning of multiple skills with continuous state and actions. In Kuipers, B., Shultz, T., Stoytchev, A., and Yu, C., editors, *IEEE International Conference on Development and Learning (ICDL2010)*. IEEE, Piscataway, NJ.

Deci, E., Koestner, R., and Ryan, R. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research*, 71(1):1–27.

Doya, K., Samejima, K., Katagiri, K.-I., and Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural Comput*, 14(6):1347–1369.

Elfwing, S., Uchibe, E., Doya, K., and Christensen, H. (2007). Evolutionary development of hierarchical learning structures. *IEEE Transactions on Evolutionary Computation*, 11(2):249–264.

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.

Hart, S. and Grupen, R. (2011). Learning generalizable control programs. *IEEE Transactions on Autonomous Mental Development*, 3(1):216–231.

Hart, S. and Grupen, R. (2012). Intrinsically motivated affordance discovery and modeling. In Baldassarre, G. and Mirolli, M., editors, *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer-Verlag, Berlin. (this volume).

Houk, J. C., Adams, J. L., and Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals ghat predict reinforcement. In Houk, J. C., Davids, J. L., and Beiser, D. G., editors, *Models of Information Processing in the Basal Ganglia*, pages 249–270. The MIT Press, Cambridge, MA.

Jacobs, R., Jordan, M., Nowlan, S., and Hinton, G. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Jung, T., Polani, D., and Stone, P. (2011). Empowerment for continuous agent-environment systems. *Adaptive Behavior*, 19(1):16–39.

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

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

Klyubin, A., Polani, D., and Nehaniv, C. (2005). Empowerment: A universal agent-centric measure of control. In *The 2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 128–135.

Lieberman, D. A. (1993). *Learning, Behaviour and Cognition*. Brooks/Cole.

Luciw, M., Graziano, V., Ring, M., and Schmidhuber, J. (2011). Artificial curiosity with planning for autonomous perceptual and cognitive development. In Cangelosi, A., Triesch, J., Fasel, I., Rohlfing, K., Nori, F., Oudeyer, P.-Y., Schlesinger, M., and Nagai, Y., editors, *IEEE International Conference on Development and Learning (ICDL2011)*, pages E1–8. IEEE, Piscataway, NJ.

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

McGovern, A. and Barto, A. (2001). Automatic discovery of subgoals in reinforcement learning using diverse density. Technial report of the faculty publication series, University of Massachusetts – Amherst, Computer Science Department.

Meunier, D., Lambiotte, R., and Bullmore, E. T. (2010). Modular and hierarchically modular organization of brain networks. *Frontiers in Neuroscience*, 4:200.

Mirolli, M. and Baldassarre, G. (2012). Functions and mechanisms of intrinsic motivations: The knowledge versus competence distinction. In Baldassarre, G. and Mirolli, M., editors, *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer-Verlag, Berlin. (this volume).

Mirolli, M., Santucci, V. G., and Baldassarre, G. (subm). Phasic dopamine as a prediction error of intrinsic and extrinsic reinforcement driving both action acquisition and reward maximization: A simulated robotic study. *Neural Networks*.

Oudeyer, P.-Y., Banares, A., and Frédéric, K. (2012). Intrinsically motivated learning of real world sensorimotor skills with developmental constraints. In Baldassarre, G. and Mirolli, M., editors, *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer-Verlag, Berlin. (this volume).

Oudeyer, P.-Y. and Kaplan, F. (2007). What is intrinsic motivation? a typology of computational approaches. *Frontiers in Neurorobotics*, 1:6.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions in Evolutionary Computation*, 11(2):265–286.

Pickett, M. and Barto, A. (2002). Policyblocks: An algorithm for creating useful macro-actions in reinforcement learning. In Sammut, C. and Hoffmann, A. G., editors, *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 506–513. Morgan Kaufmann, San Francisco, CA.

Redgrave, P. and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nature Review Neuroscience*, 7(12):967–975.

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

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

Santucci, V. G., Baldassarre, G., and Mirolli, M. (2010). Biological cumulative learning through intrinsic motivations: a simulated robotic study on development of visually-guided reaching. In Johansson, B., Sahin, E., and Balkenius, C., editors, *Proceedings of the Tenth International Conference on Epigenetic Robotics (EpiRob2010)*, pages 121–128. Lund University, Lund, Sweden.

Schembri, M., Mirolli, M., and Baldassarre, G. (2007a). Evolution and learning in an intrinsically motivated reinforcement learning robot. In Almeida e Costa Fernando, Rocha, L. M., Costa, E., Harvey, I., and Coutinho, A., editors, *Advances in Artificial Life. Proceedings of the 9th European Conference on Artificial Life (ECAL2007)*, volume 4648 of *Lecture Notes in Artificial Intelligence*, pages 294–333. Springer Verlag, Berlin.

Schembri, M., Mirolli, M., and Baldassarre, G. (2007b). Evolving childhood's length and learning parameters in an intrinsically motivated reinforcement learning robot. In Berthouze, L., Dhristiopher, P. G., Littman, M., Kozima, H., and Balkenius, C., editors, *Proceedings of the Seventh International Conference on Epigenetic Robotics*, volume 134, pages 141–148. Lund University, Lund.

Schembri, M., Mirolli, M., and Baldassarre, G. (2007c). Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot. In Demiris, Y., Mareschal, D., Scassellati, B., and Weng, J., editors, *Proceedings of the 6th International Conference on Development and Learning*, pages E1–6, Piscataway, NJ. IEEE.

Schmidhuber, J. (1991a). Curious model-building control systems. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2, pages 1458–1463.

Schmidhuber, J. (1991b). A possibility for implementing curiosity and boredom in model-building neural controllers. In Meyer, J.-A. and Wilson, S., editors, *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, pages 222–227, Cambridge, MA. MIT Press.

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

Schmidhuber, J. (2012). Maximizing fun by creating data with easily reducible subjective complexity. In Baldassarre, G. and Mirolli, M., editors, *Intrinsically*

motivated learning in natural and artificial systems. Springer-Verlag, Berlin. (this volume).

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, 36(2):241–263.

Singh, S., Barto, A., and Chentanez, N. (2005). Intrinsically motivated reinforcement learning. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, Cambridge, MA. The MIT Press.

Singh, S., Lewis, R., Barto, A., and Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):70–82.

Stout, A. and Barto, A. G. (2010). Competence progress intrinsic motivation. In Kuipers, B., Shultz, T., Stoytchev, A., and Yu, C., editors, *IEEE International Conference on Development and Learning (ICDL2010)*. IEEE, Piscataway, NJ.

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

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

Taylor, M. and Stone, P. (2009). Transfer learning for reinforcement learning domains: A survey. *The Journal of Machine Learning Research*, 10:1633–1685.

Thrun, S. and Schwartz, A. (1995). Finding structure in reinforcement learning. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in neural information processing systems 7 (NIPS1994)*, pages 385–392. MIT Press, Cambridge, MA.

Vigorito, C. and Barto, A. (2010). Intrinsically motivated hierarchical skill learning in structured environments. *IEEE Transactions on Autonomous Mental Development*, 2(2):132–143.

von Hofsten, C. (2007). Action in development. *Dev Sci*, 10(1):54–60.

Vygotsky, L. S. (1978). The development of higher psychological processes. *Mind in Society*.

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

Yao, X. (1999). Evolving artificial neural networks. In *Proceedings of the IEEE*, volume 87, pages 1423–1447. IEEE, Piscataway, NJ.