On Affect and Self-adaptation: Potential Benefits of Valence-Controlled Action-Selection

Joost Broekens, Walter A. Kosters, and Fons J. Verbeek

Leiden Institute of Advanced Computer Science, Leiden University, P.O. Box 9500, 2300 RA Leiden, The Netherlands

Abstract. Psychological studies have shown that emotion and affect influence learning. We employ these findings in a machine-learning meta-parameter context, and dynamically couple an adaptive agent's artificial affect to its action-selection mechanism (Boltzmann β). The agent's performance on two important learning problems is measured. The first consists of learning to cope with two alternating goals. The second consists of learning to prefer a later larger reward (global optimum) for an earlier smaller one (local optimum). Results show that, compared to several control conditions, coupling positive affect to exploitation and negative affect to exploration has several important benefits. In the alternating-goal task, it significantly reduces the agent's goal-switch search peak. The agent finds its new goal faster. In the second task, artificial affect facilitates convergence to a global instead of a local optimum, while permitting to exploit that local optimum. We conclude that affect-controlled action-selection has adaptation benefits.

1 Introduction

Affect influences thought and behavior in many ways [1,2,3,4]. While affective states can be complex and composed of multiple components, in this paper we use the term *affect* to refer to valence: the positiveness versus negativeness of an agent's affective state (in our case, the agent's mood: a long term, low intensity affective state) [3]. Valence can be seen as a further undifferentiated component of an affective state that defines an agent's situation as good versus bad [5].

We focus on the influence of affect on *learning*. Numerous psychological studies support the idea that enhanced learning is related to positive affect [6], while others show that enhanced learning is related to negative affect [7], or to both [8]. Currently it is not yet clear how affect influences learning. Computational modeling might give insights into the possible underlying mechanisms.

From a machine learning point of view the influence of affect on learning suggests that adaptive agents can benefit from artificial affect, once we know how to (1) simulate affect in a way useful for learning, (2) know what parts of the adaptive agent's architecture can be influenced by artificial affect, and (3) know how to *connect* artificial affect to the appropriate parts of that architecture.

We investigate the relation between affect and learning with a self-adaptive agent in a simulated gridworld. Our agent autonomously influences its action-selection mechanism. It uses artificial affect to control its amount of exploration.

J. Mira and J.R. Álvarez (Eds.): IWINAC 2007, Part I, LNCS 4527, pp. 357–366, 2007.

Our agent uses a standard form of model-based reinforcement learning (see Section 4). We present results based on two different learning tasks. In the first the agent has to cope with two different alternating goals in a two-armed grid-world. We call this the "alternating goal task". This is an important task to be able to learn. An agent with a changing set of goals that has to cope with a dynamic environment has to learn to modify its behavior in order to reflect a change in the set of goals; it has to be flexible enough to give up on an old goal and it has to be persistent enough to continue trying an active goal. In other words the agent has to decide when to explore versus exploit its knowledge, a.k.a. the exploration-exploitation problem [9].

The second task consists of learning to prefer a later larger reward (global optimum) for an earlier smaller one (local optimum). We call this task the "Candy task"; candy represents the local optimum being closest to the agents starting position, while food represents the global optimum being farther away from its starting position. The ability to learn this task is important as it enables survival with the knowledge an agent has, while trying to find better alternatives. Failure to do so results in getting stuck in local optima or slow convergence.

2 The Influence of Affect on Learning

In this section we review some of the evidence that affect influences natural information processing and learning. Some studies find that negative affect enhances learning [7]. Babies aged 7 to 9 months were measured on an attention and learning task. The main result is that negative affect correlates with faster learning. Attention was found to mediate this influence. Negative affect related to more diverse attention, i.e., the babies' attention was "exploratory", and both negative affect and diverse attention related to faster learning. Positive affect had the opposite effect as negative affect (i.e., slower learning and "less exploratory" attention). This relation suggests that positive affect relates to exploitation, while negative affect relates to exploration.

Other studies suggest an inverse relation [6], and find that mild increases in positive affect related to more flexible attention but also to more distractible attention. So it seems that in this study positive affect facilitated a form of exploration, positive affect removes attention bias towards solving the old task.

Of course, attention is not equivalent to learning. It is, however, strongly related to exploration: an important precursor to learning. Flexible distribution of attentional resources favors processing of a wide range of external stimuli. So, in the study by Dreisbach and Goschke [6] positive affect facilitated exploration, as it helped to remove bias towards solving the old task thereby enabling the subject to faster adapt to the new task. In the study by Rose et al. [7] negative affect facilitated exploration as it related to defocused attention.

Other studies, e.g., [8] show that both negative and positive affect can relate to faster learning. The authors found that both flow (a positive state characterized by a learner receiving the right amount of new material at the right speed [10]) and confusion related to better learning.

Combined, these results suggest that positive and negative affective states can help learning at different phases in the process, a point explicitly made in [8]. Our paper investigates this in an adaptive agent context.

3 Simulated Affect Influences Action Selection

To model the influence of affect on learning, we simulate affect as follows. Our agent learns based on reinforcement learning (RL), so at every time step it receives a reward r. Simulated affect is based on this r:

$$e_p = (r_{star} - (r_{ltar} - f\sigma_{ltar}))/2f\sigma_{ltar} \quad (1)$$

Here, e_p is the measure for positive affect, where e_p ranges from 0 to 1, modeling negative affect versus positive affect respectively. The short-term running average reinforcement signal, r_{star}, has a parameter *star* defining the window-size (in steps) of that running average. At every step of the agent, r_{star} is used as input to calculate a long-term running average reinforcement signal, r_{ltar}, with *ltar* a parameter again defining the window-size. The standard deviation of r_{star} over that same long-term period is denoted by σ_{ltar}, and f is a multiplication factor defining the sensibility of the measure. The standard deviation is included as a measure to normalize the affect signal based on the natural variance of r_{star}. Artificial affect measures "how well the agent is doing compared to what it is used to".

Two issues regarding natural affect are important. First, in studies that measure the influence of affect on cognition, affect relates more to long-term mood than to short-term emotion. Affect is usually induced before or during the experiment aiming at a continued, moderate effect instead of short-lived intense emotion-like effect [6,3,7]. This is reflected by the fact that e_p is based on reinforcement signal *averages*, not on r itself.

Second, affect induction (the method used in psychological experiments to investigate the influence of affect on information processing) is compatible with the administration of reward in reinforcement learning. Affect is usually induced by giving subjects small *unanticipated* rewards [1,11]. The reinforcement signal in RL only exists if there is a difference between predicted and received reward. Predicted rewards thus have the same effect as no reward. It seems that both reward and positive affect follow the same rule: if it's predicted it isn't important. This is reflected in our measure. It compares a short-term estimate r_{star} with a long-term estimate r_{ltar}. As the first, short-term average reacts quicker to changes in the reward signal than the second, long-term average, a comparison between the two yields a measure for how well the agent is doing compared to what it is used to. If the environment and the agent's behavior in that environment do not change, e_p converges to a neutral value of 0.5. This reflects the fact that anticipated rewards do not influence affect much.

Our agent uses a Boltzmann distribution to select actions:

$$p(a) = \frac{e^{V_t(s,a)\cdot\beta}}{\sum_{b=1}^{n} e^{V_t(s,b)\cdot\beta}} \quad (2)$$

Here, $p(a)$ is the probability that a certain action a out of n possible ones is chosen, and $V_t(s,a)$ is the value of action a in state s at time t. The inverse temperature parameter β determines the randomness of the distribution. High β's result in a greedy selection strategy (low temperature, small randomness). If β is zero the distribution function results in a random selection strategy, regardless of the predicted reward values (high temperature, high randomness).

In our experiments artificial affect e_p controls an agent's β parameter and thereby exploration versus exploitation. This is compatible with Doya's approach [12], who proposes that emotion is also a system for meta-learning.

3.1 Type-A: Positive Affect Relates to Exploitation

To investigate the influence of artificial affect on exploration, we study two types of relations. Type-A models that positive affect increases exploitation [7]:

$$\beta = e_p \cdot (\beta_{max} - \beta_{min}) + \beta_{min} \tag{3}$$

If e_p increases to 1, β increases towards β_{max}. As e_p decreases to 0, β decreases towards β_{min}. So positive affect results in more exploitation, while negative affect results in more exploration. In essence, our agent is autonomously adapting how selective its attention process is: "when happy, be greedy in the actions to consider, when sad: consider all possible actions equally".

3.2 Type-B: Negative Affect Relates to Exploitation

Type-B models the inverse of the previous relation (as suggested in [6]):

$$\beta = (1 - e_p) \cdot (\beta_{max} - \beta_{min}) + \beta_{min} \tag{4}$$

As affect e_p increases to 1, β decreases towards β_{min} and as ep decreases to 0, β consequently increases towards β_{max}. So positive affect results in more exploration, while negative affect results in more exploitation. In this case our agent uses a different way to adapt attention: "when sad, be greedy in the actions to consider, when happy: consider all possible actions equally".

4 Experiment and Method

Our experiments are performed in two different simulated gridworlds (Fig. 1). The first is a two-armed maze with a potential goal at the end of each arm. This maze is used for the Alternating-Goal (AG) task, i.e., coping with two alternating goals, find food or find water (only one goal is active during an individual trial, goal reward $r = +2.0$). The second maze has two active goal locations. The nearest goal location is the location of the candy (i.e., a location with a reward $r = +0.25$), while the farthest goal location is the food location ($r = +1.0$). This maze is used for the Candy task. The walls in the mazes are *lava* patches, on which the agent can walk, but is discouraged to do so by a negative

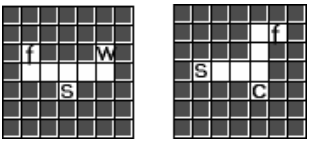

Fig. 1. The task-mazes used. Left maze is the Alternating-Goal task. Right maze is the Candy task. s denotes agent's starting position, f is food, c is candy and w is water.

reinforcement ($r = -1.0$). The agent learns by acting in the maze and perceiving its direct environment using an 8-neighbor and center metric (i.e., it senses its eight neighbors and the location it is at). An agent that arrives at a goal location is replaced to its starting location. Agents start with an empty world model and construct a Markov Decision Process (MDP) as usual (a perceived stimulus is a state s in the MDP, and an action a leading from state s_1 to s_2 is an edge in the MDP). The agent counts state occurrences, $N(s)$, and uses this count in a standard weighting mechanism. Values of states are updated using as follows:

$$R(s) \leftarrow R(s) + \alpha \cdot (r - R(s)) \tag{5}$$

$$V(s) \leftarrow R(s) + \gamma \cdot \sum_i V(s_{a_i}) \frac{N(s_{a_i})}{\sum_j N(s_{a_j})} \tag{6}$$

So, a state s has a learned reward $R(s)$ and a value $V(s)$ that incorporates predicted future reward. $R(s)$ converges to the reward for state s with a speed proportional to the learning rate α. $V(s)$ is updated based on $R(s)$ and the weighted values of the next states reachable by action $a_{1\ldots i}$ (with a discount factor of γ). So, we use a standard model-based RL approach [9]. In the Alternating-Goal task the learning rate α and discount factor γ are respectively 1.0 and 0.7, and in the Candy task respectively 1.0 and 0.8. We have fixed the artificial affect parameters *ltar*, *star* and f to 400, 50 and 1, respectively.

4.1 Learning Tasks

To analyze the difference in learning behavior of agents that use affect control of type-A, B and a control condition using static β values we did the following. In the Alternating-Goal task agents first have to learn goal one (food). After 200 trials the reinforcement for food is set to $r = 0.0$, while the reinforcement for water is set to $r = +2.0$. The water is now the active goal location (so an agent is only reset at its starting location if it reaches the water). This reflects a change in goals, of which the agent is initially unaware. It has to search for the new goal location. After 200 trails, the situation is set back; i.e., food becomes the active goal. This is repeated 2 times, resulting in 5 learning phases (phases 0 to 4 referring to learning of food, then water, food, water, and finally food). This represents 1 run, and we repeated these runs.

The setup of the Candy task is simpler. The agent has to learn to optimize reward in the Candy maze. The problem for the agent is to (1) exploit the local

reward (candy), but at the same time (2) explore and then exploit the global reward (food). This relates to opportunism, an important ability that should be provided by an action-selection mechanism [13].

All AG task results are based on 800 runs, while Candy task results are based on 400 runs. We averaged the number of steps needed to get to the goal over these runs, resulting in an average learning curve. The same was done for β (the *exploration-exploitation* curve), and the agent's quality of life (QOL) (measured as the sum of the rewards received during 1 trial). In all plots the trials are on the x-axis, while β, steps or quality of life on the y-axis.

4.2 Experiment 1: Alternating-Goal Task

Our main finding is that type-A (positive affect relates to exploitation, negative to exploration) has by far the lowest switch cost between different goals, as measured by the number of steps taken at the trial in which the goal-switch is made (Fig. 3b). This is an important adaptation benefit. All goal-switch peaks (phases 1–4) of the 4 variations of type-A (i.e., dotted lines labeled "AG dyn 3–6, 3–7, 3–9, and 2–8") are smaller than the peaks of the controls (straight lines labeled "AG static 3,4,5,6, and 7") and type-B (i.e., striped lines labeled "AG dyn inv 3–6, 3–7, 3–9 and 4–9"). Initial learning (phase 0) is marginally influenced by affective feedback (peaks at phase 0 in Fig. 3b), as can be expected: no goal switch occured before the initial learning phase. Closer investigation of the first goal switch (trial 200, phase 1) shows that the trials just after the goal switch also benefit considerably from type-A (Fig. 2b). When we computed for all settings an average peak for trial 200, 201 and 202 together, and compared these averages statistically, we found that type-A performs significantly better ($p < 0.001$ for all comparisons, Mann-Whitney, $n = 800$). Analysis of the peak at trial 800 (phase 4), reveals about the same picture. The trial in which the goal is switched benefits significantly from type-A ($p < 0.01$) for all comparisons.

All other comparisons between peaks revealed significantly ($p < 0.001$) smaller peaks for type-A (Fig. 3b). This effect is most clearly shown for the peaks of phase 3 and 4, where the relative peak-height difference between type-A peaks and static peaks ranges between 1.25 and 2. This means that using positive affect to control action-selection in the way described in type-A can result in up to a 2 fold decrease of search investment needed to find a new goal. As expected, the smallest difference between control and type-A is when β is small (3 or 4) in the control condition (small β = exploration = less tied to old goal). However, small β's have a classical downside: less convergence due to less exploitation (Fig. 3a). In contrast, type-A curves in Fig. 3a show that the agent does converge to the minimum number of steps needed to get to the goal (i.e., 4).

For completeness we show the β curves for the complete phase 1 of the control group and one type-A and one type-B (Fig. 2a). These curves confirm the expected β dynamics. For type-A, the goal-switch induces high exploration (β near β_{min}) for type-A due to the lack of reinforcement (it is going worse than expected), after which β quickly moves up to β_{max}, and then decays to average. For type-B this is exactly the inverse.

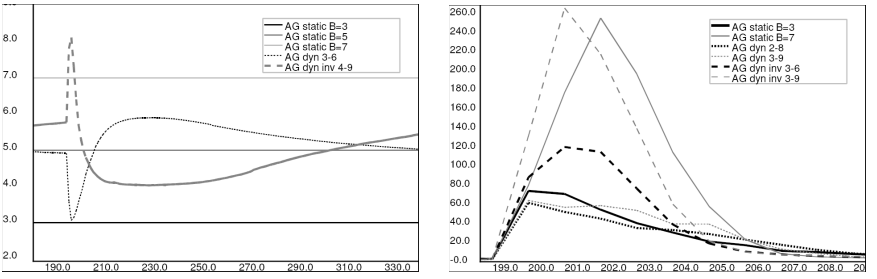

Fig. 2. (a) AG mean β for phase 1. (b) AG mean steps, detail phase 1.

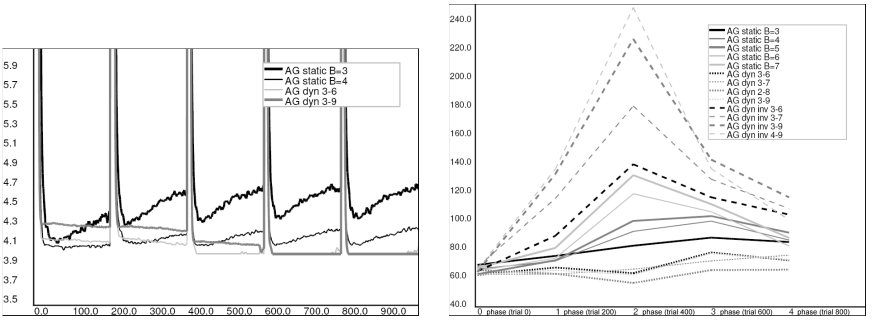

Fig. 3. (a) AG all phases. (b) AG peaks at phases 0–4 (steps in trial 0, 200, 400, etc.).

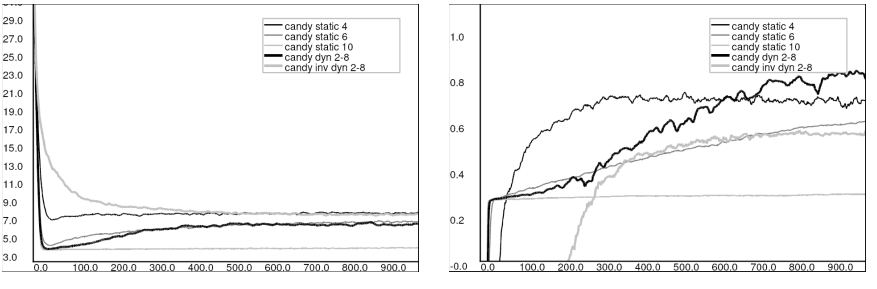

Fig. 4. (a) Candy task complete, mean steps. (b) Candy task complete, mean QOL.

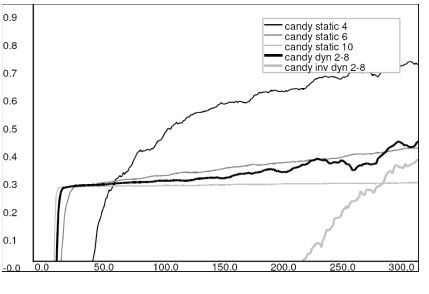

Fig. 5. Candy task starts learning, mean QOL

4.3 Experiment 2: Candy Task

Type-A agents have a considerable adaptation benefit compared to control and type-B agents. In general, type-A results in the same speed of convergence as a high β, as shown by the learning curves of the complete task (Fig. 4a). We see that the learning curves of "β=6" and "dyn 2–8" overlap considerably. Interestingly, the quality of life curves show that in the beginning the QOL of the type-A agent quickly converges to the local optimum (candy, 0.25) comparable to that of the high β control agent (Fig. 5). At the end of the task (later trials) the QOL of the type-A agent steadily increases towards the global optimum (food, +1.0; Fig. 4b). This shows that type-A affective feedback helps to first exploit a local optimum, while at a later stage explore for and exploit a global optimum. This is a major adaptation benefit resulting from affective control of β. This is specifically important for artifical and natural agents in real-world situations. One wants to exploit something good *and* search for something better.

The control agent with $\beta = 4$ does converge to the global optimum just like the type-A agent (Fig. 4b). However, due to continuous high randomness in this agents action-selection mechanism this agent does not converge nicely with regard to the steps they need to take to get that reward as compared to the type-A agent (Fig. 4b). Also due to this high randomness this agent does not learn the local reward consistently enough to quickly exploit it (Fig. 5). For these smaller βs this results in a major delay in arrival at the same level of QOL as compared to the larger βs and the type-A agent (compare the curves "β=5" and "dyn 2-8" in Fig. 5). The type-B agent does not perform well at converging or at quickly exploiting the local optimum (Fig. 4a and b, Fig. 5).

5 Discussion of Results

Although our results show that adaptation benefits from affective control of exploration versus exploitation (and specifically when positive = exploitation), several issues exist. First, we introduce new, non-trivial, parameters to control just one, β. Setting β_{min} to a very small value (i.e., close to 0) results in a problem for type-A agents: once it explores a very negative environment it cannot easily get out of *exploration mode* due to continuously receiving negative rewards that result in even lower affect and thus in even more exploration. The only way to get out of this is by getting completely used to the negative environment, which might take a long time. However, the main benefit of setting meta-parameters such as *ltar*, *star* instead of configuring β is that they can be set for a complete (set of) learning task(s) potentially eliminating the need to adapt β using other mechanism (such as simulated annealing) during learning. Therefore, configuration of these values is more efficient. And, as our results show, the β is controlled in such a way that the agent switches to exploration in the alternating-goal task when needed, and it "gets bored" in the Candy task when needed.

Second, if local and global rewards are very similar, even a type-A agent cannot learn to prefer global reward, as the difference becomes very small. So, the candy and food reward have to be significantly different, such that the agent

can exploit this difference once both options have been found. You don't walk a long way for a little gain. The discount factor is related to this issue, as a small γ results in discarding rewards in the future and therefore the agent is more prone to fall for the nearer local optimum. So γ should be set such that the agent is at least theoretically able to prefer a larger later reward for a smaller earlier one (which is $\gamma = 0.8$ in the Candy task, as compared to 0.7 in the AG task).

5.1 Related Work

Our work relates to computational modeling of emotion and motivation based control/action-selection. It explicitly defines a role for emotion in biasing behavior-selection (e.g, [14]). The main difference is that we have explicitly experimented with a psychologically plausible model of affect as a way to directly, and continuously, control the randomness of action-selection.

Although affective control of exploration is promising for adaptive behavior, our learning model is specific, and our claims hard to generalize. Other learning architectures, such as Soar or ACT-R, should be used to further investigate the mechanisms introduced here. Belavkin [15] has shown using ACT-R that affect can be used to control the search through the solution space, which resulted in better problem-solving performance. He used an information-theoretic approach towards modeling affect that is related to the rule-state of the ACT-R agent. A key difference is thus that our measure for affect is based on a comparison of reinforcement signal averages. Further we explicitly model affect based on different theoretical views on the relation between affect and information processing and compared these views experimentally. The *Salt* model [16] relates to Belavkin's approach in the sense that the agent's effort to search for a solution in its memory depends on, among other parameters, the agent's mood valence.

Schweighofer and Doya [17] used a similar measure for "how well the agent is doing compared to what it is used to"; however, they use it differently. Instead of directly controlling β, affect is used as basis for a search-based method. If a random change to β results in positive affect (agent is doing better), the new β is kept, and vice versa. Recently we have extended this work by comparing these methods on the same tasks in a different learning environment (Soar-RL) [18].

6 Conclusions

We have defined a measure for affect for adaptive agents, and used it to control action-selection. Based on experimental results with learning agents in simulated gridworlds, we conclude that coupling positive affect to exploitation and negative affect to exploration has several important adaptation benefits, at least in the tasks we have experimented with. First, it significantly reduces the agent's *goal-switch search peak* when the agent learns to adapt to a new goal: the agent finds this new goal faster. Second, artificial affect facilitates convergence to a global instead of a local optimum, while exploiting that local optimum. However, additional experiments are needed to verify the generality of our results, e.g., in continuous problem spaces, and other learning architectures (see [18]).

References

1. Ashby, F.G., Isen, A.M., Turken, U.: A neuropsychological theory of positive affect and its influence on cognition. Psychological Review **106** (1999) 529–550
2. Damasio, A.R.: Descartes' error: Emotion, reason, and the human brain. Gosset/Putnam Press, New York (1994)
3. Forgas, J.P.: Feeling is believing? The role of processing strategies in mediating affective influences in beliefs. In: Frijda, N. et al. (Eds.). Emotions and Beliefs, Cambridge, UK: Cambridge University Press (2000) 108–143
4. Phaf, R.H., Rotteveel, M.: Affective modulation of recognition bias. Emotion **15** (2005) 309–318
5. Russell, J.A.: Core affect and the psychological construction of emotion. Psychological Review **110** (2003) 145–172
6. Dreisbach, G., Goschke, T.: How positive affect modulates cognitive control: Reduced perseveration at the cost of increased distractibility. Journal of Experimental Psychology **30** (2004) 343–353
7. Rose, S.A., Futterweit, L.R., Jankowski, J.J.: The relation of affect to attention and learning in infancy. Child Development **70** (1999) 549–559
8. Craig, S.D., Graesser, A.C., Sullins, J., Gholson, B.: Affect and learning: An exploratory look into the role of affect in learning with Autotutor. Journal of Educational Media **29** (2004) 241–250
9. Sutton, R., Barto, A.: Reinforcement learning, An introduction. MIT Press, Cambridge, Massachusetts (1998)
10. Csikszentmihalyi, M.: Flow: The psychology of optimal experience. New York: Harper Row (1990)
11. Custers, R., Aarts, H.: Positive affect as implicit motivator: On the nonconscious operation of behavioral goals. Journal of Personality and Social Psychology **89** (2005) 129–142
12. Doya, K.: Metalearning and neuromodulation. Neural Networks **15** (2002) 495–506
13. Tyrrell, T.: Computational mechanisms for action selection, PhD thesis. University of Edinburgh (1993)
14. Avila-Garcia, O., Cañamero, L.: Using hormonal feedback to modulate action selection in a competitive scenario. In: Proceedings of the 8th Intl. Conf. on Simulation of Adaptive Behavior. (2004) 243–252
15. Belavkin, R.V.: On relation between emotion and entropy. In: Proceedings of the AISB'04 Symposium on Emotion, Cognition and Affective Computing, Leeds, UK (2004) 1–8
16. Botelho, L.M., Coelho, H.: Information processing, motivation and decision making. In: Proc. 4th International Workshop on Artificial Intelligence in Economics and Management. (1998)
17. Schweighofer, N., Doya, K.: Meta-learning in reinforcement learning. Neural Networks **16** (2003) 5–9
18. Hogewoning, E., Broekens, J., Eggermont, J., Bovenkamp, E.G.P.: Strategies for affect-controlled action-selection in Soar-RL, submitted to IWINAC'2007. (2007)