Evolution and Learning in an Intrinsically Motivated Reinforcement Learning Robot

Massimiliano Schembri, Marco Mirolli, Gianluca Baldassarre

Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche
Via San Martino della Battaglia 44, I-00185 Roma, Italy
{massimiliano.schembri, marco.mirolli,
gianluca.baldassarre}@istc.cnr.it

Abstract. Studying the role played by evolution and learning in adaptive behavior is a very important topic in artificial life research. This paper investigates the interplay between learning and evolution when agents have to solve several different tasks, as it is the case for real organisms but typically not for artificial agents. Recently, an important thread of research in machine learning and developmental robotics has begun to investigate how agents can solve different tasks by composing general skills acquired on the basis of internal motivations. This work presents a hierarchical, neural-network, actor-critic architecture designed for implementing this kind of intrinsically motivated reinforcement learning in real robots. We compare the results of several experiments in which the various components of the architecture are either trained during lifetime or evolved through a genetic algorithm. The most important results show that systems using both evolution and learning outperform systems using either one of the two, and that, among the former, systems evolving internal reinforcers for learning building-block skills have a higher evolvability than those directly evolving the related behaviors.

1 Introduction[1]

One important area of investigation of Artificial Life concerns the relationships existing between evolution and learning, the two key mechanisms that generate adaptive behavior in real organisms [1]. The synthetic approach of Artificial Life is an invaluable tool for investigating such a topic given the difficulties of collecting relevant empirical evidence related to it [2]. This approach already highlighted several important aspects of the relationship (for a review, see [3]), for example the fact that learning can guide evolutionary search [4] and that evolution can discover good starting conditions which can in turn facilitate learning processes during lifetime [5].

One of the most important distinctions between the two adaptive mechanisms is the time scale within which they operate [1]. In this respect, evolution has the advantage of producing various aspects of behavior 'readily available' at birth, but with the

[1] This research was supported by the EU Projects ICEA, contract no. FP6-IST-027819-IP, and MindRACES, contract no. FP6-511931-STREP.

cost that it can 'track' environmental changes only if they take place at a time scale longer than the individuals' life length. On the contrary, learning has the cost of causing inefficient behavior during the first phases of life but it allows tracking environmental changes within an individual's life span. Because of these time-scale differences, the models proposed so far, which typically used neural networks as agents' control systems, assigned to evolution the role of developing the 'general aspects' of learning systems, for example their overall architecture [6], the learning rules [7][8], the parameters regulating learning [9], and the initial connection weights [2][5], whereas they assigned to learning processes the role of updating connection weights during individuals' life. In this respect, a relevant novelty of this paper is that it proposes a reinforcement-learning system in which evolution develops some components of the system while learning uses these innate components to guide the training of the other components. The only work which carried out a study related to this issue is the pioneering work of Ackley and Littman [2]. In this work, the authors had a genetic algorithm evolving both an actor and an evaluator network, where the former was also trained during individual lifetime through a reinforcement learning algorithm on the basis of the evaluations of the latter. However, the main focus of that work was on the Baldwin effect. In contrast, the present work proposes a new hierarchical neural network architecture which learns to solve several different tasks by combining general skills acquired during an 'infancy' period. Hence, the most important novelty of the present work consists in studying the relationship between evolution and learning in the case in which *learning has a twofold nature*, and takes place on the basis of *both external and internal rewards*. This second point is directly related to a recent trend of research in the study of learning in artificial systems.

This new trend of research is inspired by the acknowledgement that when faced with new problems, organisms do not need to create solutions from scratch on the basis of low-level sensorimotor primitives but they can focus on composing and modifying previously developed general skills. Consequently, researchers in both machine learning [10][11] and developmental robotics [12][13][14] started to investigate systems with a twofold learning process. These systems acquire general skills on the basis of *internal motivations* (such as the drives to be exposed to novel/surprising/salient events), and then use these skills as building blocks to assemble more complex behaviors on the basis of *'external' rewards* (e.g. pleasure for eating and reproducing). This twofold process seems to play a fundamental role in the flexibility of behaviors exhibited by real organisms, especially the most sophisticated ones, like humans and primates in general [15][16]. Of course, understanding these processes is not only scientifically relevant but it is also one of the most important current goals of developmental robotics and machine learning, as it would allow building artificial intelligence systems having a flexibility and autonomy comparable to those of real organisms.

One of the most interesting machine learning proposals that encompass this insights, and that inspired the present work, is Intrinsically Motivated Reinforcement Learning (IMRL) [11][17]. The architecture used in IMRL is based on machine learning theory of 'options' [18]. Basically, options are sub-routines which can be invoked as any other primitive action, and include a set of initiation states where the option can be invoked, a termination condition, a policy mapping states to actions' probabili-

ties and, within the IMRL framework, an option model which maps initiation states to the probabilities of terminating the option in any other state. New options are created each time the system experiences a novel 'salient event'. A key point is that the system uses the prediction error of the option model as an *internal reward* to decide which option to invoke: the effect is that until the ability to produce the associated 'salient event' is not refined, an option continues to generate internal rewards and hence to be selected and trained (for a more detailed account of IMRL, see [17]).

The present paper (see also [19]) proposes a two-level hierarchical reinforcement-learning actor-critic architecture that represents a first attempt to solve two important drawbacks of the current implementation of the IMRL architecture: (a) the assumption of abstract representations of states and actions (e.g. grid-world environments and discrete actions), and (b) the fact that 'salient events' guiding options' formation and training must be hardwired by the programmer. The architecture tackles the problem (a) by using neural networks as components of the learning system which controls the behavior of a simulated robot, and tackles problem (b) by using a genetic algorithm to evolve neural 'reinforcers' that allow the system to autonomously associate a level of saliency to experienced states (see [14] for another solution to the same problems).

Using this hierarchical architecture, this research investigates the possible roles that evolution and learning can play when learning processes have the aforementioned twofold nature. In particular, it compares the performance (in terms of evolvability, learning speed, and maximal performance) of different versions of the system in which its two main components are either evolved or trained during life: the 'experts', which form the lower-level of the system's hierarchical architecture, and the 'selector', which forms its higher-level. The next section describes the proposed architecture, the task, the simulated robot, and the experimental conditions of the tests. Section 3 reports the results, while section 4 discusses the results and the limits of the present work and, on the basis of these, some possible directions for future research.

2 Simulated Robot, Task and Neural Network Architecture

The simulated robot is a mobile 'wheelchair' robot with a 30 cm diameter and a camera pointed towards a portion of the ground located just in front of the robot (24×8 cm). The robot perceives the ground using a grid of 6×2 sampling points associated with color-specific RGB receptors (so the system's input is a 'retina' formed by a 12×3 = 36 binary values). The robot's motor system is driven by setting the orientation variation within [-30, +30] degrees and the translation speed within [0, 2] cm. The environment is a square arena with a regularly textured floor (Fig. 1). Four different experimental conditions differing with respect to which parts of the system were evolved or trained were studied. We first illustrate the most complex experimental condition and then we explain the other conditions as variations to the former.

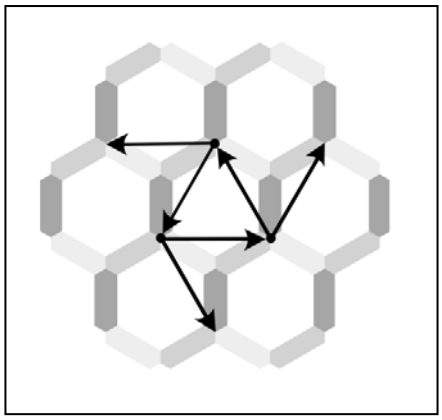

Fig. 1. The environment and the six 'adulthood' tasks. The sides of the hexagons are colored with blue (dark gray), red (gray) and green (light gray). Arrows represent the different tasks: each arrow's tail and head indicate, respectively, the starting and the target position of a task.

The robot's life is divided into two phases: 'childhood' and 'adulthood'. During childhood, the robot learns a set of basic sensorimotor skills based on intrinsic motivations. During adulthood, the robot learns to combine the acquired skills in order to accomplish six rewarded tasks (Fig. 1): in each task the robot has to reach a given target location starting from a particular position, and every time it reaches the target it receives one unit of reward and is set back to the starting position.

The controller of the robot (Fig. 2) is a hierarchical modular neural network formed by a 'selector' and three 'experts' (the quality of results did not change in tests with a higher number of experts). The selector and each experts are neural network implementations of the actor-critic reinforcement-learning model [21], which is known to have a high biological plausibility [20][22][23][24]. Each *expert* is formed by three components: (a) a *reinforcer*: a perceptron mapping the retinal input to a [-1, 1] sigmoid unit encoding the internal reward for that expert (reinforcers are evolved, see below); (b) an *actor*: a perceptron mapping the retinal input to two sigmoid units; the activation of these units sets the centre of a Gaussian function which is used to generate noisy commands issued to the motor system: the first unit sets the orientation variation of the robot, the second unit sets its translation (initial standard deviation = 0.3; noise is linearly reduced to zero during childhood); (c) *critic*: this is based on an *evaluator*, a perceptron that maps retinal input to one linear output unit encoding the expert's evaluations of states; these evaluations, together with the reward produced by the expert's reinforcer, are used to compute the *surprise* of the expert's critic in a standard way [21]. The *selector* is formed by two components: (a) *selector's actor:* a perceptron that maps the retinal input to three sigmoid output units; at each time step, the activations of these units, each corresponding to an expert, are used as pseudo-probabilities to select the expert that takes control of the motor system and (during childhood) learns; (b) *selector's critic*: analogous to the experts' critics, it uses as its reward signal either external rewards or the surprise of the expert which currently has the control (see below).

During childhood, at each time step the selector selects the expert that has the control. The selected expert: (a) selects and execute an action; (b) trains its evaluator as in standard function-approximation actor-critic models [21], but on the basis of the internal rewards delivered by its own reinforcer (discount factor = 0.99); (c) trains the actor as in [24]: if surprise is positive, the output units' activations are 'moved' (with a delta rule) towards the (Gaussian noisy) values corresponding to the executed action, whereas if surprise is negative the output units' activations are moved in the opposite direction (learning rate of evaluator and actor = 0.01). On the other hand, in order to train its own actor and evaluator, the selector uses the *surprise of the selected expert as its (internal) reward signal*. As the surprise of an actor-critic system is a good indicator of its learning progress, during this phase the selector learns to give the control to the expert which is learning at the maximum rate. Note that as surprise needs two succeeding evaluations to be computed, learning occurs only when the same expert is selected for at least two contiguous time steps.

During adulthood experts are not trained, whereas the selector is trained as in childhood, but this time not on the basis of expert's surprises, but rather on the basis of the task-related *extrinsic rewards*. During adulthood the selector's weights are reset before tackling each task in order to avoid interference between different tasks.

The genetic algorithm uses a population of 50 individuals, encoding connection weights as real variables (with initial random values in [-1.0, +1.0]), evolved for 100 generations. The duration of childhood is 150,000 time steps, while the duration of adulthood is 600,000. The fitness is computed as the number of times that the robot reaches the target divided by the theoretical maximum achievable if the robot followed the straight lines indicated in Fig. 1 at maximum speed. At the end of each generation the best 10 individuals are selected and generate 5 offspring each. Each weight of the offspring is mutated with a probability of 10% by adding to it a random value uniformly drawn in [-1.0, +1.0].

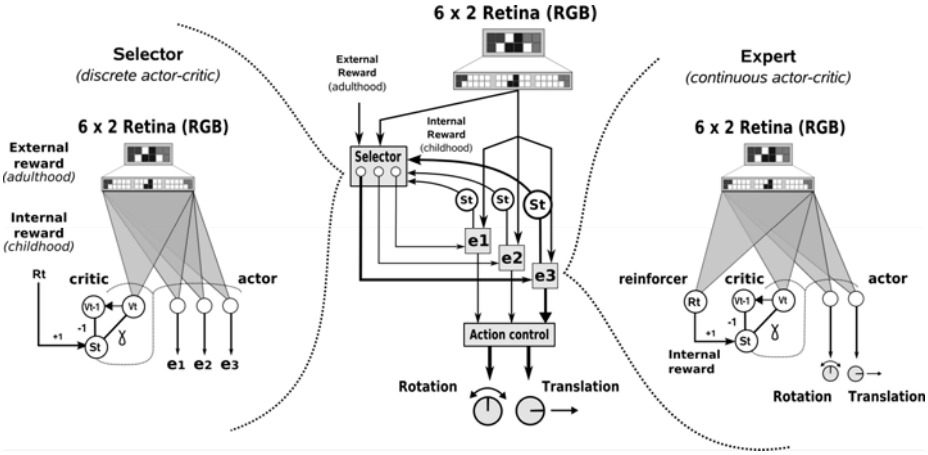

Fig. 2. Center: the whole architecture. Left: the selector's architecture. Right: one expert's architecture (see text for details)

Four different experiments were run with the following conditions:

1. Learning Experts, Learning Selector (LE-LS). This is the condition just described, in which individuals' genome encodes only the connection weights of the three experts' reinforcers.
2. Evolved Experts, Learning Selector (EE-LS). In this condition experts' actors are encoded in the genome and evolved (hence there is no childhood), while the selector is trained during adulthood as described above.
3. Evolved Experts, Evolved Selector (EE-ES). In this condition the actors of both the experts and the selector are evolved, and no learning takes place.
4. Single Learning Expert (SLE). In this condition no evolution takes place, and a simple expert is used to directly tackle each of the six adult tasks on the basis of only extrinsic rewards (weights were reset at the beginning of each task to avoid interferences between different tasks).

3 Results

Direct observation of the behavior of the evolved individuals indicates that organisms endowed with the hierarchical architecture we have presented (that is those of all but the SLE condition) tend to solve their tasks in the following way. Experts tend to specialize for following one color each, while the selector tend to compose experts' basic skills so to navigate on the colored lines and then choose the most appropriate direction at each junction (for a more detailed analysis, see [19]). This is particularly true for organisms of conditions LE-LS and EE-LS, that is the conditions in which the selector can learn during life how to make the best possible use of the experts' skills.

In order to compare the results of the four conditions, we present three kinds of statistics, which are meant to assess different properties of the various systems: (a) fitness of the best individuals along generations reveals systems' evolvability; (b) performance throughout a long learning period reveals systems' learning speed; (c) performance after a long period of learning reveals systems' steady-state ability.

Fig. 3a reports the fitness of the best individuals along 100 generations for the three conditions involving evolution: LE-LS, EE-LS and EE-ES. The most striking result is that the condition LE-LS is clearly far more evolvable than the other two conditions: it requires about an order of magnitude less generations than the other two to reach a steady state performance (about 10 vs. about 100). Moreover, the LE-LS condition has a higher reliability in different evolutionary runs (note the much smaller standard deviation in the graph). On the other side, EE-LS achieves a higher final fitness with respect to LE-LS. This happens because in the EE-LS condition, evolution is able to find highly accurate and reliable experts (data not reported), whereas the learning of the experts during childhood is always noisy, and results in the acquisition of sub-optimal basic skills. However, this limit might be reduced or even overcome by prolonging the rather short childhood phase used here and/or by optimizing the experts' learning parameters like learning rate and discount factor. Another remarkable result is that the EE-ES condition produces individuals with a quite high fitness, at the same level of the LE-LS condition (consider that in the EE-ES condi-

tion the selector is evolved, and hence robots in this condition must find a single solution for all the six different tasks). This is due to the well-known remarkable ability of evolutionary searches to find very 'smart' solution to difficult problems [25]. In particular, evolved organisms of the EE-ES condition typically produce a stereotyped behavior such that the robot follows a circular path at maximum speed which includes most of the target positions: in this way, some tasks are accomplished very efficiently, other with a reasonable efficiency, while other targets are never reached at all. This fact, together with the fact that the behavior of these robots is completely inherited, and hence fully developed from birth, explains the quite good performance reached by this condition.

Fig. 3b shows the learning curves of the three conditions involving learning: LE-LS, EE-LS and SLE over 1,000,000 cycles (for each task). The most important result is that the compositional strategies (LE-LS and EE-LS) clearly outperform the 'monolithic' strategy (SLE) in terms of learning speed. On the other hand, EE-LS and SLE outperform LE-LS in terms of final performance. In the same vein as the result on fitness discussed above, this is explained by the fact that EE-LS can evolve highly reliable experts, SLE can train its only expert during a very long period of test (1,000,000 cycles), whereas LE-LS can only sub-optimally train its three experts during the relatively short childhood phase (150,000 cycles). Finally, the higher fitness of EE-LS with respect to SLE is due to the fact that the former can solve its tasks by efficiently combining useful low-level skills rather than by relying on one single actor.

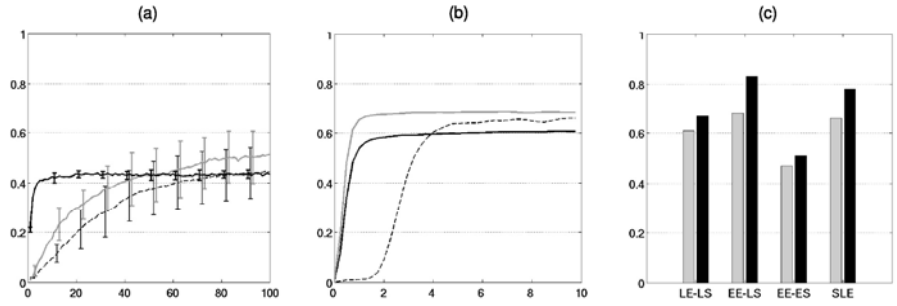

Fig. 3. (a) Evolution of the fitness of the best individuals (averaged over 10 runs) along 100 generations, for the three conditions involving evolution: LE-LS (bold line), EE-LS (gray line), and EE-ES (dashed line). The graph also reports standard deviations. (b) Average performance during learning tests lasting 1,000,000 cycles for the three conditions involving learning: LE-LS (bold line), EE-LS (gray line), and SLE (dashed line). Curves refer to the average performance (normalized number of received rewards) of the 10 best individuals of each of 10 runs on 10 tests for each of the 6 tasks (i.e. average of 10×10×6 tests). (c) Steady-state performance level of all the four conditions measured as average over the last 100,000 cycles of the data reported in graph 'b' (dark gray bars: average over 10 runs; light gray bars: best run). For the EE-ES condition the test of graph 'b' was run with no learning process taking place.

Fig. 3c shows the steady state level of performance achieved in all the four conditions at the end of learning: these tests allow to compare final performance independently from the time spent to acquire behavior. The results show that EE-ES has the lowest performance as it pays the costs of its rigid behavior. LE-LS has a performance lower than EE-LS and SLE because of the mentioned difficulty to optimize the experts in the short childhood phase. Finally, EE-LS slightly outperforms SLE because of the higher mentioned efficiency of the compositional strategy that can rely upon specialized experts.

4 Discussion and Future Work

This paper investigated the role played by evolution and learning in adaptive behavior when learning processes during life take place in two stages, one where the systems acquire flexible sensorimotor skills on the basis of intrinsic motivations (as a general drive to explore) and a second one where those skills are assembled to accomplish tasks that directly increase fitness (e.g., allow eating) on the basis of extrinsic rewards (e.g. pleasure from food). To this purpose, we used a reinforcement-learning hierarchical neural-network architecture as the control system of a simulated robot and we evaluated the effects of applying either evolution or learning to the various components of the system.

The results highlighted various interesting phenomena related to the relative strengths and limits of evolution and learning, and to their complementary roles in producing adaptive behavior. First of all, the they clearly confirmed previous seminal works (see [1][3]) indicating that evolution alone has the limit of producing rigid behaviors whereas learning alone has the limit of exposing organisms to long periods of non-adaptive behavior. On the contrary, systems that build up adaptive behavior on the basis of both evolution and learning tend to have both the flexibility and fast adaptation advantages provided by the two adaptive processes. With respect to the behavioral flexibility provided by learning, one should also consider that in the learning tests done in this paper robots were tested with the same tasks used during evolution. The advantages provided by learning would surely be much stronger if the systems were tested with tasks which have never been encountered during evolution: this might be a subject of investigation in future work.

A novel interesting finding of this work is that within 'mixed' systems, which rely on both evolution and learning, developing *innate low-level behaviors* in the course of evolution might allow achieving a higher performance. This is in line with the presence of a few but important innate behaviors even in the most complex species such as primates. These are typically behaviors which are very directly related to fitness (like the behaviors implemented by the experts of our system) and for which a ready availability at birth is very important (examples of these are the motor reflexes or basic behaviors related to feeding such as salivation and babies' suction reflex).

On the other hand, our simulations clearly demonstrate that *evolving general criteria (reinforcers) for guiding learning of building-block behaviors* is much easier than directly evolving behaviors themselves. Furthermore, the entity of this effect in the experiments presented here is so big that it suggests that such result might be caused

not only by a difference in search spaces for the two conditions (in our experiments reinforcers have half the weights of the actors) but also by the fact that, generally speaking, evolving 'goals' might be much easier than evolving the behaviors that satisfy them (a similar suggestion has also been made by [2]). Future research should investigate more in detail why this is the case.

Furthermore, and most important, our experiments clearly show that the costs of learning, namely the need to acquire behavior from scratch at every generation, can be significantly diminished if agents have a hierarchical control system architecture like the one presented here. In this case, organisms which have to tackle several different tasks during their life can accomplish this by combining general low-level abilities which might be either genetically inherited or acquired during a childhood phase. Indeed, the system that learned each behavior from scratch took nearly four times to reach a performance comparable to that of systems exploiting compositional strategies. This result strongly supports the motivations behind the Intrinsically Motivated Reinforcement Learning framework.

Although interesting, these results are preliminary in many respects, and their limits suggest important problems for future research. First, several interesting conditions have not been explored yet, for example the conditions in which: (a) the genetic algorithm evolves neither the actors of the experts (as in the EE-LS), nor their reinforcers (as in the LE-LS), but rather their evaluators (cf. [2]); (b) the whole hierarchical architecture is trained only on the basis of external task-related rewards; (c) learning and discount parameters are evolved; (d) not only expert's reinforcers, but also their number is evolved. Second, the present architecture might be improved under various respects: for example the selector, which is supposed to operate at a more abstract level with respect to experts, should not operate at the same time-scale and with the same input as them. Notwithstanding these limits, we think that the work presented here is a first important step in the investigation of the relationships existing between evolution and *compositional learning* processes.

References

1. Nolfi, S.: Learning and Evolution in Neural Networks. In: Arbib, M. (ed.): The Handbook of Brain Theory and Neural Networks. The MIT Press, Cambridge, MA (2003) 415-418
2. Ackley, D., Littman, M.: Interactions Between Learning and Evolution. In: Langton, C.G., Taylor C., Farmer J. D., Rasmussen S.: Artificial Life II. Addison-Wesley, New York (1991) 487-509
3. Nolfi, S., Floreano, D.: Learning and Evolution. Autonomous Robots 1 (1999) 89-113
4. Hinton, G., Nowlan, S.: How learning guides evolution, Complex Systems 1 (1987) 495-502
5. Belew, R., McInerney, J., Schraudolph, N.: Evolving networks: Using the genetic algorithm with connectionist learning. In Langton, C.G.: Proceedings of the Second Conference on Arificial Life. Addison-Wesley, Reading, MA (1992)
6. Di Ferdinando, A., Calabretta, R., Parisi, D.: Evolving Modular Architectures for Neural Networks. In French, R., Sougné, J.: Connectionist Models of Learning, Development and Evolution. Springer Verlag, London (2001) 253-262

7. Urzelai, J., Floreano, D.: Evolution of Adaptive Synapses: Robots with Fast Adaptive Behavior in New Environments. Evolutionary Computation, 9(4) (2001) 495-524
8. Niv, Y., Joel, D., Meilijson, I., Ruppin, E.: Evolution of Reinforcement Learning in Foraging Bees: A Simple Explanation for Risk Averse Behavior. Neurocomputing 44(1) (2002) 951-956
9. Eriksson, A., Capi, G., Doya, K.: Evolution of Meta-parameters in Reinforcement Learning Algorithms. In: Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems
10. Schmidhuber J.: A Possibility for Implementing Curiosity and Boredom in Model-Building Neural Controllers". In Meyer, J-A., Wilson, S.W.: From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior, The MIT Press, Cambridge MA (1991) 222-227
11. Barto, G., Singh, S., Chentanez, N.: Intrinsically Motivated Learning of Hierarchical Collections of Skills. In: Proceedings of the Third International Conference on Development and Learning (2004).
12. Huang, X., Weng, J.: Novelty and Reinforcement Learning in the Value System of Developmental Robots. In: Prince, C. G., Demiris, Y., Marom, Y., Kozima, H. , Balkenius, C.: Proceedings Second International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems. Lund University Cognitive Studies, Lund (2002) 47-55
13. Marshall, J., Blank, D., Meeden, L.: An Emergent Framework for Self-Motivation in Developmental Robotics. In: Proceedings of the Third International Conference on Development and Learning (ICDL 2004) (2004) 104-111
14. Oudeyer, P., Kaplan, F., Hafner, V.V.: Intrinsic Motivation Systems for Autonomous Mental Development. IEEE Transactions on Evolutionary Computation 11(1) (2007)
15. White, R.W.: Motivation Reconsidered: The Concept of Competence. Psychological Review 66 (5) (1959) 297-333
16. Berlyne, D.E.: Conflict, Arousal and Curiosity. McGraw-Hill, New York (1960)
17. Stout, G.D., Konidaris, Barto, A.G.: Intrinsically Motivated Reinforcement Learning: A Promising Framework For Developmental Robot Learning. In. Proceedings of the AAAI Spring Symposium on Developmental Robotics (2005)
18. 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
19. Schembri, M., Mirolli, M., Baldassarre, G.: Evolving Internal Reinforcers for an Intrinsically Motivated Reinforcement-Learning Robot. 6th IEEE International Conference on Development and Learning (ICDL2007) (submitted)
20. Baldassarre, G.: A Modular Neural-Network Model of the Basal Ganglia's Role in Learning and Selecting Motor Behaviours. Journal of Cognitive Systems Research 3 (2002) 5-13
21. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. The MIT Press, Cambridge MA (1998)
22. Houk, J.C., Davis, J.L., Beiser, D.G.: Models of the Basal Ganglia. The MIT Press, Cambridge MA (1995)
23. Schultz, W.: Getting Formal with Dopamine and Reward. Neuron 36 (2002) 241-263
24. Mannella, F., Baldassarre, G.: A Neural-Network Reinforcement-Learning Model of Domestic Chicks that Learn to Localise the Centre of Closed Arenas. Philosophical Transactions of the Royal Society B – Biological Sciences 362(1479) (2007) 383-401
25. Nolfi, S.: Evolutionary Robotics: Exploiting the Full Power of Self-Organization. Connection Science 10(3-4) (1998) 167-183