Simulation, Emotion and Information Processing: Computational Investigations of the Regulative Role of Pleasure in Adaptive Behavior

Joost Broekens and Fons J. Verbeek
University of Leiden
Leiden Institute of Advanced Computer Science,
Leiden, The Netherlands.
{broekens, fverbeek}@liacs.nl

Abstract

Emotion plays an important role in thinking. In this paper we focus on the regulatory influence of pleasure on information processing in simulated adaptive agents. Our agent's pleasure is a function of its performance on the tasks it executes in the environment. Our model is based on *Reinforcement Learning* and the *Simulation Hypothesis*. The main hypothesis tested is: *if action-selection-bias is induced by an amount of simulated anticipatory behavior, and if this amount is dynamically controlled by pleasure feedback, then this provides additional survival value to an agent compared to a static amount of simulation*. Experimental results illustrate that this hypothesis holds true. Dynamic adaptation results in a learning performance that at least equals static simulation strategies, and it results in a major decrease of mental effort required for this performance. This is relevant to the evolutionary plausibility of the simulation hypothesis, for increased adaptation at lower cost is an evolutionary advantageous feature. In addition, our results provide clues of a relation between the simulation hypothesis and emotion.

1 Introduction

Emotion plays an important role in thinking. Evidence ranging from philosophy [Griffith, 1999] through cognitive psychology [Frijda, *et al.*, 2000] to cognitive neuroscience [Damasio, 1994; Davidson, 2000] and behavioral neuroscience [Berridge, 2003; Rolls, 2000] shows that emotion—in whatever form—is both constructive and destructive to a wide variety of cognitive phenomena. Normal emotional functioning seems to be necessary for normal cognition.

In this research we focus on the low-level influence of emotion on information processing in simulated adaptive agents. We define emotion as a combination of pleasure and arousal factors [Russell, 2003]. The agent's arousal is based on a metadescription of its memory, e.g., prediction accuracy. Pleasure is a function of the agent's relative performance on the tasks it executes in the environment. The agent uses Reinforcement Learning (RL) [Sutton and Barto, 1996]. In this paper we focus on the influence of pleasure as feedback to control the amount of simulated anticipatory behavior the agent uses to bias action selection. This influence is measured in terms of learning performance and total effort spent on simulated and overt interaction. Thus, we investigate the influence on learning if emotion is used to control the cognitive mechanism (i.e., simulation) that biases action-selection. We do not model categories of emotions nor use such emotions as information in symbolic-like reasoning. Reasons for our low-level approach include:

First, because emotion is integrated at multiple levels of processing and higher—conscious, reflective reasoning—levels have not always existed throughout evolution, one would expect an evolutionary advantage to integration at levels close to reward systems and behavioral control. On higher levels, emotion *regulates* information processing. Could emotion play such role at lower levels?

Second, from a computational point of view lower levels tend be more generic. Therefore, regulative mechanisms found can be applied to a wider area of disciplines including cognitive science and machine learning, for example meta-learning—how to autonomously monitor and, if necessary, adapt the learning mechanism used by the agent in order to better cope with the current task. If emotion is considered as a meta-learning system [Doya, 2000], it can be used to enhance artificial adaptive agents in a generic way. Regulative mechanisms that operate on higher cognitive levels may need a more complex concept of emotion or a dedicated cognitive architecture, and are therefore less generic.

Third, a low-level interpretation allows us to stay close to behavioral control and action-selection mechanisms thereby avoiding philosophical debates about emotion. Consequently, we use a modest—but broadly usable and less controversial—concept of emotion as basis for the research.

Fourth, Montague *et al.* [2004] recently argued that computational models of RL can be used to model and understand behavioral control, and to gain insights into the neurophysiological aspects of psychiatric disorders. By computationally studying how emotion relates to information processing and reinforcement we hope to extend the analogy between RL and behavior.

To study the low-level regulatory influence of emotion on information processing, we use a computational RL model. Besides RL, our approach is based upon the following hypotheses. **1.**) The *Simulation Hypothesis*, which assumes that thinking is internal simulation of behavior using the same sensory-motor systems as those used for overt behavior [Hesslow, 2002] **2.**) *interactivism*, stating that thinking emerges from continuous interaction with the environment [Bickhard, 2001].

These hypotheses have several important characteristics in common [Broekens, 2005b], amongst which the following are particularly important for this paper:

a.) These hypotheses are primarily about neuronal systems, but do allow connectionist but non-neuronal modeling, the basis of our model.

b.) Emotion plays a role in information processing.

c.) These hypotheses closely relate to Damasio's [1994] concept of thinking as an "as-if body loop", involving simulated actions that are evaluated by their *somatic markers*, emotional impact estimators. Four systems are critically involved: the body; the somato-sensory cortex (SSC), the emotional marker system that receives information from the body; the sensory and association cortexes (SC/AC); and the ventromedial prefrontal cortex (VM-PFC), the system that stores relations between factual representations stored in the SC/AC and somatic markers stored in the SSC. Interaction with the environment enables the VM-PFC to learn these links. Two important processing mechanisms are the "body-loop" and the "as-if body loop". When facts about a situation are recognized, the SC/AC activate the VM-SSC, and links between the situational facts and emotional outcomes are activated. In the "body-loop", the VM-SSC activates the body, and the SSC that stores somatic-markers is organized according to the body. This loop thus involves the emotional evaluation of action. In the "as-if body loop", the VM-PFC signals the SSC to reorganized itself directly without signaling the body. This loop thus involves the emotional evaluation of simulated action. The "as if" loop produces imagined future factual-emotional states, and the somatic marker part of such states is the state's predicted accumulative emotional outcome (reward/punishment). This marker signal is used to bias decision-making [Damasio, 1994]. Even though we do not model the body of the agent, we use the somatic marker concept to understand the relation between reinforcement learning (RL), emotion and decision-making.

In this paper we first introduce our computational approach without emotional feedback. Next, we introduce our concept of emotion and pleasure in more detail, and we explain how pleasure is used to control the amount of anticipatory simulation of the agent. Finally, we discuss our results, related work and give directions for future research.

2 Computational Approach

Our experiments are performed in a gridworld, a two-dimensional grid with positively and negatively reinforced locations, in our case, lava (negative reinforcement of –1), roadblocks (–0.5), food (+1.0) and empty cells (Figure 1). The agent can move everywhere, but is discouraged to walk on the lava (by a negative reinforcement). The agent's perceptual field has either a chessboard, 8 neighbor (Figure 1b), or a cityblock, 4 neighbor metric (Figure 1a, c). In, e.g., Figure 1c, the agent would perceive "eleee" representing the (l)ava left of the agent and the (e)empty cells above, right, beneath, and below the agent.

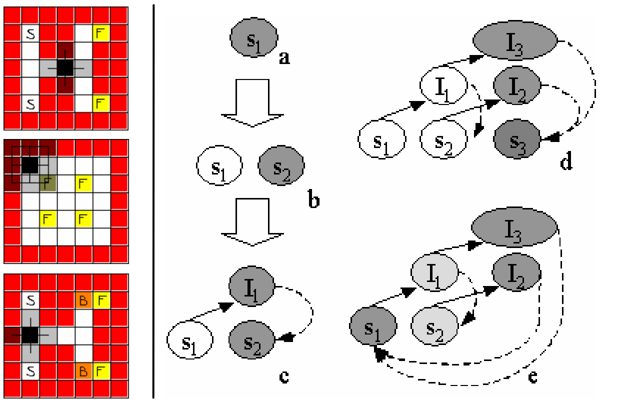

Figure 1 (left) and 2 (right). Fig. 1: three different experimental settings: agent (black), lava (dark gray, red), possible food (*F*), roadblock (*B*), possible start location (*S*). Tasks from left to right: *find food*, *forage*, *invest*. Fig. 2: examples of the agent's memory.

2.1 Hierarchical-State Reinforcement Learning

We first explain the basic model without emotional feedback. The agent's memory structure is modelled by a directed graph. The memory is adapted while the agent interacts with its environment (online learning) in the following way. The agent selects an action, $a \in A$, from its set of potential actions $A=\{u, d, l, r\}$, executes the action in the gridworld and perceives the result of that action, *p*. This is combined into a *situation*, $s=<a, p>$, that is stored in the agent's memory according to a basic rule: *if a situation s occurs, the agent creates a node in the graph if and only if there does not exist a node for s*. For example in Figure 1c, if the agent has moved down, "d", and perceives "eleee". In an initially empty model a node is created to represent the situation $s_1=<d,eleee>$ (Figure 2a), because the graph does not yet contains this node. Now the agent moves again, resulting in a new situation, e.g., $s_2=<d,elele>$, resulting in a new node that represents s_2 (Figure 2b). To model that s_2 follows s_1 (or s_1 predicts s_2), the previous situation, s_1, is now connected to the current situation, s_2, by creating a new node, an *interactron,* between s_1 and s_2 with edges as shown in Figure 2c. This process continues, never violating the basic rule. Also, the process is recursively applied to active interactrons. Active in this case means that an interactron corresponds to the history of observed situations, e.g., node I_1 in Figure 2c. If situation s_2 is followed by s_3, the resulting memory structure is shown in Figure 2d, with active nodes s_3, I_2 and I_3. If, on the other hand s_2 is followed by s_1, the resulting structure is shown in Figure 2e, with active nodes s_1, I_2 and I_3.

If at a later time the sequence of situations s_1s_2 is again observed then, according to the rule, I_1 is not created again. Instead, a counter υ, the *usage* of interactron I_1, that is initially zero is increased by one. This υ can be used to calculate the probability $P(s_2 \mid s_1)$ using the following more generic formula:

$$\mathrm{P}(x \mid y) = \upsilon_x \Big/ \sum_{i=1}^{|X_y|} \upsilon_{x_i}$$

,where y is an active interactron or situation, $x \in X_y = \{x_1, \ldots, x_n\}$ the set of predicted situations by y (represented by their corresponding interactrons, e.g., I_1 representing the prediction of s_2). This formula is true, for I_1 is conditionally active upon s_1, and υ is only increased if an interactron is active and multiple sequences other than s_1s_2, e.g., s_1s_3, s_1s_4 etc., have their own interactron attached to s_1 with its own υ increased if and only if the corresponding sequence is observed. Furthermore, we define a threshold, θ, representing the minimal "survival probability" for an interactron. If $P(x \mid y) < \theta$, the corresponding interactron is forgotten and removed from the memory, including its dependencies. This corresponds to Bickhards [2000] notion of interaction (de)stability based on consistent confirmation of predicted interactions, see also [Broekens and DeGroot, 2004].

The memory maintains a distributed, hierarchical prediction of the next situation. Every active interactron predicts potential next situations, k of these interactrons can be active, and the 1st till k-th interactron predict potential next situations with a history of length 1 till k respectively (e.g., I_3 is a $k=2$ interactron with history s_1s_2). Learning in the context of this memory can be seen as the online learning of $1 \ldots k$-th order Markov Decision Processes in parallel.

In addition to a predictive probability, every interactron has a reinforcement value, called a *marker*, μ, with $\mu = \lambda + \nu$, where λ is the interactron's *direct reinforcement* value and ν is a back-propagated *indirect reinforcement* value. Thus, the value of an interactron is a function of it's own reward and the rewards of those situations it predicts. More specific, first, all k active interactrons are reinforced by a signal from the environment, r^t, at time t. For every such interactron y, λ_y, is adapted according to the formula:

$$\lambda^{t+1}{}_y = \lambda^t{}_y + (r^t - \lambda^t{}_y) \times \rho$$

, where ρ is the agent's learning rate. Second, for every interactron y, ν_y, is calculated as follows:

$$\nu^{t+1}{}_y = \sum_{i=1}^{|X_y|} \mu^t(x_i \mid y) \times \mathrm{P}(x_i \mid y)$$

, where $\mu^t(x_i \mid y)$ is defined as the marker of interactron x_i, with x_i predicted by y. This indirect part of an interactron's (say y) value is thus the weighted average of the markers belonging to the interactrons X_y that represent the situations that y predicts, where weighted is according to the probability distribution $P(x_i \mid y)$ over all i.

Action-selection is based on the parallel inhibition and excitation of actions in the set of actions, A. The inhibition/excitation originates from the k active interactrons and is calculated using the formula:

$$l^t(a_h) = \sum_{i=1}^{k} \sum_{j=1}^{|X_{y_i}|}{}^{*} \mu^t(x^i{}_j \mid y_i) \times \mathrm{P}(x^i{}_j \mid y_i)$$

, where $l^t(a_h)$ is defined as the level of activation of an action $a_h \in A$ at time t, y_i an active interactron, and x^i_j predicts action a_h. This last clause is needed, for the memory stores action-perception pairs and any of these pairs that are predicted by any of the k active interactrons should inhibit (negative marker) or excite (positive marker) the corresponding action, but not other actions. Additionally, of all good actions (any $l^t(a_h)>0$) the best action a_h, i.e., $l^t(a_h)=max(l^t(a_1),\ldots,l^t(a_{|A|}))$, is always selected. If there are only bad actions (all $l^t(a_h)<0$) a stochastic selection is made based on $(l^t(a_1),\ldots,l^t(a_{|A|})$; the action with the highest activation therefore has the highest chance of being chosen resulting in a probabilistic Winner-Take-All action-selection.

The process described in this section is our agent's "body loop". Next, we describe our agent's "as-if" loop, its simulation mechanism. For a discussion on the relation between Damasio's somatic marker hypothesis and our computational model, see [Broekens, 2005b].

2.2 Internal Simulation and Action-Selection Bias

To study anticipatory simulation we add the following capability to our model: after every real interaction with the environment, the model simulates one time-step ahead. Instead of selecting an action based on past interactions the following process is executed:

1.) *Interaction-selection*: at time t select a subset of to-be-simulated interactions from the set of interactions predicted by all k active interactrons.

2.) *Simulate*: send the subset of selected interactions to the model as if they were real interactions. The memory advances to time $t+1$.

3.) *Reset-state*: to be able to select an appropriate action, reset the memory's state (the active interactrons) to the previous timestep, i.e., time t.

4.) *Action-selection*: select the next action using the standard mechanism described above. Thus, the propagated markers of the simulated predicted interactions directly bias action-selection. Our anticipation mechanism is best understood as *state anticipation* [Butz *et al*, 2003].

5.) *Reset-markers*: reset μ, λ and ν of the interactions that were changed at step 2 (simulation) to the values of μ, λ and ν of these interactions before step 2.

Step 1 selects predicted interactions to be simulated, and is a critical component in our simulation mechanisms since it defines the amount of internally simulated information. In a previous experiment [Broekens, 2005] we used four static selection criteria (also referred to as *simulation strategies*).

a.) No simulation (NON). The actions are selected as described in the previous section and the 5-step simulation procedure is not executed. **b.**) Simulation of the predicted best interaction (BEST). The winning interaction of the WTA selection resulting from step 1 is sent to the model for simulation (step 2). Any real interaction is accompanied by a reinforcement signal. As this is a simulation we lack such a signal. Instead, this signal is simulated using the μ of the winning interaction as reinforcement. We simulate the predicted interaction and its associated value. **c.**) A selection of the predicted 50% best interactions, i.e., a more balanced selection, (BEST50). Again we simulate the reinforcement signal using the μ's of the simulated interactions. **d.**) All of the predicted interactions (ALL).

In essence, NON, BEST, BEST50 and ALL simulate different values for the *selection threshold* of the WTA interaction selection ranging from infinite (NON) to high (BEST) to medium (BEST50) to low (ALL). This threshold filters the set of predicted interactions used to simulate. The final result of simulation is a bias to the predicted rewards of the set of next possible interactions, with action-selection based on these biased rewards (Figure 3). This means that our model of internal simulation influences action-selection in a way that is compatible with the somatic marker hypothesis [Damasio, 1994] and the simulation hypothesis [Hesslow, 2002]. For more on the compatibility between our model and the simulation hypothesis see [Broekens, 2005].

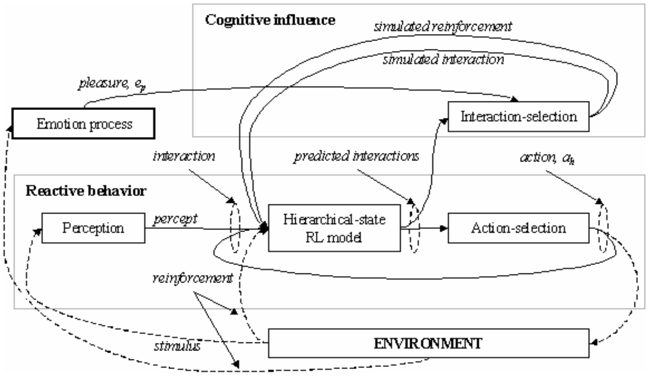

Figure 3. Architecture of the different components in our model.

2.3 Differences in Performance of Static Simulation Strategies Motivate the Feedback Control

In a previous study [Broekens, 2005] we showed that simulation in general, and simulation of all possible next interactions (ALL) in particular, has a clear adaptive advantage. The agent learns the tasks quicker and converges better to the solution. The agent had to learn three tasks (Figure 1):

1.) Continuously try to *find* a randomly changing food location, thereby learning the optimal route to both possible food locations in the gridworld maze (Figure 1a).

2.) Learn to *forage* (Figure 1b). Now, the agent is initially placed in the environment, after which it should explore and find food. Again, food locations are randomly selected.

3.) The same as the first, but the agent additionally had to learn to accept an initial negative reinforcement (roadblock in Figure 1c) in order to get to a larger positive one (food in Figure 1c). With this task we wanted to test how the different simulation strategies handle *investment*, which is a relevant problem for natural adaptive agents [Doya, 2002].

Intuitively it is not really a surprise that ALL "wins", as it is the heuristic using the most information. However, for some experimental settings BEST or BEST50 do result in a better performance (i.e., a smaller amount of simulation results in a better performance). This suggested a relation between the parameters of the experimental setting, and the effect of the amount of simulation used by the agent.

Analysis of this relation revealed that the *goal orientedness* of the task and the *complexity of the task* influence this performance. When the agent is solving a goal oriented task (*find food*, *invest*), it benefits from a narrow (i.e., BEST) simulation strategy with a high learning rate, while in an uncertain or more exploratory task (*forage*) it benefits from a broad (i.e., BEST50 or ALL) simulation strategy.

Simple goal-oriented tasks are solved by quickly propagating the delayed reward to the beginning, specifically if there is "just one hill to climb". Local solutions converge to a global solution. The faster the convergence the quicker the global solution is found, as reflected by previous results.

If a task is complex, the agent benefits from broader simulation, for this allows it to mentally explore multiple options and make a more balanced choice. This relates to the exploration-exploitation problem [cf. Doya, 2002]. Essentially our agent has to vary its *simulation strategy* (instead of its action selection) between mental exploitation and mental exploration.

These findings suggested that it is beneficial to the agent to dynamically adapt simulation to accommodate the task. Additionally, we hypothesized that dynamic adaptation of simulation could outperform any of the four static strategies tested, for dynamic adaptation could be beneficial to the agent at *different stages of learning a task*. The main hypothesis addressed in this paper is: *if action-selection-bias is induced by an amount of simulated anticipatory behavior, and if this amount is dynamically controlled by pleasure feedback, then this provides additional survival value to an agent, compared to a static amount of simulation*. Our approach is compatible with Cañamero's [2000] view on why and how emotion systems should be designed.

3 Emotion as Pleasure and Arousal Factors That Control Information Processing

Before describing how we add emotional feedback to the simulation mechanism, we present some rationale for our concept of emotion. Emotion influences thinking. This influence is found at low and high levels of information processing and is both positive as well as negative. For example, at the neurological level malfunction of certain brain areas not only destroys or diminishes the capacity to have (or express) certain emotions but also has the same effect on the capacity to make sound decisions [Damasio, 1994] and on the capacity to learn new behavior [Berridge, 2003], which indicates that these areas are linked to emotions as well as "classical" cognitive and instrumental learning phenomena. At the cognitive psychological level a person's beliefs about something are updated according to the emotion. The current emotion is used as information about the perceived object [Clore and Gasper, 2000; Forgas, 2000], and emotion is used to make the belief resistant to change [Frijda and Mesquita, 2000]. Emotions are "at the heart of what beliefs are about" [Frijda *et al.*, 2000]. For example, your belief about roller coasters tells you something about the emotion attached to your cumulative experiences with roller coasters.

More specifically, emotion is related to the regulation of adaptive behavior and to information processing. Emotions can be defined as states elicited by rewards and punishments [Rolls, 2000]. Behavioral evidence suggests that the ability to have sensations of pleasure and pain is highly connected to basic mechanisms of learning and decision-making [Berridge, 1998; Cohen and Blum, 2002]. Behavioral neuroscience teaches us that positive emotions reinforce behavior while negative emotions extinct behavior, so at this lower level one type of regulation of behavior has already been established—i.e., approach versus avoidance. The emotion resulting from an unconditioned natural stimulus is associated with the conditioned stimulus or with a specific action. In the future, upon presentation of the conditioned stimulus to the animal, this association results either in more actively *choosing the action* that leads to the unconditioned stimulus (rats' lever pressing behavior) or in *behavior that is associated* with the unconditioned stimulus (Pavlov's dog producing saliva). At this lower level, emotion has a direct—mostly associative—effect (but also other effects are reported [Dayan and Balleine, 2002]).

At the higher level of cognitive psychology, evidence suggests that the processes involved in emotion are crucial for both evaluating the world around us at different levels of abstraction [Scherer, 2001] as well as actually taking action [Frijda, 2000]. Emotion also plays a role in the regulation of cognitive processes. Scherer [2001] argues that emotions are related to the continuous checking of the environment for important stimuli. More resources are allocated to further evaluate the implications of an event, only if the stimulus appears important. This suggests that certain emotions are related to regulation of the *amount of information processing*. This finding provides an important clue to our approach of adding emotional control to the amount of simulation used by the agent. Furthermore, in the work of Forgas [2000] the relation between emotion and information processing strategy is explicit: depending on the strategy used, the influence of mood on thinking changes.

Although many different emotions (and emotion theories) exist, and emotion consists of many different components—e.g., facial expression, a tendency to act, subjective evaluation of the situation—, the *core-affect* theory of emotion states that emotion (mood) consists of two fundamental factors, *pleasure* and *arousal* [Russell, 2003]. Pleasure relates to emotional valence, while arousal relates to action-readiness, or activity, of the organism. Many different situations can be emotionally described using these two factors, for example, winning the lottery (a high arousal high pleasure emotion), or losing a friend (a low arousal and low pleasure emotion). Although Mehrabian [1996] argues for dominance as a third factor, he agrees with, and shows considerable evidence for, the pleasure and arousal factors.

Certain cognitive appraisal theories argue that pleasure and arousal can be produced by very simple stimulus checking functions. This suggests that low-level mechanisms like intrinsic pleasantness checks and suddenness checks are involved [Scherer, 2000].

The suggestion that pleasure and arousal factors are fundamental to emotion, that these factors can be produced by simple mechanisms and that these factors can influence further information processing inspired us to look at how these two factors could result from low-level features of the agent's memory structure and its performance, and subsequently how these factors could then influence information processing in a way that is compatible with cognitive appraisal theory. In this paper we focus on the pleasure factor.

3.1 Pleasure As a Measure for Relative Task-Performance

According to cognitive appraisal theory positive emotions are related to top-down goal oriented processing while negative emotions are related to bottom-up stimulus oriented processing [Fiedler and Bless, 2000]. Furthermore, emotion is often seen as an indication of the current performance of the agent [Clore and Gasper, 2000]. To capture these findings we measure pleasure in the following way:

$$e_p = (\bar{r}_{star} - (\bar{r}_{ltar} - f\sigma_{ltar}))/2f\sigma_{ltar}$$

The current pleasure, e_p, of the agent is the short-term running average over the reinforcement signal, r, with a window size of *star* steps, normalized around the agent's long-term running average over the same reinforcement signal with a window size of *ltar* steps. This value is normalized using f times the standard deviation of the long-term distribution of reinforcement signals σ_{ltar}. So, e_p is a continuous measure for how well the agent is currently performing on a task, relative to what it is used to, according to the recent past. A large f results in smaller fluctuations around 0.5, while a small f results in larger fluctuations around 0.5. Also, e_p is clipped between 0 and 1. Information processing can be influenced by e_p in the following way (Figure 3). When $e_p=1$, interaction-selection (Step 1) selects only the best interactions for simulation, i.e. a high selection threshold. When $e_p=0$ it selects all interactions, i.e., a low selection threshold. The agent thus varies between BEST and ALL depending on its pleasure. It can be argued that our use of pleasure relates more to mood than to emotion, due to its timescale. Moods typically occur at longer timescales, while emotions are short complex reactions to events. Pleasure in our case is measured over multiple interactions and does not react to one interaction in particular. Even if e_p is interpreted as the agent's mood, the modeled effects of positive versus negative emotion is consistent with the previously mentioned ideas about top-down versus bottom-up processing related to respectively positive and negative emotions as well as to the concept of emotion influencing the amount of processing needed. If the agent goes well, little processing (focussed attention) is needed, if it goes bad more processing (broad attention) is needed.

4 Experimental Setup

To test our hypothesis we created a combined task in which simple and complex elements are present as well as goal oriented and exploratory behavior is needed. The first half consists of the *find food* task (Figure 1a), and the second half consists of the *invest* task (Figure 1c). The agent is unaware of this change; it is abruptly replaced in a slightly different environment and has to learn about this change by interacting with the environment. The hypothesized effect is that the agent dynamically adapts the amount of simulation according to the change in complexity and goal orientedness. We predicted the following changes to simulation during the task: BEST→ALL→BEST. BEST performs best on the goal-oriented *find food* task. The change to the *invest* task induces a pleasure decrease, resulting in simulation close to ALL: mentally explore the new task. During learning of the *invest* task, simulation should return to one that is close to BEST because the agent's pleasure increases, resulting in goal oriented behavior of the agent.

f:	*1*		*1.5*		*2*	
star:	*50*	*100*	*50*	*100*	*50*	*100*
ltar:	200	400	200	400	200	400
	250	500	250	500	250	500
	375	750	375	750	375	750
	500	1000	500	1000	500	1000
	750	1500	750	1500	750	1500

Table 1: *ltar*, *star*, and *f* configurations used in the experiment.

One experimental setting is a combination of *f*, *star*, *ltar,* θ and ρ. These parameters are varied as follows: the forgetting rate θ*=(0, 0.01, 0.03, 0.05)*, learning rate ρ*=(1, 0.8)* and *ltar*, *star* and *f* according to Table 1. For every experimental setting the agent had 255 *trials* (defined as one *run*) to get to the food. It had to learn the task within these 255 trials, which showed to be enough to conclude convergence.

For every experimental setting, we recorded the agent's total number of actions needed to complete a run (i.e. 255 trials), and averaged over 15 runs. This resulted in averages for *5×6=30* (*f*, *star*, *ltar*) configurations per (θ, ρ) configuration. The goal of these experiments is not to find out what the exact parameters are to get the best dynamic result, but to investigate the potential benefit of pleasure controlling simulation effort in general. We assume that there should be an *overall benefit* to emotional feedback. Therefore, averaging again aggregates these 30 averages. The result is one value per (θ, ρ) depicted by the red (gray) lines in Figure 4. Red lines should be interpreted as the average performance of an agent that uses emotional feedback to dynamically control the amount of simulation (DYN). Performance is in terms of the total *number of interactions* needed to complete a run (Figure 4a and c), and *mental effort* in terms of the total number of simulated interactions needed to complete a run (Figure 4b and d). Black lines show the corresponding performance of the static strategies (NON, BEST, BEST50, ALL) averaged over 30 runs per (θ, ρ) configuration.

5 Results

The performance of our dynamically adapting agent is comparable to (Figure 4a and c), and in several special cases even better than (Figure 4e, result of one setting averaged over 30 runs instead of 15), the performance of our static agents. If this effect is put in light of total simulation (mental) effort, it is even more dramatic. DYN uses about 33% of the mental effort needed for ALL and about 70% of the effort needed for BEST50 but performs comparably. The predicted effect of the pleasure feedback is confirmed. Figure 5 depicts a typical pleasure flow (15 run e_p average) of an agent that uses DYN. Just after the task switch (at trial 128) a steep decrease of pleasure is observed, this results in more simulated interactions, i.e., broader attention. While exploring, the agent improves at the *invest* task, and pleasure gradually increases, resulting in goal-directed simulation.

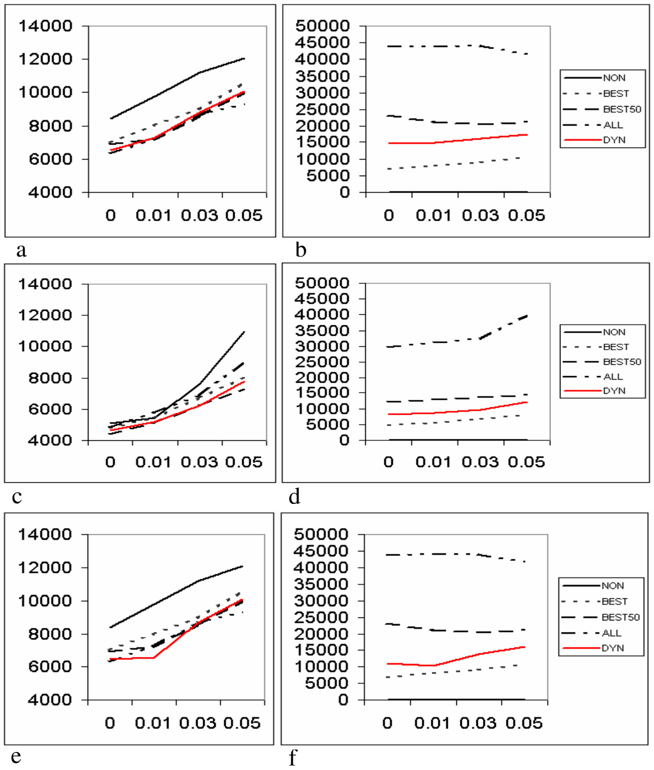

Figure 4. Figure 4a and b ρ=0.8, 4c and 4d ρ=*1*. Figure 4e,f, DYN (*star=100, ltar=1500, f=1*) performing better (one-tailed *t*-test, *n*=30, α=0.05) than static strategies with ρ=*0.8 and* θ=*0.01.*

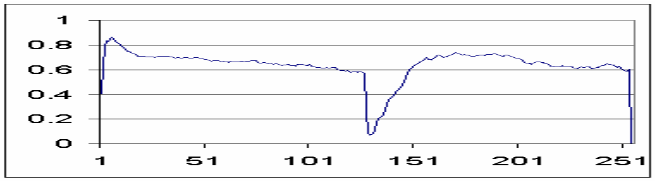

Figure 5. Pleasure flow during one run, averaged over 15 runs.

6 Discussion and Conclusions

Under the assumption that total simulation effort positively correlates with total energy consumption of the agent, decrease of mental effort reduces the energy need for information processing, thereby saving energy for occupancies other than foraging. If dynamic adaptation reduces mental effort and if this is an hereditary feature, it becomes evolutionary advantageous. This suggests that dynamic adaptation of the amount of simulation has a strong evolutionary drive.

Our results show that the relation between (1) positive emotions and top-down goal oriented thinking, and (2) negative emotions and bottom-up stimulus driven thinking could result from the feedback of a simple measurement of the performance of the agent to the selection threshold of the simulation mechanism. These results show one possible relation between emotion and the simulation hypothesis, as well as provide experimental evidence for the fact that even simple emotional integration processes can be used to adapt cognitive processes.

6.1 Related Work

Our work is highly related to Gadanho's [2003] work on the "Alec" architecture. However, in their RL based adaptive system, stochastic action-selection is biased by a fixed value produced by a rule-based cognitive system. In our system this value is dependent on the predicted states and the cognitive process is not separated from the adaptive system. We chose not to separate the cognitive system from the reactive system, as this is important for the evolutionary continuity between simulating and non-simulating agents [Broekens, 2005; Cruse, 2002; Hesslow, 2002].

The "Salt" model by Botelho and Coelho [1998] relates to ours 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. Our approach differs in that we focus on simulation of behavior (not specifically targeted towards search), we use a dynamic influence to link emotion to the cognitive system (not a rule-based system), and we specifically define how our agent's mood is produced.

Our work relates to emotion and motivation based control/action-selection, in that it explicitly defines a role for emotion in biasing behavior-selection [Avila-Garcia and Canamero, 2004; Canamero, 1997; Velasquez, 1998]. The main difference is that in these studies emotion directly influences action-selection (or motivation(al states)), while we have studied the indirect effect of emotion as a metalearning parameter affecting information processing that on its turn influences action-selection (cf. Gadanho [2003]).

Up until now our agent is unable to learn the representation of a goal (what is a goal) and thus is unable to consider different goals in its final action selection. We learn from behavioral neuroscience that rats adapt learned behavior contingent on their drives (i.e., lever-pressing when hungry versus button-pushing when thirsty) [Dayan and Balleine, 2002]. They argue that the rat's motivation acts as a gate between the learned predictive state and the incentive value associated with it. Such a mechanism can be implemented using a Markov Decision Process [Smith *et al,* 2003]. They model a conditioning task whereby the learned reward is multiplied by an artificially varied "gating factor", i.e., a simulated dopamine signal that is necessary for the agent to see the consequences of its actions.

However, implementations such as [Smith *et al*, 2003] are still limited since many animals develop multiple complex goals, suggesting that they can learn to use many representations as gating factor for the predicted reinforcement signal in a certain situation. In this case, a learned goal can influence behavior without the behavior being directly associated with a positive or negative reinforcement signal. Learned goals could even become reinforcers by themselves. This approach relates to one proposed by Singh *et al.* [2004], where multiple different reinforcement techniques are used to learn hierarchical collections of skills that function as intrinsically motivating actions for the agent. Further, it relates to work by Gadanho [2003], where multiple goals—related to homeostatic variables—determine the reinforcement for the adaptive system, and to work on emotion learning by, for example, Botelho and Coelho.

6.2 Future work

We have investigated one way in which pleasure can influences information processing. Combining *arousal and pleasure* as feedback to control simulation might give additional insights into the relation between these two factors, as well as introduce a second learning metaparameter.

To measure arousal, the agent could compare to what extend the predicted environment equals the actual environment. This measurement is called the stimulus predictability check [Scherer, 2000]. We can implement this in our model by comparing the probabilities of next interactions with the actually occurring interactions.

Another way to measure arousal is the stimulus familiarity check [Scherer, 2000]. This check measures how much of the environment is actually known. In our model we can count the number of active interactions in the state hierarchy (high number = familiar, low number = unfamiliar).

These two arousal measurements can be integrated into one signal, say e_a that, e.g., influences the absolute amount of effort put in simulation (information processing). A high e_a results in a large amount of effort put into simulation, while a low e_a results in a low amount of effort. The e_a factor combined with e_p results in a distribution of maximum available simulation steps over the potential next interactions. Along these lines, we plan to adapt our model so that it is able to simulate multiple steps ahead depending on a cut-off depth based on the total amount of effort available for that specific branch. This approach is highly similar to planning and algorithms for depth-first, breadth-first and iterative deepening search. We hope that techniques following from our research are generic in terms of their ability to modify solution-search behavior in these kinds of algorithms.

A different way to influence simulation is by letting e_a control the amount of randomness in the interaction selection process. This is analogous to the role of noradrenaline as proposed by Doya [2002].

6.3 Conclusions

Experimental results show that if pleasure is used to dynamically adapt the amount of simulation, this results in a learning performance that, at least, equals static simulation strategies. Importantly, our results show a major decrease of mental effort required for this performance. This observation is relevant to the understanding of the evolutionary plausibility of the simulation hypothesis, as increased adaptation at lower cost is an evolutionary advantageous feature. In addition, our results provide clues of a relation between the simulation hypothesis and emotion theory.

Acknowledgements

We thank Jeroen Eggermont for useful discussions, and the MNAS reviewers and Walter Kosters for useful suggestions.

References

[Avila-Garcia and Cañamero, 2004]. O. Avila-Garcia and L. Cañamero. Using hormonal feedback to modulate action selection in a competitive scenario. In: *From Animals to Animats 8: Proc. 8th Intl. Conf. on Simulation of Adaptive Behavior.* MIT Press, Cambridge, Massachusetts.

[Berridge, 2003] K. C. Berridge. Pleasures of the brain. *Brain and Cognition 52.*

[Botelho and Coelho, 1998]. L. M. Botelho and H. Coelho. Information processing, motivation and decision making. In: *Proc. 4th International Workshop on Artificial Intelligence in Economics and Management.*

[Butz *et al.*, 2003] M. V. Butz, O. Sigaud and P. Gerard. Internal models and anticipations in adaptive learning systems. In: *Anticipatory Behavior in Adaptive Learning Systems.* Springer (LNAI 2684).

[Bickhard, 2000] M. H. Bickhard. Motivation and emotion: an interactive process model. In: *The Caldron of Consciousness*. John Benjamins, New York.

[Broekens and DeGroot, 2004] J. Broekens and D. DeGroot. Emergent Representations and Reasoning in Adaptive Agents. In: *Proc. ICMLA'04*. IEEE.

[Broekens, 2005] J. Broekens. Internal simulation of behavior has an adaptive advantage. In: *Proc. CogSci'05*. (in press).

[Broekens, 2005b] J. Broekens. Computational Investigations of the Regulative Role of Pleasure in Adaptive Behavior. TR 2005-06, LIACS, Leiden University. http://www.liacs.nl/~broekens/BroekensTR2005-06.pdf.

[Cañamero, 2000]. D. Cañamero. Designing emotions for activity selection. *Dept. of Computer Science Technical Report DAIMI PB 545*. University of Aarhus, Denmark.

[Clore and Gasper, 2000] G. L. Clore and K. Gasper. Feeling is believing: some affective influences on belief. In: *Emotions and Beliefs,* Cambridge Univ. Press, Cambridge, UK.

[Cohen and Blum, 2002] Jonathan D. Cohen and Kenneth I. Blum. Reward and decision. *Neuron 36.*

[Cruse, 2002] H. Cruse. The evolution of cognition: a hypothesis. *Cognitive Science 27.*

[Damasio, 1994] A. R. Damasio. *Descartes' error: Emotion, reason, and the human brain.* G.P. Putnam, New York.

[Dayan and Balleine, 2002] P. Dayan and B. W. Balleine. Reward, motivation, and reinforcement learning. *Neuron 36*.

[Davidson, 2000] R. J. Davidson. Cognitive neuroscience needs affective neuroscience (and Vice Versa). *Brain and Cognition 42*.

[Doya, 2000] K. Doya. Metalearning, neuromodulation, and emotion. In: *Affective Minds*. Elsevier Science B.V.

[Doya, 2002] K. Doya. Metalearning and neuromodulation. *Neural Networks 15.*

[Fiedler and Bless, 2000] K. Fiedler and H. Bless. The formation of beliefs at the interface of affective and cognitive processes. In: *Emotions and Beliefs.* Cambridge Univ. Press, Cambridge, UK.

[Forgas, 2000] J. P. Forgas. Feeling is believing? The role of processing strategies in mediating affective influences in beliefs. In: *Emotions and Beliefs.* Cambridge University Press, Cambridge, UK.

[Frijda and Mesquita, 2000] N. H. Frijda and B. Mesquita. Beliefs through emotions. In: *Emotions and Beliefs.* Cambridge Univ. Press, Cambridge, UK.

[Frijda *et al.*, 2000] N. H. Frijda, A. S. R. Manstead and S. Bem. The influence of emotions on beliefs. In: *Emotions and Beliefs.* Cambridge Univ. Press, Cambridge, UK.

[Gadanho, 2003] S. C. Gadanho. Learning behavior-selection by emotions and cognition in a multi-goal robot task. *Journal of Machine Learning Research 4*.

[Griffith, 1999] P. E. Griffith. Modularity & the psychoevolutionary theory of emotion. *Mind and Cognition: An Anthology*

[Hesslow, 2002] G. Hesslow. Conscious thought as simulation of behaviour and perception. *TICS 6*.

[Mehrabian, 1996] A. Mehrabian. Framework for a comprehensive description and measurement of emotional states. *Gen., Soc. and General Psych. Monographs 121*.

[Montague *et al.*, 2004] P. R. Montague, S. E Hyman and J. D. Cohen. Computational roles for dopamine in behavioural control. *Nature 431*.

[Rolls, 2000] E. T. Rolls. Precis of The brain and emotion. *Behavioral and Brain Sciences 23*.

[Russell, 2003] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological Rev. 110.*

[Scherer, 2001] K. R. Scherer. Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, Methods, Research*. Oxford Univ. Press, New York.

[Smit *et al.*, 2003] A. Smith, S. Becker and S. Kapur. From dopamine to psychosis: a computational approach. In *Proc. KES'03*. Springer (LNAI 2773).

[Singh *et al.*, 2004] S. Singh, A. G. Barto and N. Chentanez. Intrinsically motivated reinforcement learning. *Proc. NIPS'04*. MIT Press, Cambridge, Massachusetts.

[Sutton and Barto, 1996] R. S. Sutton and A. G. Barto. *Reinforcement Learning: an introduction*. MIT Press, Cambridge, Massachusetts.

[Velasquez, 1998]. J. D. Velasquez. A computational framework for emotion-based control. In: *SAB'98 Workshop on Grounding Emotions in Adaptive Systems*.