Valency for Adaptive Homeostatic Agents: Relating Evolution and Learning.

Theodoros Damoulas, Ignasi Cos-Aguilera, Gillian M. Hayes, and Tim Taylor

IPAB, School of Informatics, The University of Edinburgh,
Mayfield Road, JCMB-KB, EH9 3JZ Edinburgh, Scotland, UK
{T.Damoulas, I.Cos-Aguilera, G.Hayes, T.Taylor}@ed.ac.uk
http://www.ipab.inf.ed.ac.uk

Abstract. This paper introduces a novel study on the *sense of valency* as a vital process for achieving adaptation in agents through evolution and developmental learning. Unlike previous studies, we hypothesise that behaviour-related information must be underspecified in the genes and that additional mechanisms such as valency modulate final behavioural responses. These processes endow the agent with the ability to adapt to dynamic environments. We have tested this hypothesis with an *ad hoc* designed model, also introduced in this paper. Experiments have been performed in static and dynamic environments to illustrate these effects. The results demonstrate the necessity of valency and of both learning and evolution as complementary processes for adaptation to the environment.

1 Introduction

The relationship between an agent's ability to monitor its internal physiology and its capacity to display adaptation to its environment has received little attention from the adaptive behaviour community. An aspect of this relationship is the sense of *valency*, a mechanism evolved to foster the execution of beneficial actions and to discourage those whose effect is harmful. Agents endowed with this sense exhibit some advantage over others which cannot anticipate the effect of some actions in their decision making. This facilitates the life of an agent in a competitive environment. Formally, we have defined the sense of valency as the *notion of goodness or badness attached by an individual to the feedback from the environment resulting from the execution of a behaviour*. We therefore view valency as a process occurring in a framework of interaction relating perception, internal bodily dynamics and behaviour arbitration. We have implemented these elements in a simulated animat, consisting of an artificial internal physiology [6, 5, 15], a behaviour repertoire, a selection module and a valency module. The goal of this agent is to survive, ergo to maintain its physiological parameters within their viability zone [2].

Previous work [1, 4] hypothesised genes to encode the valency of stimuli and the behavioural responses to stimuli (represented as an evaluation network or as a motivation system, respectively). Both studies use the valency as a feedback loop that assesses and corrects their behavioural patterns. These studies focused

on the interaction between learning and evolution via the Baldwin effect [3, 9], where certain action-related genes dominate and *shield* other genes encoding the valency of related actions. They argue that random mutations allow developmental knowledge to be transferred to the genome, which may deteriorate the valency-related ones. As stated by [1]: *The well-adapted action network apparently shielded the maladapted learning network from the fitness function. With an inborn skill at evading carnivores, the ability to learn the skill is unnecessary.* However, we argue that this may be necessary in a variety of cases, e.g. if the environment constantly changes, it does not seem reasonable to encode volatile information in the genes (this may lead to the extinction of the species). Instead, it seems wiser to genetically encode action-related information in an underspecified manner to be completed via interaction with the environment (via reward driven learning). If as a result of the combination of both processes this information is transferred to the next generation, this would endow the next generation with the necessary knowledge to survive while maintaining the flexibility for a range of variation within its environment.

A model to test this hypothesis is introduced next with three different versions. Section 2 introduces the agent's internal model. Section 3 presents the three approaches examined, their corresponding elements and the results for static and dynamic environments. Finally, Section 4 discusses the results obtained.

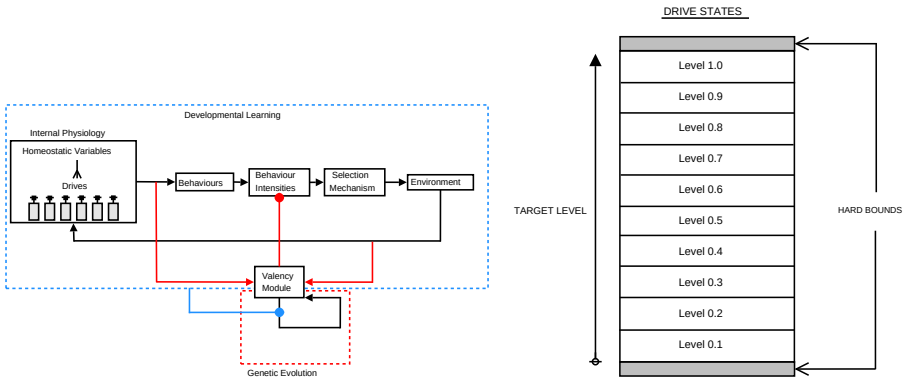
2 Model Architecture

2.1 Internal Agent Structure

The agent's *internal physiology* is a simplified version of the model proposed by Cañamero [5]. In this model the agent's internal resources are represented as *homeostatic variables*. These are characterised by a range of operation and by an optimal value or set point. They exhibit their status of deficit or excess via a set of related *drives* [10], which together with the external stimuli configure the agent's motivational state [19]. For our case we are only interested in the agent's internal interpretation of the effect. Therefore, it is possible to simplify the environment to its minimal expression: the feedback or effect per interaction. This allows us to combine homeostatic variables and drives in a single parameter: *homeostatic drives*. These drives decay according to

$$Level_{(t)} = Level_{(t-1)} \times 0.9 + \sum_{j} < effect\ of\ \ action >_{t_j} \tag{1}$$

where *level* is the value of a drive and *effect of action* the value of the effect of executing a certain behaviour (an incremental transition of +0.1, 0.0 or -0.1 on the drives). To simplify, the drives are generic (e.g. energy related) since we are mostly concerned with them in terms of their functionality (e.g. decay, discretised states, etc.). Figure 1(b) shows a schematic view of a single drive with its discretised states and the hard bounds.



(a) General Schema (b) Generic Homeostatic Variable

Fig. 1. Left: General Architecture. Right: A typical drive with 10 discretised states from 0.1 to 1.0. The drive has hard bounds below 0.1 and above 1.0 (ad-hoc) to ensure that agents operate within limits.

The *selection mechanism* consists of choosing the action exhibiting the largest valency. As we shall see, the association action-valency is learned during the lifetime of the agent.

2.2 Lifetime Learning

Valency is interpreted by the agent as the relative value of executing a behaviour. This association is learned during lifetime via the *valency module* (cf. center-bottom in Fig. 1(a)) and directly affects the *behaviour intensities* according to the effect that executing a behaviour has on the *internal physiology*.

The learning of the agent is modeled as a 'full' reinforcement learning problem [17]. Every action and state transition of the agent's physiological space is evaluated according to the reward function that is provided by genetic evolution. The learning has been modeled as a Temporal Difference (TD) algorithm [16], since this learns by experience and without *bootstrapping*, i.e., lacking a model for the environment. This should be of advantage in dynamic environments.

The Q-learning algorithm was used with the Q-Values representing the valency of actions and the update rule (2) indirectly associating the effect of an action to a specific valency through the individual reward function.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \, [r_{t+1} + \gamma \, \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]. \qquad (2)$$

2.3 Genetic Evolution

The *valency module* is part of both processes, developmental and genetic. It acts as a link between the two, using genetic information and lifetime experience in order to create an individual *sense of valency*. According to our implementation

the core element of valency is the reward function which is the genetically encoded information. This is *independent* of the behaviours and could be encoded into the genome in a biologically plausible manner.

The *reward function* is evolved by a standard GA [12]; it is either directly encoded in the animat's chromosome or indirectly encoded as the weights of a neural network. In other words, each animat is genetically "defined" to assign a *utility* to each change in its physiological state due to the execution of a behaviour.

The role of genetic evolution and developmental learning in the mechanism of valency, the evolutionary nature (direct or indirect encoding) of the reward function and their effect on adaptation to dynamic environments are, respectively, the issues we have addressed with three different models, introduced in the next section.

3 Experiments and Results

In order to examine the effect of valency in the developmental and genetic processes, this approach has been implemented with direct and indirect encoding of the reward function (Models 1a and 1b), and compared to a model that uses genetic evolution only (Model 2). Models 1a and 1b are used to demonstrate that the instabilities of Darwinian models in dynamic environments [11, 14] are due to having *action selection* (as opposed to just the reward function) encoded in the genome. Model 2 is used to examine the necessity of developmental learning in stable and dynamic environments.

Models 1a and 1b test different evolutionary encodings of the reward function. In Model 1a the reward function is directly encoding on the chromosome, whereas in Model 1b the chromosome encodes synaptic weights of a neural network that estimates the reward function. This second encoding method has been extensively used and tested in previous work [4, 8, 13, 14, 18].

Finally, we examine the above approaches in both stable and dynamic environments in order to observe their effect on the adaptability of our animats.

3.1 Experimental Setup

The *environment* has been modeled in terms of deterministic reward. Every time the agent performs an action, the environment feedbacks a physiological effect, which is unique for each behaviour. A *static environment* is characterised by relating to each behaviour a unique effect, which is maintained throughout generations (e.g. action 1 has always a -0.1 effect). In contrast, the effect related to each behaviour is inverted every generations for *dynamic environments* (action 1 changes effect from -0.1 to +0.1).

The *Q-Values* represent the value of selecting a specific action in a given state. Q-Values model the valency of actions and qualify an execution as successful or not. Since for every drive we have 10 discrete states and in every state the same action set is available, the Q-Value table for describing every state-action will be

a matrix of dimensions $10 \times (\#Actions\ per\ state \times \#Drives)$. The initialization of the Q-Values has always been performed by setting them to 1 and the update rule 2 converged those values to the individual *valency* of each agent based on its reward function.

The *learning episode* of selecting actions for a number of steps is repeated for at least 10,000 cycles where convergence of the best *possible* behaviour according to the individual reward function is ensured. A *competition* procedure was used to assign the fitness value at each agent at the end of the learning cycle (if applicable). The agent was initialized on a minimum drive level, it was allowed to use its action-selection mechanism (converged Q-values) for a specific number of steps and it was scored according to its overall performance. The target was to reach satiation on its drive(s) and to remain at that level (1.0).

The *metrics* used in our study are the average and maximum fitness progressions through generations of animats. The maximum fitness score each time (10 for single drive and 20 for double drive cases) indicates a perfect performance over the competition cycle and a successfully evolved/developed sense of valency.

3.2 Learning & Evolution with Indirect Encoding of Action Selection

As has been shown previously [1, 11, 14], direct encoding of action selection leads to animats that are behaviouraly predisposed. Consequently, their fitness progressions in dynamic environments suffer from relative instabilities. To overcome these limitations, our **Model 1a (RL & GA)** was investigated (Fig. 2), where the genome is not directly encoding action-selection. Instead it carries information (reward for state transitions) used to build the *behaviour-selection* mechanism via developmental learning.

An alternative version of the above implementation, **Model 1b (RL & GA with NN**), which uses a neural network for the provision of the reward function (Fig. 3), was also examined. The genome is still *indirectly* encoding action selection.

3.3 Strictly Evolutionary Approach

The final model (Model 2) was used to test the necessity of lifetime learning as a complementary process to genetic evolution. **Model 2 (Q-Evolution)** is strictly evolutionary (Fig. 4).

3.4 Learning & Evolution, or Evolution Alone?

Static Environment The *base case* considered first is that of a static environment with an animat endowed with a single drive. As seen in Fig. 5, every approach manages to produce ideal solutions, i.e., animats with a sense of valency are able to guide selection toward the satiation of the drive. The results confirm our hypothesis that a combined framework of learning and evolution through

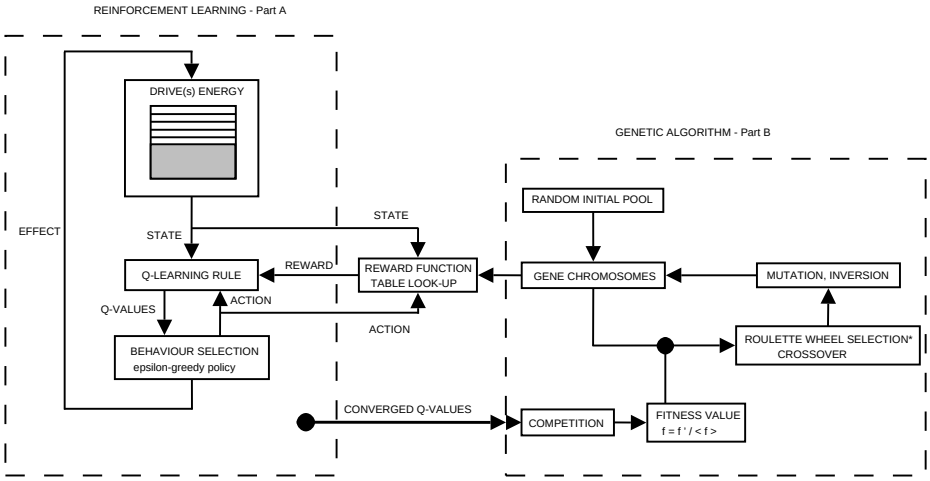


Fig. 2. **Model 1a (RL & GA) of combined Learning and Evolution. The chromosome encodes the reward function of each agent (magnitudes in the range 0-10)and the Q-learning update rule is used to create the action selection mechanism through lifetime. Step-size parameter α=0.1 and ϵ=0.1**

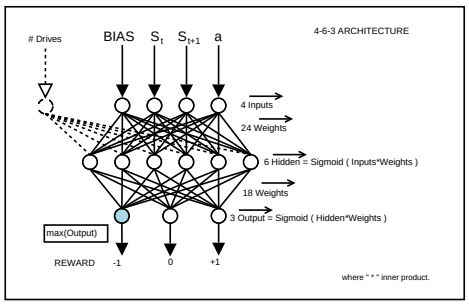


Fig. 3. **The Neural Network architecture used in Model 1b (RL & GA with NN). The input is the states, the bias, and the action whereas the output is the magnitude of reward. In this case a simple set of reward was used with +1, 0 or -1 possible values. The bias is set to 1 and the network operates with a standard sigmoid function. In the case of more then one drives an additional node inputs that information.**

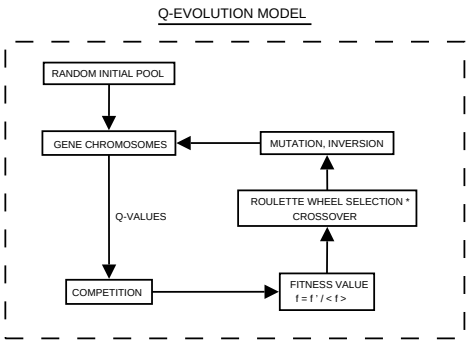
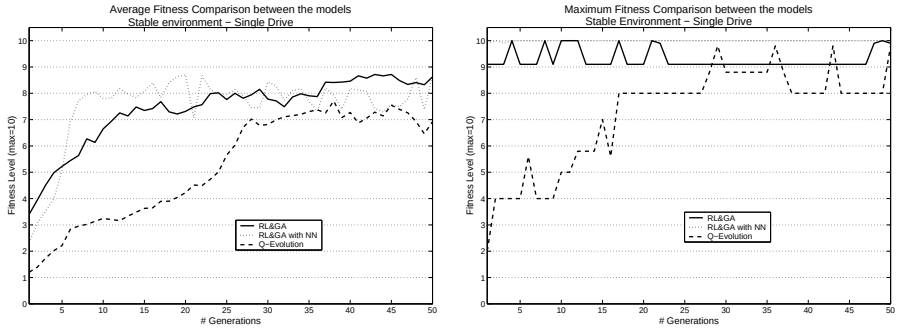


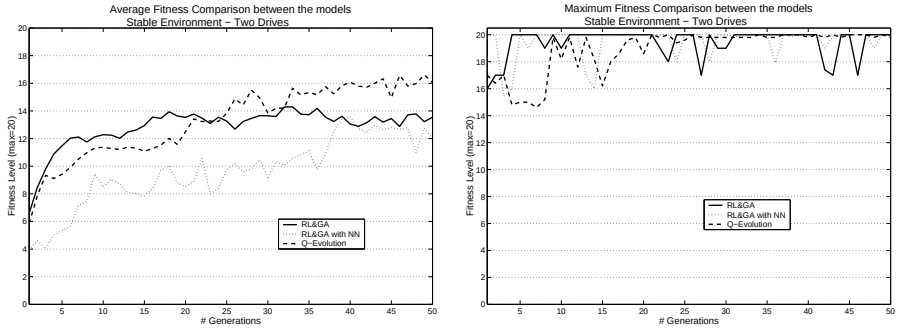
Fig. 4. **Model 2 (Q-Evolution), implementing only standard Evolutionary techniques without a Learning cycle. The valencies of actions (Q-Values) are directly evolved instead of the reward function and hence the genome of agents encodes information that is directly connected to the action selection mechanism. The model is operating without a valency module.**



(a) Average (b) Maximum

Fig. 5. **Average and Maximum Fitness results for the three models on a stable environment where the animats have a single drive. The fitness function requires optimal behaviour selection in order to achieve maximum status. Notice how the models utilizing developmental learning achieve higher fitness levels in a few number of generations.**

the valency module performs better than those lacking it. However, the strictly evolutionary approach (Q-Evolution model) still manages to achieve, in certain occasions, maximum fitness and to increase the average of the population. The approach that directly evolves the reward function (RL & GA) achieves a higher average fitness but is less stable in the maximum fitness development compared with the alternative evolution of synaptic weights (RL & GA with NN).



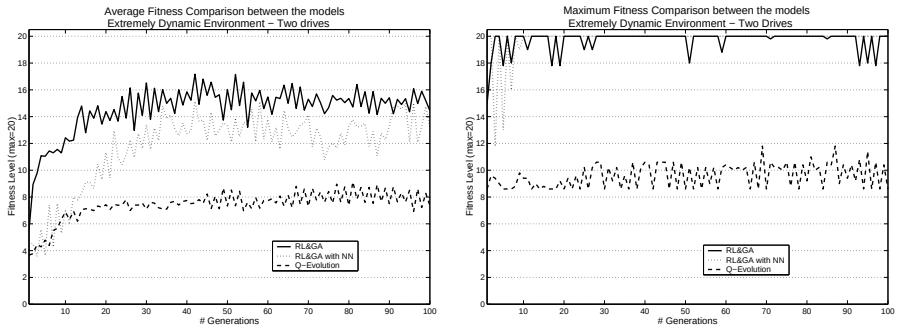
(a) Average (b) Maximum

Fig. 6. **Average and Maximum Fitness results for the three models on a stable environment where the animats are utilizing two drives. The results are for a "loose" fitness function that allows suboptimal behaviour selection.**

The *double drive case* in a stable environment increases the difficulty of the task and explores the capabilities of all the approaches. The results in Fig. 6 compare the models on a "loose" fitness function (excess competition steps) that allows for suboptimal behaviour selection (the animat can achieve maximum fitness without selecting always the best possible action). For a fitness function

requiring optimal behaviour selection (that is, always to choose the best behaviour), the strictly evolutionary approach fails to produce a perfect solution even after 50,000 generations [7].

Dynamic Environments The effect of the *dynamic* environment on the adaptability of the animats is shown in Fig. 7. The extreme dynamic case is considered where the effect of actions is changing at every generation. Under these circumstances the models implementing a combined framework of learning and evolution via an indirect encoding of action-selection, manage to produce animats able to adapt to the environment overcoming the initial fluctuations on the maximum fitness of the population.



(a) Average (b) Maximum

Fig. 7. **Average and Maximum Fitness results for the three models on a dynamic environment where the animats are utilizing two drives. Every 1 generation the environment changes, causing the effect of actions to be inverted. Only the models implementing both developmental and genetic approaches are adaptable to the changes and able to achieve consecutive maximum fitness solutions. The Q-Evolution model is unstable.**

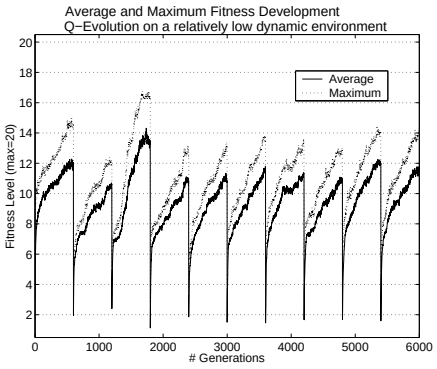


Fig. 8. **The Q-Evolution model implementing an evolutionary approach without developmental learning fails to adapt even to a relative low-dynamic environment that changes every 600 generations. The limitation of the model is due to the lack of a complete valency module. Whenever the environment suffers a change there is a sudden drop on both the average and maximum fitness level of the population**

In contrast, the Q-Evolution model, which implements a strictly evolutionary approach, is unable to adapt to the dynamic environment as it is shown by the low-value and fluctuating average and maximum fitness developments. Even in a dramatically less severe environment, where the changes occur every 600 generations (Fig. 8), evolution alone is unable to follow the changes of the environment and both the average and maximum fitness of the population have a sudden drop at the instance of the change.

3.5 Direct or Indirect Encoding of Action Selection?

Contrary to the results of [11, 14], the average fitness progression of the combined learning and evolution approach does not suffer from large oscillations every time the environment changes. This is due to the fact that action selection is underspecified in the genes and hence the animats do not have to unlearn *and* relearn the right behaviour. They just have to learn it during their lifetime. This demonstrates and proves our hypothesis that underspecified encoding of action selection, in a combined framework of developmental learning and genetic evolution, endows animats with a further adaptive skill that facilitates their survival.

In contrast, animats with an "inborn" skill for selecting and executing a behaviour have to re-learn it at every change of the feedback from the environment. This is a dramatic disadvantage, leading to the animats' extinction when the genetically encoded behaviour becomes a deadly option.

4 Discussion and Conclusion

In the present study we have examined the role of valency as a process relating developmental learning and genetic evolution to assist adaptation. We implemented two different approaches, one that is strictly evolutionary and one that makes use of both developmental and evolutionary mechanisms in order to compare and draw conclusions on the nature of valency. Furthermore, we have tested their performance on both stable and dynamic environments in order to investigate their adaptability.

It has been demonstrated that in both stable and dynamic environments a combined framework of learning and evolution performs better, since agents achieve higher fitness in fewer generations. In the case of an animat equipped with two drives, or in a dynamic environment, evolution alone fails to find a perfect solution, implying that a valency mechanism is necessary if the animats are to adapt at all. Furthermore, we have shown that action selection has to be underspecified in the genome for the sake of adaptation. Instead of directly encoding action selection (as in [1, 14, 11]), the genes should *indirectly* encode that information in order to avoid becoming predisposed toward the execution of a behaviour that could later become harmful.

References

1. Ackley, D. and Littman, M.: Interactions between learning and evolution. In C. Langton, C. Taylor, D. Farmer, and S. Rasmussen, editors, Proceedings of the Second Conference on Artificial Life. California: Addison-Wesley, 1991.
2. Ashby, W.R.: Design for a Brain: The Origin of Adaptive Behaviour. Chapman & Hall, London, 1965.
3. Baldwin, J.M: A new factor in evolution. The American Naturalist, 30 (June 1896):441-451, 536-553, 1896.
4. Batali, J. and Grundy, W.N: Modelling the evolution of motivation. Evolutionary Computation, 4(3):235-270, 1996.
5. Cañamero, D.: A hormonal model of emotions for behavior control. In VUB AIMemo 97U06, Free University of Brussels, Belgium. Presented as poster at the Fourth European Conference on Artificial Life. ECAL 97, Brighton, UK, July 28-3, 1997.
6. Cos-Aguilera, I., Cañamero, D. and Hayes, G.: Motivation-driven learning of object affordances: First experiments using a simulated khepera robot. In Frank Detje, Dietrich Dörner, and Harald Schaub, editors, The Logic of Cognitive Systems. Proceedings of the Fifth International Conference on Cognitive Modelling, ICCM, pages 57-62. Universitäts-Verlag Bamberg, April 2003.
7. Damoulas, T.: Evolving a sense of Valency. Masters thesis, School of Informatics, Edinburgh University, 2004.
8. Harvey, I.: Is there another new factor in evolution? Evolutionary Computa- tion, 4(3):313-329, 1996.
9. Hinton, G.E. and Nowlan, S.J.: How learning can guide evolution. Complex Systems, 1:495-502, 1987.
10. Hull, C.: Principles of Behaviour: an Introduction to Behaviour Theory. D. Appleton-Century Company, Inc., 1943.
11. Mclean, C.B.: Design, evaluation and comparison of evolution and reinforcement learning models. Masters thesis, Department of Computer Science, Rhodes University, 2001.
12. Mitchell, M.: An Introduction To Genetic Algorithms. A Bradford Book, The MIT Press, Cambridge, Massachusetts, London, England, 1998.
13. Nolfi, S.: Learning and evolution in neural networks. Adaptive Behavior, (3) 1:5-28, 1994.
14. Sasaki, T. and Tokoro, M.: Adaptation toward changing environments: Why darwinian in nature? In P. Husband and I. Harvey, editors, Proceedings of the Fourth European Conference on Artificial Life, pages 378–387. MIT Press. 1997.
15. Spier, E. and McFarland, D.: Possibly optimal decision-making under self-sufficiency and autonomy. Journal of Theoretical Biology, (189):317-331, 1997.
16. Sutton, R.S.: Learning to predict by the method of temporal difference. Machine Learning, 3(1):9-44, 1988.
17. Sutton, R.S. and Barto, A.G.: Reinforcement Learning. MIT Press, 1998.
18. Suzuki, R. and Arita, T.: The baldwin effect revisited: Three steps characterized by the quantitative evolution of phenotypic plasticity. Proceedings of the Seventh European Conference on Artificial Life (ECAL2003), pages 395-404, 2003.
19. Toates, F. and Jensen, P.: Ethological and psychological models of motivation - towards a synthesis. In Jean-Arcady Meyer and Stewart W. Wilson, editors, Proceedings of the First International Converence on Simulation of Adaptive Behaviour, A Bradford Book., pages 194-205. The MIT Press, 1990.