Intrinsic motivation mechanisms for competence acquisition

Vieri G. Santucci
Ist. di Scienze e Tecnologie della Cognizione (ISTC)
Laboratory of comput. embodied neuroscience (LOCEN)
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
vieri.santucci@istc.cnr.it

Gianluca Baldassarre and Marco Mirolli
Ist. di Scienze e Tecnologie della Cognizione (ISTC)
Laboratory of comput. embodied neuroscience (LOCEN)
Consiglio Nazionale delle Ricerche (CNR)
Via San Martino della Battaglia 44, 00185 Roma, Italia
{marco.mirolli,gianluca.baldassarre}@istc.cnr.it

***Abstract*—In the computational literature intrinsic motivations have been connected to the possibility of developing more autonomous and versatile agents. Despite the growing theoretical understanding of the distinction between functions and mechanisms of intrinsic motivations, the implications of the distinction have not been exploited in specific models. In particular, knowledge-based mechanisms are widely used to implement intrinsic motivations signals for the acquisition of competences, leading to inappropriate learning signals. In this paper we analyse and compare, with the support of simple grid-world simulations, different mechanisms that can be used to implement competence acquisition through intrinsic motivations, describing their limits and strengths and highlighting which features are best suited for the acquisition of competence.**

I. Introduction

The concept of intrinsic motivations (IM) [1] has been introduced during the 1950s in animal psychology to explain experimental data (e.g. [2], [3], [4]) showing how stimuli not related to (extrinsic) primary drives were able to provide a reinforcing value suitable for the acquisition of instrumental responses.

In the computational literature (e.g. [5], [6], [7], [8], [9]) IM has been linked to the possibility of building autonomous and versatile agents (especially reinforcement learning – RL – agents) that are able to self-generate reward signals. In particular, IM signals can drive the learning of new knowledge and skills that are not immediately extrinsically reinforced (i.e. not directly related to main tasks or fitness pressures, for organisms, or to the user's needs, for robots and intelligent machines); later, the acquired abilities will be exploited to obtain extrinsic rewards [10]. To this purpose, IM signals have to be *transient*: they have to persist during the learning process and disappear when it is completed, so that the system can move to acquire new knowledge and skills [11].

Depending on the *mechanisms* they rely upon, IM have been divided into two main categories: knowledge-based IM (KB-IM) mechanisms and competence-based IM (CB-IM) mechanisms [12]. KB-IM mechanisms generate learning signals based on the acquisition of *knowledge*, for example based on the improvement of the prediction capability of a predictor (i.e., a forward model of the world). CB-IM mechanisms, instead, generate learning signals based on the acquisition of *competence*, for example based on the capacity of achieving a certain desired state (e.g., the capacity of an inverse model or of a state-action controller to achieve a goal state). Importantly, KB-IM and CB-IM mechanisms can be *both* used for two distinct functions [13]: (a) the acquisition of knowledge, for example the acquisition of better prediction capabilities or the formation of object representations; (b) the acquisition of competence, i.e. the capacity to act so as to achieve a state of the world when it becomes desirable.

Despite the growing theoretical understanding of the differences between functions and mechanisms of IM (e.g. [13]), their implications have not been fully exploited in specific models. In particular, as underlined in [14], KB mechanisms are widely used to implement IM signals for the acquisition of competence, leading to inappropriate learning signals.

In this paper we analyse some of the main mechanisms used to implement IM signals for the acquisition of competences, describing their limits and strengths and highlighting which features are best suited for the acquisition of competence. We support this analysis with simple grid-world simulations that compare the different mechanisms and allow us to capture the principles on which those mechanisms are built.

II. KB mechanisms

One of the pioneering works on computational IM is presented in [15]. In this work a RL agent is driven to explore the environment by a self-generated reward signal. The mechanism used to implement this signal is a predictor that models the state transition function: it receives as input the current state together with the planned action and learns to predict the next state perceived by the agent. The prediction error (PE), i.e. the mismatch between the prediction and the next state, determines the size of the reward signal given to the system.

However, this kind of signal cannot cope with unpredictable situations: if it is not possible to anticipate what will be the future state, or the predictor has limited computational capabilities, the prediction errors will not disappear thus providing reward signals to the system that will so get stuck. To avoid this problem, in [5] a second mechanism was introduced that generates a reward signal based on the *prediction error improvement* (PEI): the signal will be positive only with an improving predictor while it will approach zero with unpredictable or too difficult portions of the environment.

A. Testing KB mechanisms for competence acquisition

KB-IM mechanisms generate signals based on the knowledge of the system. In many IM systems this knowledge is the capacity of a predictor to anticipate future states. We tested the behaviour of a system driven by the signal of a KB-IM mechanism similar to the one presented in [5] to verify if it could drive the acquisition of competence. Note that such a mechanism is one of the most used IM mechanisms within the ICDL community. The test used is based on a very simple simulated environment: a 1x10 grid where a simulated agent can move leftward and rightward for 5000 time steps. A performed action leads to its intended state with 0.95 probability and to the opposite one with 0.05 probability.

The agent is an actor-critic RL model trained through the standard TD-learning rule [16]. The input to the system is the actual position of the agent coded through a vector of 10 binary units (a value of 1 is given to the unit corresponding to the position of the agent, 0 to the others). All the weights of the system are initialised to 0.

The critic has a single linear output unit, fully connected with the input vector, which evaluates the current state. The learning rate is set to 0.02. The actor has two linear output units, fully connected with the input vector, which determine the displacement of the agent on the next time step: if the activation of the first unit is higher than the one of the second unit the agent moves leftward, otherwise rightward. A random noise value uniformly drawn in [-0.2, 0.2] is added to the activation of the two units. The learning rate is set to 0.8.

The IM reward signal is implemented as the PEI of a one-step-ahead predictor of future states. The input to the predictor encodes a combination of the actual position and the planned action of the agent with a 2x10 binary unit matrix. The output of the predictor is a vector of 10 linear units, fully connected with the input units. The weights of the predictor are updated through a standard delta rule where the position observed after action execution is used as teaching input. The learning rate is set to 0.08.

The PE is calculated as the absolute value of the mismatch between the prediction vector (PV) at time $t-1$ and the actual state vector (S) at time t (j is the vector element index):

$$PE_t = \sum_{j=1}^{10} |S_{j_t} - PV_{j_{t-1}}|$$

The PEI at time t is calculated as the difference between average PE calculated over a period T of 5 time steps:

$$PEI_t = \frac{\sum_{i=t-(2T-1)}^{t-T} PE_i}{T} - \frac{\sum_{i=t-(T-1)}^{t} PE_i}{T}$$

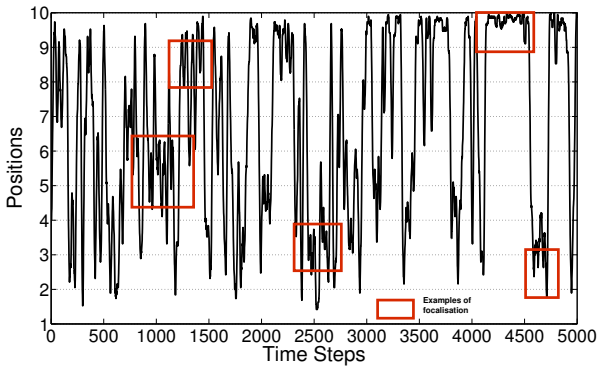

Fig. 1. Agent's position (y-axis) during the simulation (x-axis).

B. Results and analysis

Fig. 1 shows the behaviour of the agent during the simulation. The KB reward signal generated by the predictor improvement is able to drive the agent to explore the environment and to focus on different portions of it for relatively long times (up to 350 time steps).

However, the behaviour of the agent highlights two problems connected with the use of a pure KB mechanism in the perspective of competence acquisition. First, as expected, the system focuses on different areas over time but what it learns is not very useful in terms of competence. Indeed, the agent's actor learns *one-step stimulus-response associations* which represent a competence difficult to exploit. Indeed, a system working at the fine level might in theory learn all the possible combinations between all possible states and all possible primitive actions (10x2 in our case): these, which represent one-step policies, would however be no more useful for solving future extrinsically rewarded tasks than the original primitive actions.

Second, because of its architecture the system does not even cash the acquired one-step policies to form a repertoire of different skills. Indeed, at different stages of life the agents learns to associate different actions to the same state. This problem impairs one of the main functions related to the use of IM: learning and storing different skills in order to exploit them in the future to boost learning speed.

Some authors have shown how this type of mechanism, despite these problems, can be exploited for the acquisition of competences, for example in [7], [9], [17], [18], [19]. However, in many of these works the KB-IM mechanisms are used in support of, or are supported by, other learning processes and aiding mechanisms. For example, in [17], [18] the IM signal is used to improve the acquisition of behaviours mainly driven by extrinsic rewards. However, as argued in [8], [11], designing agents able to acquire competence without extrinsic rewards is a fundamental step towards having fully autonomous cumulative learning versatile systems.

In a different work [17], Schmidhuber implemented a system provided with a fixed reward value for the achievement of a desired state in addition to the reward signal generated by the predictor. However, this is not a good solution for at least two reasons: (a) if the reward is extrinsic, this is not in line with the idea of providing systems with IM to improve their autonomy and versatility; (b) as described in [21], providing a fixed reward signal would drive the system to get stuck on the rewarding activity, preventing it from learning different abilities and forming an rich repertoire of actions.

Oudeyer et al. [7], in one of the most influential works on IM applied to robot learning, focus on competence acquisition using a system based on the KB mechanism described in Sec. II-A. However, they used the predictor to predict few (three) high-level abstract important states (visual detection of an object; activation of a biting sensor; perception of an oscillating object). These high-level states represent few relevant states among a huge number of non-interesting states, and each of them can be achieved only with sequences of actions. In doing so, the authors, although not explicitly, deviate from the type of KB-IM mechanism they refer to and make an important step towards the CB-IM mechanisms we illustrate in Sec. III.

In summary, if one takes into account these extensions, KB-IM mechanisms can be important means to acquire competence. However, because of the two problems described above and for reasons and results indicated in Sec. III and Sec. IV, we think there is opportunity for important improvements.

III. Goals and hierarchies

In the two sub-sections below we describe two solutions for coping with the two problems of KB mechanisms highlighted in Sec. II-B, thus moving towards more suitable mechanisms for the acquisition of competence through IM.

A. Goals

If we are simply interested in widening the knowledge of a system (e.g. its ability to predict), we do not need a specific target in the learning process: everything that is not expected (predicted) is relevant for the system.

On the other hand, if we are looking at competence acquisition we have to carry out an important conceptual shift: while knowledge can be generic, competence is always *competence in doing something*. In particular, when we are trying to learn a new competence, what we are doing is improving our ability to reach a *specific target state* (or set of states) that we are considering important for some reasons: that *specific target state* is a *goal* (see Sec. V for a discussion on the possible origin of goals). A key implication is that if we look at the skills necessary for the achievement of a goal we need to shift from low-level motor primitives to temporally-extended activities. This is, for example, the key insight incorporated in the option theory [20] (see [6] for an example of the use of options with IM). The system can thus focus on acquiring multi-step policies directed to accomplish few critical goals, so avoiding the first of the two problems described in Sec. II-B related to focusing on learning every possible state-action combination.

If we translate this conceptual shift to the computational framework described in Sec. II, we have to reconsider the implementation of the IM signal. What the predictor has to anticipate is no longer any possible next state, but the *achievement of the goal*: given the actual state and the next performed action the predictor has to learn if the system will reach such a goal. In this way a reward signal (RS) is not generated at every time step, but *only when the goal is achieved* and on the basis of the PE of this achievement (cf. [6]):

$$RS = 1 - P$$

or in the case of the PEI:

$$RS = (1 - P_{t-1}) - (1 - P_t)$$

where P is the prediction, with a continuos value in [0, 1].

B. Hierarchical architectures

The architectural limits described in Sec. II-B, which prevent systems from cashing learned skills, do not allow the accumulation of different competences to be exploited in the future. To cope with this problem a good solution is to implement systems endowed with a *hierarchical architecture* able to retain previously learned abilities while discovering and acquiring new ones (see [22] for a review).

For example, the architecture in [8] is formed by different modules (*experts*), each focusing on learning a specific goal, and a *selector* that decides which expert will guide action and learn at each time step based on a CB-IM signal (see [14] for a thorough discussion of this, and [23] for a hierarchical architecture using similar IM mechanisms). In this way, the single expert can learn and retain the acquired competence while different competences are acquired by other experts. See [6] for another notable example of a model capable of acquiring multiple skills based on IM and a hierarchical architecture.

IV. Testing IM signals for competence acquisition assuming a hierarchical architecture

In this section we focus on mechanisms that can provide a proper signal for the selector of a hierarchical architecture similar to the one introduced in Sec. III-B. We first analyse signals determined by the *PE* and then those determined by the *PEI*. Although PE signals suffer from the unpredictability problem discussed in Sec. II, we believe that they can still play a significant role in driving competence acquisition. Indeed, not all environments present a relevant stochasticity and in some setups it is reasonable to assume that some goals are learnable by the system (e.g., self-determined goals, see Sec. V). Moreover, signals based on PE are more robust than those based on the PEI (as it is explained below in relation to Fig. 2 and Fig. 3).

To test these mechanisms we used the same simple scenario described in Sec. II-A, with some modifications. The agent

now has a goal, that is reaching the final position of the 1x10 grid. The simulation is run for a maximum of 100,000 trials, where each trial ends when the goal is reached or after a time-out of 20 time steps. At the beginning of every trial, the agent is positioned randomly on the 1x10 grid between positions 1 and 9; position 10 is the goal.

As we are interested in finding IM learning signals that can be used within hierarchical architectures, but at the same time we want to keep the simulations simple and focused on comparing different signals, we consider the actor-critic architecture of the previous system as one of the experts of an hypothetical hierarchical architecture. In this condition, what is needed to acquire a repertoire of skills is that the selector trains an expert to achieve high competence in accomplishing a goal (e.g., the goal considered here) and then, when no additional IM reward signal is generated, it selects a different expert to accomplish other goals. To capture this process using only one expert, we set a threshold (0.01) for the reward signal related to the goal, and when the signal goes below such threshold we stop the learning process. This mimics in an abstract fashion the fact that the selector finds more reward in selecting a different expert/goal and thus stops training the only expert forming our system. When the threshold is reached, we test the competence of the expert in reaching the goal by positioning the agent on every position of the grid (except the target position) for 30 times. In this way we test if the signal generated by the IM mechanism is able to guarantee the learning of a satisfying competence related to the goal.

The main problem related to the acquisition of competence based on KB reward signals is the decoupling between the ability of the predictor and the level of acquired competence (or: rate of prediction improvement and rate of competence improvement). This leads to the possibility that the predictor learns, and hence decreases the leaning signal, before the agent has acquired a full competence in achieving the goal. For this reason, we tested the IM mechanisms illustrated below with different learning rates for the predictors (the rates ranged in [0.002, 0.8]).

For all the mechanisms, the output of the predictor is a single linear unit fully connected with the input units. Connection weights are updated as described in Sec. II-A. The achievement of the final goal, or the failure to do so, are used as teaching input encoded with 1 and 0, respectively. The reward signal (RS) at time t is calculated as the average of PE calculated during a period T of 25 time steps:

$$PE_t = \frac{\sum_{i=t-(T-1)}^{i=t} PE_i}{T}$$

We tested three different mechanisms, which vary in the composition of the input to the predictor. Critically, the type of input is one of the differences that make the mechanism a KB-IM mechanism (measuring knowledge) or a CB-IM mechanism (measuring competence).

1) KB goal-oriented predictor: The first IM mechanism is similar to what we can find in [7], [9]: a predictor getting the input typical of the KB-IM mechanisms described in Sec. II-A (i.e., the actual state and planned action) but *used to predict the achievement of a goal rather than any state*.

2) CB one-state-ahead predictor: This predictor is a CB-IM mechanism. Differently from KB mechanisms, it does *not have actions as input*: the only input to the predictor is the actual position of the agent. In this way the prediction is directly connected to the competence of the system in reaching the final goal from the position that precedes the target one. Indeed, the prediction can improve only if the competence improves: in particular, only if the policy chooses an action in the state before the goal with an increasingly higher confidence. The effect of this is that the reward signal persists until there is competence to acquire. This kind of mechanism is analogous to what has been used, for example, in [6].

3) CB first-state predictor: This mechanisms establishes a stronger coupling between the learning of the predictor and the ability of the agent to reach the goal. The mechanism is inspired to what done in [24] where an IM signal is generated on the basis of the expansion of the region of the state space from which an options succeeds to achieve its goal. The input to the predictor is the first state encountered by the agent when randomly initialised at the beginning of each trial. This input determines a prediction that the goal will be achieved within the fixed time corresponding to the time out of the trial. This prediction is verified at the end of the trial (the success or failure are used as teaching input). In this way, the prediction is strictly connected to the capacity of the agent to achieve the goal from any possible initial state, and for this reason it can be considered a better measure of the competence of the system.

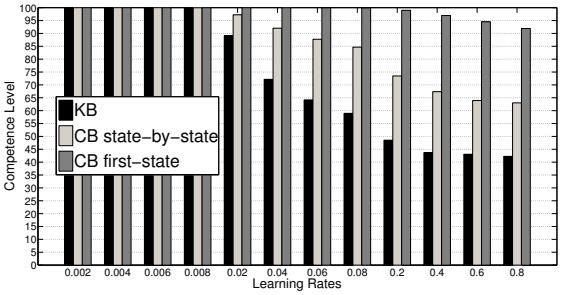

Fig. 2. Competence achieved by the agent in reaching the goal (y-axis) when driven by the PE signal generated by the three different IM mechanisms, corresponding to different learning rates of the predictor (x-axis).

A. Results and analysis

Fig. 2 shows the competence achieved by the system when the reward signal generated by the three different IM mechanisms goes under the threshold. The data are an average of 20 replications of each experiment.

With very low values of the learning rate (between 0.002 and 0.008) all the mechanisms are able to guarantee high competence in the achievement of the goal before terminating learning. However, when we raise the values of the learning

rate the three mechanisms determine different results in the level of competence acquired by the system. In particular, when driven by the KB mechanism, the achieved competence is very sensitive to the learning rate of the predictor: the higher the learning rate, the faster the predictor learns to anticipate the achievement of the goal, the sooner the learning process will stop. Even with low values of the learning rate (between 0.02 and 0.08) few presentations of that stimulus-response association are sufficient to significantly improve the knowledge of the predictor: in this way the reward signal becomes very low and the system terminates learning without having acquired a reliable competence. This result shows how KB mechanisms tend to produce signals which are inadequate for the acquisition of competence.

The signal generated by the CB one-step-ahead predictor is more robust to the variation of learning rates: indeed, now the system is able to reach a higher competence than with the KB mechanism. However, also in this case increasing the value of the learning rate determines a drop in competence acquisition, especially with high learning rates (between 0.2 and 0.8). The reason is that although the learning of the predictor is more closely coupled to the competence of the agent in achieving the target position, the signal generated by this mechanism is a good measure of the agent competence in achieving the goal *only from the states before the goal* itself. This is a limitation because, especially for complex sequences, it is possible for a system to have learnt to systematically reach the goal from positions close to it but still be unable to reach with sufficient ability those positions. For this reason, also the signal generated by this type of CB-IM mechanism is inadequate to properly drive competence acquisition, although it can be considered an improvement in relation to KB mechanism.

The signal determined by the CB first-state predictor mechanism is able to drive the system in achieving high level of competence independently of the learning rate values: even with high values it guarantees the acquisition of a competence level higher than 90%. The reason of these results is the fact that the improvement of the prediction ability of this mechanism is fully coupled with the ability of the agent to reach the goal *from any possible starting position*. In this way, the predictor is able to anticipate the achievement of the goal only when the whole sequence of actions that lead to the target position has been learned by the agent. The signal generated by this mechanism can be considered a good measure of the competence of the agent and for this reason this mechanism seems the most suitable for the acquisition of competence through IM.

B. Improvement in the prediction error signals

In order to have a more precise analysis of the different mechanisms, we tested the KB predictor and the CB first-state predictor (the CB one-step-ahead predictor can be considered an imperfect CB mechanism) also for the generation of a reward signal based on the PEI. Because of the differences of this signal with respect to the PE signal (noisier and weaker), we changed some of the conditions of the experiment. The threshold for the signal was set to 0.001 while the PE was calculated as an average over a period T of 75 trials. The PEI was calculated as described in Sec. II-A.

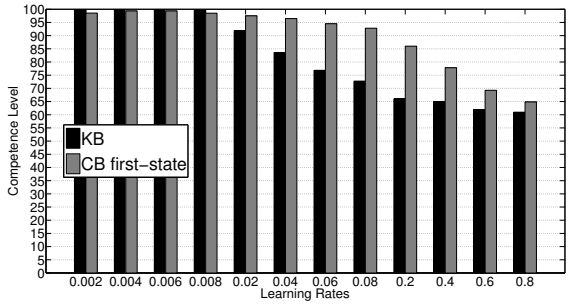

Fig. 3. Comparison between the competence achieved by the system in reaching the goal (Y axis) when driven by the improvement in prediction error signal generated by KB and CB mechanisms at the varying of the learning rate of the predictor (X axis)

Fig. 3 shows the average results of 20 replications of each experiment. Even using the PEI as the reward signal, when driven by the KB mechanism the competence achieved by the system suffers from the variations of the learning rates of the predictor, decreasing in a sensible way for values higher than 0.02. These results confirm the analysis of Sec. IV: KB-IM mechanisms produce inadequate signals for the acquisition of competence through IM.

When driven by the signal generated by the CB-IM mechanism, the system reaches a higher performance. Although PEI is more sensitive to the variations of the learning rate, the CB first-step predictor is able to allow the system to achieve competence levels higher than 90%, except for very high learning rate values. This confirms the fact that the learning of this type of mechanism is strictly connected to the improvement of the competence of the system in achieving desired goals. For this reason, it can be considered a good solution for the generation of an IM reward signal for competence acquisition.

V. Conclusions and future work

In this work we investigated which are the IM mechanisms that are best suited for guiding the acquisition of competence. In Sec. II we showed how systems driven by signals determined by every unpredicted state are not able to learn useful policies for the achievement of competence. In Sec. III-A and Sec. IV we highlighted and verified how shifting to a goal-directed perspective is a fundamental passage for the achievement of competence acquisition through IM.

Our analysis, supported by simple simulations, highlights how KB-IM mechanisms generate signals that are not fully suitable for competence acquisition. In contrast, CB-IM mechanisms, and in particular CB predictors whose learning development is fully coupled with the competence of the system, generate reward signals that guarantee the acquisition of skills.

Implementing hierarchical architectures, as described in Sec. III-B, is another key element for the achievement of competence acquisition through IM. A crucial point related to

this is the possibility for an agent to autonomously set its own goals. In this work we did not investigate this problem, but this is certainly a core topic for future research. An hypothesis that we suggest is the possibility of using KB-IM mechanisms, for example predictors of sensor activation (similar to those of [9], [21]), to signal unexpected changes in the environment and highlight interesting states that could become the desired goals of the system. However, detailed ways to implement this solution need to be further investigated.

Looking at different types of reward signals for guiding the acquisition of competence, we underlined how the error of a predictor is a robust signal that can be suitable for competence acquisition. However, this signal suffers from the well know problem of getting stuck in unpredictable situations (Sec. II). The prediction error improvement provides a solution to this problem but at the same time generates a small signal that is easily corrupted by noise and this can negatively influence the learning process.

A different solution to the problem of unpredictability is given by using the TD-error signal generated by the critic of an actor-critic architecture as an IM reward signal [8] (see also [14]). Note that this type of signal suffers from the same signal-noise problem of the PEI. To properly analyse these issues the implementation of a hierarchical architecture is necessary, which was not done here because of the aim of this work.

Future research will have to focus on the implementation of a goal-directed hierarchical architecture to better analyse a wider range of IM mechanisms and test the impact of different reward signals in the development of complex agents that deal with complex environments. Moreover, a hierarchical architecture would possibly allow one to fully attest how CB mechanisms overcame KB mechanisms in producing a proper reward signal for competence acquisition through IM.

ACKNOWLEDGMENT

The authors want to thank Prof. Andrew Barto for his advices and for the fruitful discussions on the topics of this paper. 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".

REFERENCES

[1] Ryan, Deci: Intrinsic and extrinsic motivations: Classic definitions and new directions. Contemporary Educational Psychology **25**(1) (2000) 54–67

[2] Harlow, H.F.: Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys. Journal of Comparative and Physiological Psychology **43** (1950) 289–294

[3] Kish, G.B.: Learning when the onset of illumination is used as reinforcing stimulus. Journal of Comparative and Physiological Psychology **48**(4) (1955) 261–264

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

[5] Schmidhuber, J.: Curious model-building control system. In: Proceedings of International Joint Conference on Neural Networks. Volume 2., IEEE, Singapore (1991) 1458–1463

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

[7] Oudeyer, P., Kaplan, F., Hafner, V.: Intrinsic motivation system for autonomous mental development. In: IEEE Transactions on Evolutionary Computation. Volume 11. (2007) 703–713

[8] Schembri, M., Mirolli, M., Baldassarre, G.: Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot. In Demiris, Y., Mareschal, D., Scassellati, B., Weng, J., eds.: Proceedings of the 6th International Conference on Development and Learning, Imperial College, London (2007) E1–6

[9] Santucci, V.G., Baldassarre, G., Mirolli, M.: Biological cumulative learning through intrinsic motivations: A simulated robotic study on the development of visually-guided reaching. In Johansson, B., Sahin, E., Balkenius, C., eds.: Proceedings of the Tenth International Conference on Epigenetic Robotics. Number 121–, Lund University Cognitive Studies, Lund (2010)

[10] Baldassarre, G.: What are intrinsic motivations? a biological perspective. In Cangelosi, A., Triesch, J., Fasel, I., Rohlfing, K., Nori, F., Oudeyer, P.Y., Schlesinger, M., Nagai, Y., eds.: Proceedings of the International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob-2011), IEEE, New York (2011) E1–8

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

[12] Oudeyer, P.Y., Kaplan, F.: What is intrinsic motivation? a typology of computational approaches. Front Neurorobot **1** (2007)

[13] Mirolli, M., Baldassarre, G.: Functions and mechanisms of intrinsic motivations: The knowledge vs. competence distinction. In Baldassarre, G., Mirolli, M., eds.: Intrinsically Motivated Learning in Natural and Artificial Systems. Springer-Verlag, Berlin (in press)

[14] Baldassarre, G., Mirolli, M.: Temporal-difference competence-based intrinsic motivation (td-cb-im): A mechanism that uses the td-error as an intrinsic reinforcement for deciding which skill to learn when. In Baldassarre, G., Mirolli, M., eds.: Intrinsically Motivated Learning in Natural and Artificial Systems. Springer-Verlag, Berlin ((in press))

[15] Schmidhuber, J.: A possibility for implementing curiosity and boredom in model-building neural controllers. In Meyer, J., Wilson, S., eds.: Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats, MIT Press/Bradford Books, Cambridge, Massachusetts/London, England (1991) 222–227

[16] Sutton, R., Barto, A.: Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA (1998)

[17] Schmidhuber, J., Zhao, J., Schraudolph, N.N.: reinforcement learning with self-modifying policies. In Thrun, S., Pratt, L., eds.: Learning to Learn. Kluwer (1997)

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

[19] Baranes, A., Oudeyer, P.Y.: R-iac: Robust intrinsically motivated exploration and active learning. IEEE Transactions on Autonomous Mental Development **1**(3) (2009) 155–169

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

[21] Santucci, V.G., Baldassarre, G., Mirolli, M.: Cumulative learning through intrinsic reinforcements. In Cagnoni, S., Mirolli, M., Villani, M., eds.: Evolution, Complexity and Artificial Life. Springer-Verlag, Berlin (in press)

[22] Barto, A.G., Mahadevan, S.: Recent advances in hierarchical reinforcement learning. Discrete Event Dynamic Systems **13** (2003) 341–379 10.1023/A:1025696116075.

[23] Stout, A., Barto, A.G.: Competence progress intrinsic motivation. In Kuipers, B., Shultz, T., Stoytchev, A., Yu, C., eds.: IEEE International Conference on Development and Learning (ICDL2010). IEEE, Piscataway, NJ (2010) Ann Arbor, MI, August 18-21, 2010.

[24] Hart, S., Grupen, R.: Intrinsically motivated affordance discovery and modeling. In Baldassarre, G., Mirolli, M., eds.: Intrinsically Motivated Learning in Natural and Artificial Systems. Springer-Verlag, Berlin (In press)