Mapping Elements of Reinforcement Learning to Human Emotions

Elmer Jacobs

Master of Science Thesis

TU Delft Delft University of Technology

Interactive Intelligence

Mapping Elements of Reinforcement Learning to Human Emotions

Master of Science Thesis

For the degree of Master of Science in Computer Science at Delft University of Technology

Elmer Jacobs

October 4, 2013

Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS) · Delft University of Technology

Delft University of Technology
Department of
Interactive Intelligence (II)

The undersigned hereby certify that they have read and recommend to the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS) for acceptance a thesis entitled

Mapping Elements of Reinforcement Learning to Human Emotions

by

Elmer Jacobs

in partial fulfillment of the requirements for the degree of

Master of Science Computer Science

Dated: October 4, 2013

Supervisor(s):

dr. J. Broekens

Reader(s):

prof.dr. C.M. Jonker

prof.dr. J-J.Ch. Meyer

Abstract

Considerable overlap exists between emotion and Reinforcement Learning (RL). Emotion influences action selection while RL selects actions based on their anticipated result. Emotions also provide feedback on a situation, reflecting if the situation is desirable or not. The same type of feedback is given in RL based on the results of a state change. Finally, emotion drives adaptations in behaviour while RL continuously updates its policy based on newly gained experience. Because of this overlap, we theorized that a mapping exists from elements of RL to emotions, such that the occurrence and development of these RL emotions matches that in humans. Theory on emotions shows that complex emotions develop later and habituation can be observed in joy and fear. Further supported by theory on mapping situations to emotions, we mapped joy, hope/fear and confirmation emotions. We showed mathematically and in simulations that the development and occurrence of RL emotions with the mapping we created matches expectations based on emotion theory.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I dedicate this thesis to my parents, who have been supportive throughout all my years in university and without whom I would have never been able to start on a second master's degree at all. The knowledge that they are always available and willing to help in any way has provided me with the tranquility to finish this thesis.

Particular gratitude goes out to dr. Joost Broekens for coordinating my work, but especially for providing inspiration, motivation, enthusiasm, long discussions about and lots of new insights into the grand world of emotions, as well as a continuous drive to keep (improving my) writing.

I would also like to thank Prof. dr. Catholijn Jonker for taking over for Joost while he was gone, finding time for evaluating my thesis, and helping me find a way through the bureaucracy that seems inherent to any big organization.

I also thank dr. Judith Redi for her willingness to be in my committee and take a look at my work. Through an unfortunate combination of circumstances, she can not be present at my defense, which I sincerely regret.

I thank prof. dr. John-Jules Meyer for his presence in my committee and being prepared to evaluate the work of a student completely unknown to him.

I thank Anita Hoogmoed for her ability to find the holes in the schedules of the people I needed to meet and arranging countless other important things.

I would also like to thank my fellow students, especially Alex, Mark, Arvind, Allard and Rien for making life at the 12th floor that much more interesting.

Finally, a special mention goes out to the people at D.S.V.V. Punch and especially my own teammates, for all the good times we have experienced and will experience in the future.

“One day they’ll have secrets... one day they’ll have dreams.”
— *Dr. Alfred Lanning*

Chapter 1

Introduction

Emotions play an important role in human behaviour. They drive adaptations in behaviour and are therefore often coupled to learning [1]. Emotions emerge from situations in this world and similar situations usually result in similar types of emotions. An emotion can stimulate humans to take actions which preferably improve the situations. Furthermore, expected emotions can be used to strife for a specific situation through behaviour. With growing experience, emotions may change [2], meaning behaviour can be adapted on the basis of these emotions.

Reinforcement Learning (RL) [3] is inspired by psychology on human behaviour and relies on a mechanism similar to operant conditioning. In RL, a state represents the situation at a specific time instance, and a value is attributed to each state. The value is based on neighbouring states and rewards, the last which may be received upon transition to another state. A behavioural policy can be defined through the use of these values and by continuously updating the values on the basis of experience an optimal policy may be reached.

We see that RL and emotions share similar traits, a notion that has inspired earlier researchers to try and combine the two [4, 5, 6]. These similarities lead us to believe that it is possible to map human emotions using elements of Reinforcement Learning. A correct mapping would result in these Reinforcement Learning emotions to develop in the same way as human emotions.

Since human emotions result from situations, emotions in RL should be have the same way, also originating from situations or states. Mappings from situations to emotions can be found in the field of appraisal theory. We combine appraisal theory and theory about the development and occurence of emotion (among which is habituation [7, 8, 9]) to create a mapping from RL to emotions.

Based on the mathematics behind this mapping, we expect the resulting emotions to behave similar to humans, which we also demonstrate in a simulation where an RL-driven agent collects rewards in a maze. Apart from habituation and development we also test other hypotheses grounded in human theory. These hypotheses concern the addition of a penalty for making wrong moves and changes in the predictability of action results.

This thesis is organized as follows. In Chapter 2 we discuss theory about the functions and development of human emotions. Chapter 3 introduces theory on Reinforcement Learning and goes deeper into the overlap between emotions and elements of RL. Chapter 4 presents the basics of appraisal theory which maps situations to emotions. Appraisal theory is used in Chapter 5 to create a mapping from situations in Reinforcement Learning to emotions. In Chapter 6, we give several hypotheses based on emotion theory and clarify the setup of experiments. We also show the results of the experiments. Chapter 7 presents a discussion and conclusion on the results and gives recommendations for future work.

Chapter 2

The functions and development of human emotions

The emotions that we experience influence our behaviour and the nature of this influence needs to be understood in order to draw a parallel between emotions and Reinforcement Learning. In order to evaluate the accuracy of the mapping we wish to create, we also need a background on the development of emotions and possible mechanics that influence this development. In this chapter, we investigate both of these items.

2-1 Theories on the function of emotions

Emotions are usually regarded as a drive for action selection and we discuss two theories on how this works. The first theory about the influence of emotions on action selection is that they influence action readiness. Action readiness is the readiness of an individual to engage in interaction with the environment [10]. Emotionally relevant events increase the readiness for specific actions, with different events resulting in different actions being readied. Apart from different situations, action readiness may also vary among individuals. For example, meeting an animal in the wild that is perceived as potentially dangerous may induce a feeling of readiness for either fighting or fleeing. Action readiness reflects a desire for a specific change in a situation, even if no actions are available to reach that desire. In general, action readiness reflects the behavioural and psychological response process resulting from the evaluation of an event, which provides a preference for undertaking specific actions. The emotion elicited by the current situation therefore directly influences the action selection. In Figure 2-1, the process is summarized in a flow diagram, where we can see that the arrival in a new situation gives rise to emotions that affect actions.

Another theory about the nature of the connection between emotions and actions assumes that emotions function as a feedback signal [1]. According to this theory, emotions are coupled to previously experienced situations, which resulted from certain actions. The action selected is the one that has the most positive emotion attributed to its probable result. This

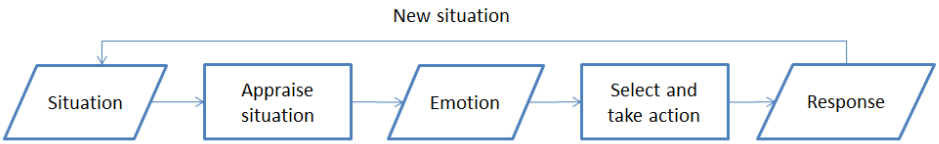

Figure 2-1: Flow diagram of Frijda's theory of action readiness.

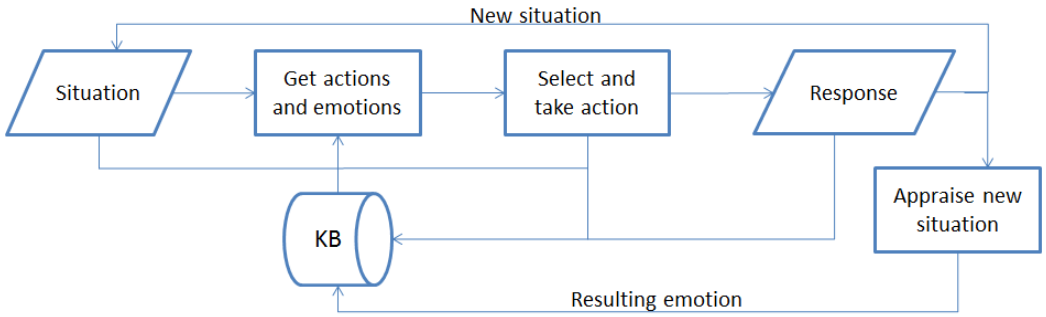

Figure 2-2: Flow diagram of Baumeister's theory of emotion as feedback. The knowledge base stores experience in the form of the previous state and action, the resulting state and resulting emotion. This experience is used later to select an action.

requires some previous experience, which may be gained through trial-and-error or learned from external resources, such as books, movies or teachers. Results that are distinctly different from what was expected influence the knowledge base of experiences, effectively forming a feedback loop to the action selection. In other words, the emotion elicited by possible future situations influences the action selection. A flow diagram capturing this process is given in Figure 2-2.

These theories show two ways in which emotions can influence action selection. They both come down to evaluating the preferred action based on a desired state change, either based on the current state or possible future states.

2-2 Development of human emotions and habituation

Knowledge about the function of emotions has provided us with a foundation for the mapping, but we also need to learn about the development of emotions. As behaviour changes and becomes more complex, the same goes for emotions. In the first months of infancy, children exhibit a narrow range of emotions, consisting of distress and pleasure. Distress is typically expressed through crying and irritability, while pleasure is marked by satiation, attention and responsivity to the environment [2]. Joy and sadness emerge by 3 months, while infants of that age also demonstrate a primitive form of disgust. This is followed by anger which is most often reported between 4 and 6 months. Anger is thought to be a response designed to overcome an obstacle, meaning that the organism exhibiting anger must have some knowledge about activity toward a goal. This reflects the child's early knowledge of its own abilities,

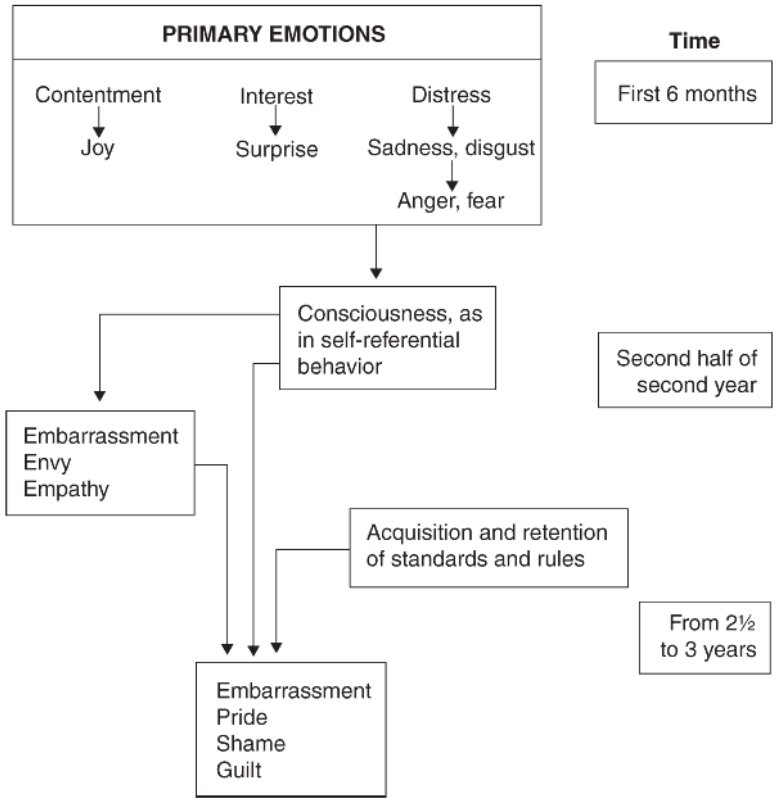

Figure 2-3: Development of emotions over the first 3 years of life. Adapted from [13], Figure 18.1.

since it knows when a goal is being thwarted. Anger is followed by fearfulness, usually first reported at 7 or 8 months. Since fearfulness requires comparing the frightening event to some other event, to determine the unfamiliarity or discrepancy of the situation [11]. Next to these emotions, surprise can also be noted within the first 6 months of life. The six emotions anger, disgust, fear, happiness, sadness and surprise were found by Ekman [12] to be universal across humans and are in general seen as the basics upon which other emotions are built.

More complex emotions start arising as the self-consciousness of the child develops. The ability to put events into perspective with reference to oneself forms the basis for motions such as envy, empathy and embarrassment. Later, the acquistion and retention of standards and rules allows for more complex forms of embarrassment, as well as pride, shame and guilt [13]. This set of emotions is complete after about 3 years and forms the basis for the development of any further emotions. It is summarized in Figure 2-3.

2-2-1 Habituation

Apart from the development of emotions, habituation is also an important mechanic to study emotion. Habituation is the decrease in intensity of the response to a stimulus resulting from that stimulus being repeatedly received. It has been shown to exist for both the joy and fear emotions [7, 8, 9, 14]. Next to affecting emotions, habituation is also present in learning [15],

since the learning signal originating from a certain event also decreases in strength if that event happens more often.

2-3 Summary

This chapter has provided us with several insights into the theory behind human learning. First, we have examined two ways in which emotions influence human action selection. Next, we have created a basis for comparing later simulations using our set of represented emotions to the real world. A specific order of developing emotions as well as the mechanism behind habituation was discussed to this end. In the next chapter, we take a look at Reinforcement Learning (RL), examining basic elements and solution methods. We also compare these elements to human behaviour and take a look at previous work touching upon the relation between emotions and RL.

Chapter 3

Reinforcement Learning and human learning

Reinforcement Learning (RL) is a type of machine learning aimed at maximizing some cumulative reward by choosing appropriate actions. We require a through understanding of this type of learning before using its elements to represent emotions. Both the basic concepts of RL and a number of solution methods are discussed. These solution methods revolve around certain characteristics, which we compare to similar characteristics in human learning. Previous work on using emotions to drive RL as well as research on the link between human learning and Reinforcement Learning is discussed afterward.

3-1 Basic elements of Reinforcement Learning

Reinforcement Learning takes place in an environment that has a state $s \in \mathcal{S}$, where $\mathcal{S}$ is the set of possible states. An agent present in that environment selects an action $a \in \mathcal{A}(s_t)$ to perform based on that state, where $\mathcal{A}(s_t)$ is the set of possible actions when in state s_t at time t. Based on this action, the agent receives a reward $r \in \mathcal{R}$ once it reaches the next state, with $\mathcal{R}$ the set of rewards.

The action the agent executes is based on its policy π, with $\pi_t(s, a)$ the probability that $(a_t = a)$ if $(s_t = s)$. In RL, this policy gets updated as a result of the experience of the agent such that the total reward received by the agent is maximized over the long run.

Within the environment, rewards are predefined. This allows us to predefine the goals of the agent by setting the most preferable states. Note that the actions are specifically not predefined; these should arise naturally from the policy. The total expected return R at time t can be defined by taking the sum of rewards , with r_T the reward at the final time step:

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + \cdots + r_T \tag{3-1}$$

This sum is finite in applications with a natural notion of a final time step, but can become infinite in applications where such an endstate does not exist. To deal with such situations, a discount factor γ was introduced, where $0 \leq \gamma \leq 1$, discounting rewards that are further in the future to ascertain a finite sum if all rewards are finite:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=o}^{\infty} \gamma^k r_{t+k+1} \tag{3-2}$$

Thus, with higher values of γ, future rewards become more important, while lower values lead to a bias toward nearer rewards. In this way, the parameter partially determines the balance between exploration and exploitation, with the higher values favouring exploration.

In order to simplify the prediction of the next state, given the current state and some action, we need to assume that the RL task satisfies the Markov property. The Markov property states that the probability distribution of the future state depends only on the previous state and action; in that way, a specific action executed in a specific state always has the same probability distribution, regardless of any previous states. This simplification allows a more clear formulation of the optimal solution. RL tasks satisfying the Markov Property are commonly referred to as a Markov Decision Process (MDP).

If the state and action spaces of an MDP are finite, it is called a finite MDP. These are specifically important to RL, as most of modern RL theory is based on these finite MDPs. For a finite MDP, we have two important quantities. First of all, we have the transition probabilities $P^a_{ss'}$:

$$P^a_{ss'} = Pr\{s_{t+1} = s' | s_t = s, a_t = a\} \tag{3-3}$$

Denoting the probability of each next possible state s' given current state s and action a. Strongly related are the expected values of the next rewards $R^a_{ss'}$:

$$R^a_{ss'} = E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\} \tag{3-4}$$

In general, we can use a state transition graph to represent an MDP, in which we denote both these quantities. An example transition graph is given in Figure 3-1.

3-1-1 Policies

We introduced the concept of policies earlier and examine them in more detail here. When following a policy π, we need some way to evaluate the worth of each state. For this, we use the following notation to denote a value $V^\pi(s)$ under a certain policy:

$$V^\pi(s) = E_\pi\{R_t | s_t = s\} = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s \right\} \tag{3-5}$$

That is, a value is the discounted sum of the expected rewards from all states following the current one. Alternatively, we can evaluate actions instead of states, which leads to action values $Q(s,a)^\pi$:

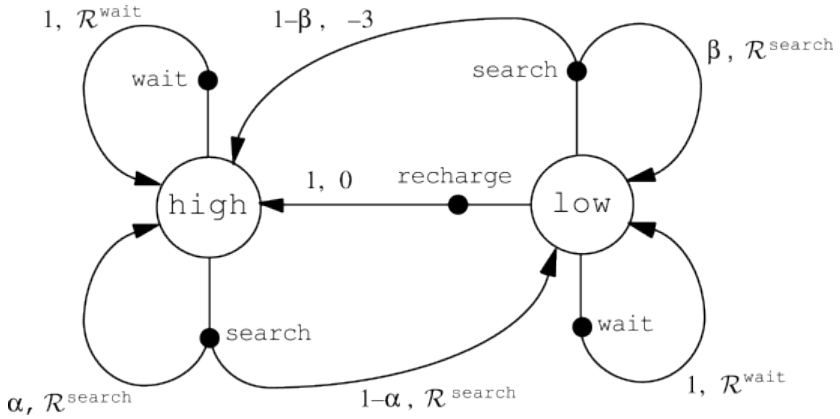

Figure 3-1: A state transition graph. Small circles are actions, large circles are states. Rewards and transition probabilities are denoted per transition. Adapted from [3], Figure 3.3

$$Q^{\pi}(s,a) = E_{\pi}\{R_t | s_t = s, a_t = a\} = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\right\} \tag{3-6}$$

Where Q^{π} is the value for a state-action combination (s,a) when following policy π. Both of these values may be estimated from experience. Methods of estimating the values for each state using averages over a long number of runs are called Monte Carlo methods. These methods are inefficient for large numbers of states. Therefore, some equation that provides a direct solution is preferred. We can do this for values by rewriting V^{π} to:

$$V^{\pi}(s) = \sum_{a} \pi(s,a) \sum_{s'} P^{a}_{ss'} \left[R^{a}_{ss'} + \gamma V^{\pi}(s')\right] \tag{3-7}$$

Which is known as the Bellman equation for V^{π}. The value function is the unique solution to its Bellman equation, allowing us to compute, approximate and learn this function. If we want the optimal policy in terms of total return, we need to maximize that return over all policies:

$$V^{*}(s) = \max_{\pi} V^{\pi}(s,a) \tag{3-8}$$

Which can also be written as:

$$V^{*}(s) = \max_{a \in A(s)} \sum_{s'} P^{a}_{ss'} \left[R^{a}_{ss'} + \gamma V^{*}(s')\right] \tag{3-9}$$

Or, in case of action-values:

$$Q^{*}(s,a) = \sum_{s'} P^{a}_{ss'} \left[R^{a}_{ss'} + \gamma \max_{a'} Q^{*}(s',a')\right] \tag{3-10}$$

Once we have these optimal values, it is easy to find the optimal policy as the agent simply can choose the action that results in the greatest value, or the highest-valued action in case

of action values. Of course, actually finding the complete set of optimal values from scratch may require vast amounts of computational power as all possible state action sequences have to be considered. Furthermore, during the learning process, we can not always choose the maximum value if the complete range of values is still unknown. Choosing the maximum in this case could result in the agent getting stuck in local optima. The action-selection method which always chooses the maximum value is called Greedy, but if we want to avoid any local optima, an alternative is ϵ-Greedy, where ϵ denotes the probability of picking a random action instead of the maximum-valued option. Another alternative may be found in the set of methods called softmax, which essentially vary the probability of choosing an action based on the expected value. One of the most common softmax methods uses a Gibbs or Boltzmann distribution, where the probability of picking action a at time t is given by:

$$\frac{e^{\beta Q_t(a)}}{\sum_{b=1}^{n} e^{\beta Q_t(b)}} \tag{3-11}$$

where β is a positive parameter called inverse temperature. If $\beta \longrightarrow \infty$ in the limit, we have the Greedy selection, while a value of 0 constitutes random action selection.

To deal with settings with large numbers of states and actions, we need to use a method to find the optimum without explicitly considering all state-action combinations. Different solution methods are discussed in the following section.

3-1-2 Solution methods

There are three main solution methods for the basic RL problem; Dynamic Programming (DP), Monte Carlo (MC) methods and Temporal Difference Learning (TD). We discuss them here shortly. In all cases we assume that the environment is a finite MDP.

Dynamic programming

If we want to determine the state-value function for an arbitrary policy, we can simply use the Bellman equation as described before, which leads to a set of linear equations the size of the state space if the dynamics of the environment are completely known. Solving these becomes quite tedious in larger spaces, so an incremental method is preferred. We can use the Bellman equation as an update rule instead, which after arbitrarily initialising the values leads to:

$$V_{k+1}(s) = \sum_a \pi(s,a) \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V_k(s') \right] \tag{3-12}$$

This function is guaranteed to converge to V^π if it exists, that is, if either $\gamma < 1$ or eventual termination is guaranteed from all states under policy π. This allows us to evaluate all values given a certain policy. Of course, we also need to evaluate different policies, comparing them to each other in order to end up with the optimal policy. We can do this by applying policy iteration, which works by attempting to determine a new policy after evaluation of the old

one. That is, the iteration determines V^π first, after which a new policy π' is determined by maximizing the following expression:

$$\sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V^\pi(s') \right] \tag{3-13}$$

over all possible actions. If the new policy is different from the old one it can be proven to always be an improvement, while equality between policies means that the optimal policy has been found. This method requires iterations within iterations (each policy improvement requires an iterative policy evaluation), which can take quite a long time to converge. Combining policy improvement and evaluation brings us to value iteration, which can be denoted by the following equation:

$$V_{k+1}(s) = \max_a \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V_k(s') \right] \tag{3-14}$$

After convergence, the policy simplifies to:

$$\pi(s) = \arg\max_a \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V(s') \right] \tag{3-15}$$

In both formerly discussed approaches, DP requires a complete sweep of the state-space, which can quickly become very costly if those spaces are large. An alternative to this is asynchronous DP, which may use old information as basis for the calculations, rather than updating every state per iteration. The advantage of DP is that the value estimates are updated based on other value estimates, otherwise known as bootstrapping, which allows for fast convergence towards the optimum. Of course, a model of transitions and expected rewards is in general not available, in which case it has to be learned by experience. We call this model-based RL.

Monte Carlo methods

Contrary to DP, we do not assume complete knowledge of the environment for MC methods. Instead, these methods learn by experience by averaging the return, which is the expected cumulative discounted reward. In order to ensure the availability of well-defined returns, MC methods are defined for episodic tasks only. Despite this, the methods are quite similar to DP.

Two MC methods are the every-visit and first-visit ones. The every-visit averages the returns that follow all visits to s in a set of episodes, while the first-visit does the same but only for the first visit to s each episode. Obviously, when a state can only occur once during an episode, the two methods are exactly the same.

As before, using any of these methods, we can define a form of policy iteration once again. Starting out with some arbitrary policy, the corresponding action values are evaluated after which the old policy can be improved based on these values. Once no improvement is noted anymore, we have reached the optimal policy. Since we use action values, improving the policy can simply be done by choosing

$$\pi(s) = \arg\max_a (Q(s,a)) \tag{3-16}$$

The problem is that convergence can only be assured under two assumptions; each episode must have exploring starts (that is, each state-action combination is visited infinitely often in the limit) and there must be an infinite number of episodes. This last assumption may be dealt with by not doing full policy evaluation before returning to the improvement. The first assumption can be removed if the policies are not completely greedy anymore, rather choosing the less greedy actions with some non-zero probability as well, for instance using an ϵ-Greedy policy as discussed before. That way, all states will be visited at some time. The disadvantage of this is that the convergence will be toward a policy that is best among the ϵ-Greedy policies only.

There are also some off-policy MC methods, which separate the behavioural policy and the policy to be improved (estimation policy). This allows one to choose a behavioural policy that has specific properties required for convergence, while the estimation policy is improved based on the results that flow forth from the use of the behavioural one.

In general, the convergence properties of MC methods are not clear. The most important difference with DP is that MC methods can learn without a model, as they use sample experience. On the other hand, MC methods do not bootstrap as value estimates are not based on other value estimates.

Temporal Difference learning

Bootstrapping and learning from experience are effectively combined in the type of algorithms applied in TD. While MC methods estimate the return and DP methods estimate the value of the next state, TD actually estimates both. By adding an eligibility trace factor $0 \leq \lambda \leq 1$ (see Section 3-1-2), it is often possible to increase the learning rate [3]. Using the notation TD(λ) for denoting the value of λ in the basic TD method, the simplest TD-method, TD(0), updates according to:

$$V(s_t) \leftarrow V(s_t) + \alpha \left[r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \right] \tag{3-17}$$

with α a constant step-size parameter. Where MC methods must wait for the final return, TD immediately updates after a number of steps, in this case one. We discuss the mathematics behind eligibility traces later.

As with MC, TD also has the possibility of either on-policy or off-policy learning. One form of on-policy learning is called Sarsa, which updates action-values according to:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right] \tag{3-18}$$

where a_{t+1} is determined by the policy. The problem with Sarsa (and with on-policy methods in general), is that convergence properties are largely unknown, although, among others, Singh made some effort towards this which showed convergence in specific cases [16], but no general results were produced yet. Convergence in the off-policy case on the other hand was proven. This method is better known as Q-learning, for which the update formula is:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right] \tag{3-19}$$

Contrary to Sarsa, Q-learning does not take the action selection of the policy into account. This means that although Q-learning finds the values of the optimal policy, the results following that policy may be worse than the one found with Sarsa if the policy is stochastic. That is, the results of Q-learning are unaffected by the possibility of taking a detrimental choice as the result of chance.

A third family of TD-methods consists of the actor-critic method. This is an on-policy method that separates the policy from the value function. The policy is regarded as the actor, while the estimated value function is the critic. The critic evaluates the performance of the actor by determining a TD error δ:

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{3-20}$$

The Temporal Difference Error is used to increase the probability for picking an action, thus adapting the policy. The values are still updated as discussed earlier, but are not explicitly used to select the action anymore, only to calculate the error.

Eligibility traces

As all three described methods have advantages, effort has been put into unifying these. This has resulted in several different other solution methods. The most important of these techniques is the use of the eligibility trace λ , which we shortly mentioned earlier. The eligibility trace can be seen as a bridge between MC and TD-methods; where MC updates only after the complete return, one-step TD updates based on the next return; the n-step TD version updates based on the next n steps. By giving positive weights to all steps and making sure the total sum of the weights equals 1, the update still remains valid. The variable λ is actually such a weight factor that changes proportionally based on the step number it is applied to. By using a normalization factor, all separate weights always amount to 1. With this knowledge, the formula for the return becomes:

$$R_t^\lambda = (1 - \lambda) \sum_{n=1}^{\infty} \lambda^{n-1} R_t^{(n)} \tag{3-21}$$

So for $\lambda = 1$, we have $R_t^\lambda = R_t$, which is the MC-return and for $\lambda = 0$ we get the one-step TD return discussed earlier. In a similar way, Sarsa, Q-learning and actor-critic methods may also be adapted.

Eligibility traces allow efficient tuning between MC and one-step TD-methods. As both methods have good efficiency in different types of situations, implementing traces allows for one solution method to cope with lots of different worlds.

3-2 Overlap between Reinforcement Learning and emotions

The discussed solution methods have given us insight into the influence of RL on the behaviour of an agent. This allows us to study overlap between the influence of emotions and behaviour and the influence of RL. First of all, emotions influence action selection, refelecting a preference or drive to to a specific action in a given situation. Frijda [10] explained this using his theory of action readiness, where emotions increase the readiness for specific actions. Action readiness reflects the behavioural and psychological response process resulting from the evaluation of an event. In RL, actions are selected based on their anticipated results, which are valued according to previous experience. A second function of emotion is providing feedback on the situation that results from an action. Baumeister [1] explained that the emotion attributed to a situation can be used to assess the benefits of the action that lead to that situation. RL uses the results from a state change to update the evaluation of the previous state, changing the preference to be in that state. These functions show emotion as a drive for adaptive behaviour, since it can alter both the selection of an action and the evaluation of an anticipated situation. The core mechanic of agents using Reinforcement Learning is that they adapt their behaviour based on knowledge gained about the environment, which based on theory also appears to be the core mechanic of emotions.

3-2-1 Emotion-driven Reinforcement Learning policies

Using emotions to drive intelligent agents has been suggested earlier. A new computational model of emotions, FLAME, was suggested in [17]. Using a fuzzy logic approach, mathematical formulas are provided for calculating the intensity of a number of emotions. By allowing human users to provide feedback, the value of action can be determined via a learning algorithm, which allows the model to calculate emotions attributed to those actions. Via surveys it is shown that the behaviour of an animal using the FLAME model is considered more natural than one without. Marinier attempts to draw a comparison between RL with and without emotion [18]. However, the concept of emotions and mood are comprised of extra state-information that is added to the given simulation in a maze-setting. It does result in faster learning, but one could argue that that is simply the result of added state information, which could be used without calling it an emotion. Sequeira [6] investigated the use of a richer reward signal based on emotions and showed that RL agents benefit from such an addition to the basic scheme.

3-2-2 Linking the human hormonal system to Reinforcement Learning elements

Apart from emotions that drive policies, some research has also been done on comparing the human hormonal system to RL. Events activate specific parts of the brain that influence learning in animals[19]. Researchers have attempted to find a link between specific hormones and computational models for learning, among which is RL. Several of these links have been studied and are summarized here.

Dopamine and the Temporal Difference error

The effects of dopamine are studied in relation to processing reward information, specifically in the form of prediction and detection. Unpredicted positive rewards result in activation of dopamine neurons, while disappointing rewards suppress neurons. A reward that is repeatedly received yields a decreasingly strong response, while rewards that do not meet expectations do activate the respective neurons resulting in adaptations in the prediction system. Dopamine responses are shown to have "characteristics of a teaching signal postulated by reinforcement learning theories" [20]. The effect of dopamine is therefore believed to be similar to that of the TD error δ described in the actor-critic model [21, 22].

Norepinephrine and inverse temperature

Norepinephrine (or noradrenaline) has been repeatedly shown to affect decision making. It controls arousal and relaxation and appears correlated to the accuracy of action selection. Making the best known decision is often favored, especially in urgent situations. However, more relaxed environments should allow exploration into unknown areas to search for unexpected rewards. This balance between exploration and exploitation is governed by the inverse temperature β in the Boltzmann action-selection. Therefore, the link was made between norepineprhine and the inverse temperature [22]. The benefits of balancing exploitation vs exploration using elements of RL to represent a form of emotion were investigated and demonstrated in [4] and [5].

Serotonin and discount factor

Serotonin is closely connected to the reinforcement of addiction and impulsive behaviour. Both are elements of the memory, where long-term consequences are ignored in favor of short-term gains. In other words, serotonin determines how far in the future the consequences of actions are considered. The discount factor γ or the eligibility trace λ can both decrease the effects of future rewards on the evaluation of the current state, following a mechanism similar to that of serotonin. Serotonin and the discount factor are hypothesized to be related, where a higher level of serotonin constitutes a higher discount factor [22].

Acetylcholine and learning rate

Acetylcholine is seen as an important modulator for memory, balancing between storage and update. It determines how much of an experience is stored as new information and how much is used as update signal to old information. In this manner, events may have a large or smaller learning effect dependent on the amount of acetylcholine. The learning rate α determines the magnitude of the update effect of new information, hence it is comparable to the release of acetylcholine [22].

3-2-3 Conclusion

Research on connecting emotions to the action-selection RL has not yet yielded any definitive results. The successful attempts at drawing a comparison between parameters of RL and the

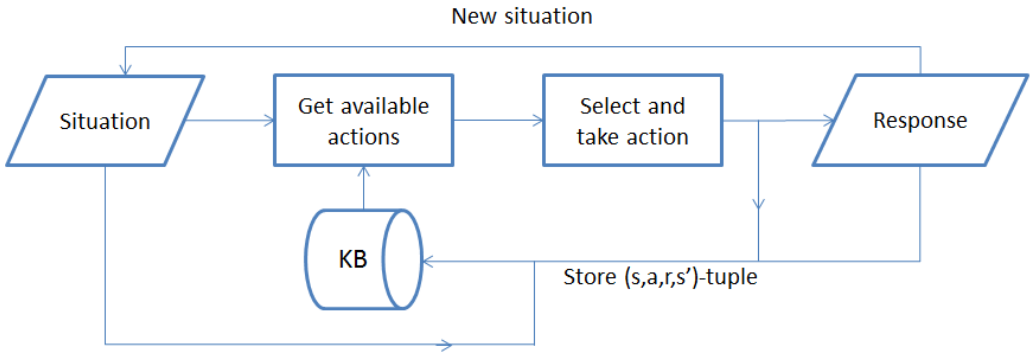

Figure 3-2: Flow diagram of basic Reinforcement Learning. The (s, a, r, s')-tuple contains the previous state, action taken, resulting reward and resulting state.

hormonal system active in human learning do show that both are comparable to each other. Discussion remains on the function of at least some of the discussed hormonal systems [23], but the existence of resemblances can not be denied. The findings therefore support RL as an accurate model for human learning, strengthening our conviction that emotions can be represented using elements of RL. We compare specific characteristics of RL to human behaviour in the following section, in order to determine what we need for a correct representation and what type of simulation is best suited to perform tests.

3-3 Elements of Reinforcement Learning vs elements of human behaviour

The basics of RL can be summarized in a flow diagram to represent how the action selection is influenced. This flow diagram is given in Figure 3-2. If we compare this diagram with the two diagrams based on the theories by Frijda and Baumeister, Figure 2-1 and Figure 2-2 respectively, we see that the flow of processes is similar to that of Baumeister. This suggests that the representations of emotions should be based on the RL elements involved with the (prediction of the) new situation, as depicted in Figure 3-3. Mapping situations to emotions is the field of appraisal theory. An appraisal model based on elements of RL is therefore a strict requirement if we want to correctly map emotions in RL. The situation is the most important element for determining emotions and we can not create a mapping without an appraisal model.

3-3-1 Characteristics of Reinforcement Learning and human learning

The solution methods discussed previously have demonstrated that there are many ways to solve an RL-problem. The choice for the most efficient solution method depends mainly on four characteristics, namely:

- Using state or action values
- The type of action selection

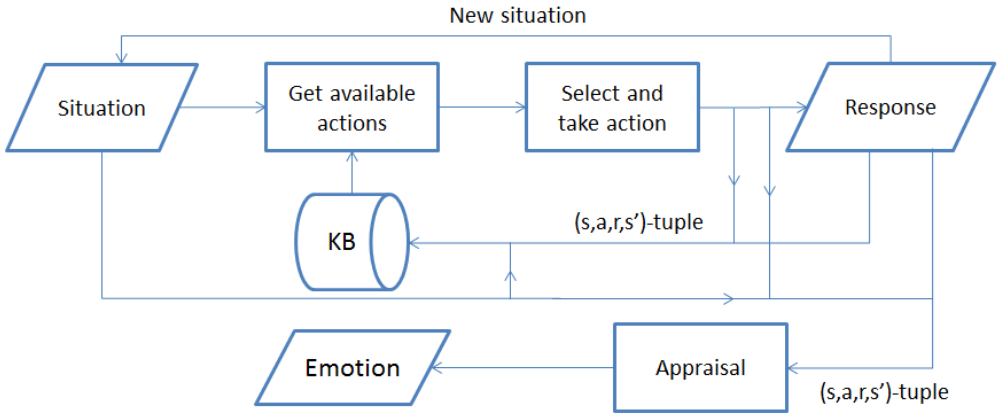

Figure 3-3: Flow diagram of Reinforcement Learning, with added appraisal process for representing emotions. The (s, a, r, s')-triple contains the previous state, action taken, resulting reward and resulting state.

- On- or off-policy learning
- Model-based or model-free learning

We can link these characteristics to elements of human behaviour by studying what their effect is on learning in RL and translating that to human learning. Since these are characteristics of the problem itself, this comparison provides a foundation for the type of simulation that is closest to the real world. All these characteristics influence the learning method, which can result in different emotional development. More precisely, the changes and updates to the state or action values are directly dependent on these four items, meaning convergence may be altered.

Choosing between state and action-values

Values and action values, as a prominent part in any RL system, are bound to play an important role in the emotion-mapping. Essentially, both types represent an expected return. In the case of state values, the state itself is the point of interest and choices are made to arrive in the most profitable state. For action-values, the agent looks at the list of available actions and chooses the most profitable action. As the action-values under an optimal policy may be expressed in the state values under the optimal policy, there is no mathematical difference in the policy that results.

Once we map to values to emotions, we need to be a bit more careful. If we use state values, this means a state itself results in an emotion. In the human world, this means the situation itself is evaluated and determines the emotion, such as being happy if the weather is nice. Using action values means the focus lies on the set of available actions which may cause the emotion. For humans this reflects an emotion induced by the scala of actions available, such as despair if everything you can think of doing results in some sort of penalty. Both possibilities exist in the human world so the use of both action-values and state values is valid in this regard.

Determining the type of action selection

Action selection determines which action is taken in any situation. It can range anywhere from completely random to Greedy. A human using random action selection constantly is not bound to live for a long time. It removes the point of learning, as nothing is done with the results of experience. If an agent takes Greedy decisions, this results in making the choice that is considered best at each time, which is an accurate depiction of general human behaviour. Similar to Greedy, ϵ-Greedy acts Greedy most of the time, but random at times with probability ϵ. In other words, this represents a greedy human who sometimes does something unexpected that does not directly increase his own well-being. The Softmax method, finally, is a method that determines the probability of choosing certain actions based on their expected value. That is, the action that is expected to benefit the agent the most has the highest probability of being chosen. Using the Boltzmann distribution, it is possible to effectively tune the action selection between a Greedy and random action selection by means of the parameter β, commonly referred to as the inverse temperature. Furthermore, behaviour similar to ϵ-Greedy may also be approximated by setting β correctly, so Boltzmann action-selection actually suffices for representing all types. In fact, research also suggests that this type of action-selection is similar to that of humans [24, 25, 26, 27].

Using on- or off-policy learning

Learning on- or off-policy has effects on the development of values and therefore may also influence emotions. For off-policy learning, the policy that is evaluated is different from the one that is being used to choose actions. As off-policy learning in general uses the maximum of all known next states for the sake of learning, which is comparable to a human assuming both complete control over his choices and correct knowledge of their effects. This is no strange assumption in simple environments, but there are also scenarios where such a predictability of the world is often not present, such as at the start of a simulation. When little is known about the world, exploration is important for finding rewards, meaning an agent should not always perform the best action. In that case, it is better off using on-policy learning since this takes into account chosen actions. We can conclude that both types of learning are valid for representing human learning, as there are circumstances where both are used.

Modelling or not modelling the environment

Learning a model of the environment parallel to learning its dynamics has clear advantages; most importantly, it allows accurate prediction of state transitions, which can be used to increase the speed of convergence toward correct values. Regardless of these advantages, many RL solution methods do not require a model. This holds for both MC and TD methods, which can converge to the right values without ever learning transition probabilities. For humans, the model seems to be used for a general impression of the environment, possibly simplifying the action selection; estimating the outcomes of actions in deciding what action to take. Furthermore, the model allows a comparison of expected and actual transition resulting from an action, which is important for determining surprise. We estimate that a model will be required for correctly mapping emotions.

3-3-2 Conclusion

Our comparisons have demonstrated that each variant has a counterpart in human learning. This means that any combination of the different choices for the aforementioned characteristics provides a scenario that is comparable to some real life example. We can therefore not make a definitive choice for any of these characteristics here. Either we have to test all possibilities in similar environments or decrease the number of options on other grounds later.

3-4 Summary

In this chapter, we have explained the basics of Reinforcement Learning. Several different solution methods were discussed in detail. The comparison between RL and human learning has demonstrated that situations need to be evaluated in order to determine emotions. Therefore, we need to determine an appraisal model based on elements of RL, since this is the only way to correctly map emotions. It should be our first priority before we do anything else.

Apart from studying the importance of situations, we also attempted to draw a comparison between characteristics of RL and the human world in order to determine the boundaries of the simulation. We found that most variants of these characteristics have some counterpart in the human world, so we were not able to make a definitive choice as to the setup for the simulation. Previous work demonstrates that emotions can improve the performance of RL-agents. Furthermore, the findings in linking the human hormonal system to different RL parameters have shown that RL and human learning are indeed comparable. This supports our theory that some way exists to represent emotions using elements from RL. Before starting on simulations, we discuss the basics of appraisal theory which are required in order to create an appraisal model that uses elements of RL, since we can not map emotions without it.

Chapter 4

Mapping situations to emotions: Appraisal Theory

Situations are an important factor in determining emotions and appraisal models provide a mapping from situations to emotions. In order to accurately map elements from Reinforcement Learning (RL) to emotions we need to have a good understanding of appraisal theory. There are several appraisal models avaialable, two of which we study in detail this chapter, demonstrating the different emotions that can be defined through each of these models.

4-1 The OCC model

In 1988, Ortony, Clore and Collins described in their book how they believed that appraisals are structured such that different combinations result in different emotion types [28]. They view emotions as "valenced reactions to events, agents, or objects, with their particular nature being determined by the way in which the eliciting situation is construed". By analysing different emotion types, they finally arrive at a rather simple structure for the different types, depicted in Figure 4-1. The emotion resulting from some appraisal may be found by simply moving through the tree, following the correct branches based on the appraisal type and noting what emotions may result throughout that branch. For instance, when we want to find the emotion resulting from a reaction to the consequence of an event, while the prospects of the event are irrelevant, we arrive at the emotion types referred to as well-being. This means the attributed emotions are either joy or distress, which in turn fall in the category of pleased or displeased. The intensity of an emotion is determined by local variables, which are different for each of the emotion types. These variables later form the basis of the mapping we create.

Next to the local variables, the OCC model defines four global variables affecting intensity; the sense of reality, proximity, unexpectedness and arousal. Sense of reality describes how real an event feels, while proximity is a measure for how close the event has taken place, whether that be in time or space or anything else. Unexpectedness is the degree to which an event,

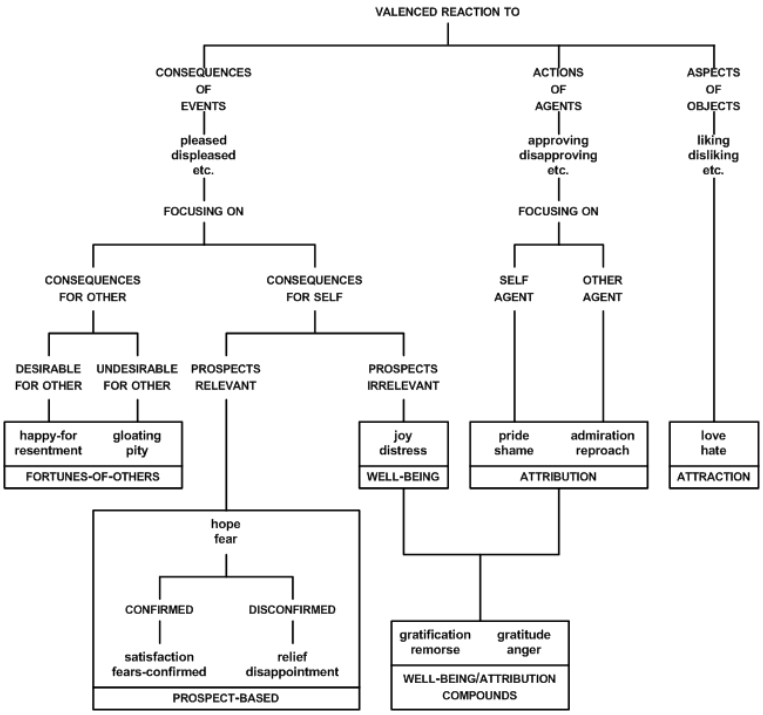

Figure 4-1: The OCC model, adapted from [28], Figure 2.1

after it has happened, was unexpected and usually positively correlated with the intensity of an emotion. It differs from the local likelihood variable, in the sense that likelihood anticipates events and affects intensity based on the anticipation. The final global variable, arousal, is a physiological status that may also intensify emotions.

4-1-1 Emotion types and intensity variables

Fortunes-of-others emotions

Fortunes-of-others emotions are being happy and sorry for someone else, as well as resentment and gloating. They are influenced by four internal variables. The first two are the degree to which an event for the other is desirable for oneself, and the degree to which that event is desirable for the other. The other two are the degree to which the person deserves the event and the degree to which the person is liked. These emotions specifically focus on someone else, judging whether they get what they deserve or had coming.

Well-being emotions

The well-being emotions are joy and distress, representing the individual being (dis)pleased about an event. The intensity of this emotion depends on only one internal variable; the desirability of an event. Well-being emotions result from a focus of the person on only the desirability, as a focus on other aspects results in other types of emotions, as we come across later. As a global variable, unexpectedness has a particularly large effect on well-being as the

focus on the desirability only results in a strong increase of intensity if the event was highly unexpected.

Attribution emotions

Attribution emotions are pride, shame, admiration and reproach and are reactions to the actions of oneself or another agent. They depend on the degree of judge praiseworthiness, the strength of the cognitive unit and the deviations of the actions of the agent with respect to expected behaviour. The strength of the cognitive unit represents the strength of the connection of the agent doing the action to oneself. Obviously, this is much higher for people close to you then for those you do not know. Deviations of expected actions are a form of unexpectedness, but are explicitly named here since they are an important part of attribution emotions.

Attraction emotions

Attraction emotions are like and dislike. These emotions are attributed to objects rather than agents. The intensity dependson the appeal of the object and the familiarity with it. The disposition towards some familiar object plays a large part in the actual emotion attributed to that object, increasing intensity if the object meets the standards of the individual.

Prospect-based emotions

Prospect-based emotions are hope and fear, which are reactions to the prospect of an event. In this group we also find the confirmation emotions, which are based on the (dis)confirmation of a prospect; satisfaction, fears-confirmed, relief and disappointment. The intensity of hope and fear are dependent on both the desirability of the event that is foreseen and the likelihood that the event will actually take place. Confirmation emotions are a retrospect evaluation of the actual outcome compared to the expected outcome. Their intensity depends on the attributed prospect emotion, the effort expended in attaining or preventing the event, and the degree of realization of the event. It should be noted that unexpected events may result in well-being emotions insteaf of dis(confirmation), as the feeling of joy is much stronger than satisfaction if the result was far better than expected.

Well-being / Attribution compounds

Compound between well-being and attribution are gratitude, anger, gratification and remorse type emotions. They depend on both the well-being and the attribution variables for intensity. These types of emotions are typically attribution emotions, with an added effect of the well-being attributed to the related event. The attribution toward an action of someone is combined with the well-being that results from that action. For example, a pride attribution for an action combined with a joyous feeling for the result of that action forms a feeling of gratification or self-satisfaction.

4-1-2 Uses of the OCC model

Because of its clear structure, the OCC setup allows a rather simple translation to programming and is thus particularly well-suited for mapping variables to emotions. It has been applied in several different applications. For instance, in the research by André [29], the model is used in two applications for determining the affective state of the character . One of these applications is used for teaching children how emotional states change behaviour and how behaviour can induce emotions, while the other concerns improving the believability of a virtual character. A similar application is described by Bartneck [30], who also tests the OCC model by letting a character express emotions based on that model. He does note the need for the quantification of some of the OCC variables (specifically likelihood, realization and effort of events) through a history function. Furthermore, he suggests the use of an interaction function to mix the affective values of the environment with the emotional state of the character to create a more accurate emotion, as well as some personality designer for setting standards and attitudes for the character. Finally, he mentions the possibility of simplifying the OCC model based on the abilities (for expression) of the character. Kshirsagar [31] also uses the emotion categories given in the OCC model combined with Ekman's 6 basic facial expressions [32] as a guideline for creating facial expression for several different emotions. In this way, a system is developed that allows users to design personalities for virtual humans by simply manipulating a number of parameters, such as the mood transition matrix which determines how likely it is that a transition from one mood to another takes place. The above examples demonstrate how the OCC model may be used and show its effectiveness and is some cases shortcomings. A further review of the model, from a programmer's perspective, is conducted by Steunebrink et al. [33]. They identify a number of ambiguities that arise when the diagram is strictly interpreted as a class-diagram with inheritances. To solve these ambiguities, they devise a more explicit model which follows basic inheritance conventions from software engineering.

4-1-3 Summary

We have discussed the OCC model in detail, showing that it provides a simple way of deriving emerging emotions from appraisals. At first glance, it appears to be a good candidate for a computational approach to emotion. This also becomes clear from the type of applications in which the model is used. Although there has been some critique, the model is easy to use, providing a straightforward appraisal structure.

4-2 The model by Scherer

In 1987, Leventhal and Scherer came up with a model for appraisal that assumes that an event is evaluated both sequentially on several levels of importance and on different levels of processing [34]. The lowest level of processing is the sensory-motor level, which is determined by innate feature detection and reflex systems. It is followed by the schematic level, where checks occur according to learned schemata. The highest level is the conceptual level, where evaluation takes place based on propositional memory storage and conscious processing is active. The levels of importance are ordered in the following groups:

1. Relevance: does the event affect me or my reference group?
2. Implication: do the consequences of the event affect my well-being and immediate or long-term goals?
3. Coping potential: can I deal with the consequences of the event?
4. Normative significance: what is the impact of the event on my self-concept and how does it relate to the social norms and values?

These groups contain a number of notions, referred to by Scherer as a Stimulus Evaluation Check (SEC). All SECs take place on one of the earlier mentioned processing levels. For example, the relevance group contains three different novelty checks: suddenness (sensory-motor level), familiarity (schematic level) and predictability (conceptual level).

The separate SECs are rated on an appropriate scale. In most cases, that means they are evaluated between low and high. Exceptions are, among others, the conduciveness, which is either obstructive or somewhere between low and high and the discrepancy from expectation, which is either consonant or dissonant. In recent work [35], Scherer produced a table denoting his hypothesis for the values of all SECs for a number of different emotions, reprinted in figures 4-2 and 4-3. We use this table later as a reference, even though only some of it has been proven experimentally correct. Emotions in the Scherer model can thus be found by checking all different SECs and finding the closest matching emotion(s) by tabular lookup.

4-2-1 Evaluation groups

Relevance

In the relevance group, we find the novelty, intrinsic pleasantness and goal / need releveance SECs. Novelty, subdivided in suddenness, familiarity and predictability, represents in how far a stimulus is novel and deserving attention. It can be anything from very low to very high, with higher values usually attributed to more intense emotions. Intrinsic pleasantness evaluates whether a stimulus is likely to result in something positive or negative. It is the fundamental reaction to the stimulus, without considering future consequences and ranges from very low to very high. Goal / need relevance represents the relevance of stimulus for short- or long-term goals and is defined from low to high.

Implication

The implication group contains two cause SECs, one for agency and one for the motive. Also belonging to this group are outcome probability, discrepancy from expectation, conduciveness and urgency. The cause checks for agency checks who or what is the cause of the stimulus, which can be either another agent, the individual itself or natural. Similarly, the motive of the stimulus is also checked, which can either be intententional, by chance or through negligence. Outcome probability is the likelihood with which consequences of an event are to be expected and varies from very low to very high, though most emotions are the result of a relatively high outcome probability. Discrepancy from expectation reflects upon expectations and can

	ENJ/HAP	ELA/JOY	DISP/DISG	CON/SCO	SAD/DEJ	DESPAIR	ANX/WOR
CRITERION							
Relevance							
Novelty							
Suddenness	low	high/med	open	open	low	high	low
Familiarity	open	open	low	open	low	very low	open
Predictability	medium	low	low	open	open	low	open
Intrinsic pleasantness	high	open	very low	open	open	open	open
Goal/need relevance	medium	high	low	low	high	high	medium
Implication							
Cause: agent	open	open	open	other	open	oth/nat	oth/nat
Cause: motive	intent	cha/int	open	intent	cha/neg	cha/neg	open
Outcome probability	very high	very high	very high	high	very high	very high	medium
Discrepancy from expectation	consonant	open	open	open	open	dissonant	open
Conduciveness	high	very high	open	open	obstruct	obstruct	obstruct
Urgency	very low	low	medium	low	low	high	medium
Coping potential							
Control	open	open	open	high	very low	very low	open
Power	open	open	open	low	very low	very low	low
Adjustment	high	medium	open	high	medium	very low	medium
Normative significance							
Internal standards compatibility	open	open	open	very low	open	open	open
External standards compatibility	open	open	open	very low	open	open	open

Figure 4-2: Predicted appraisal patterns for some major modal emotions, adapted from [35], Table 5.4

CRITERION	FEAR	IRR/COA	RAG/HOA	BOR/IND	SHAME	GUILT	PRIDE
Relevance							
Novelty							
Suddenness	high	low	high	very low	low	open	open
Familiarity	low	open	low	high	open	open	open
Predictability	low	medium	low	very high	open	open	open
Intrinsic pleasantness	low	open	open	open	open	open	open
Goal/need relevance	high	medium	high	low	high	high	high
Implications							
Cause: Agent	oth/nat	open	other	open	self	self	self
Cause: Motive	open	int/neg	intent	open	int/neg	intent	intent
Outcome probability	high	very high	very high	very high	very high	very high	very high
Discrepancy from expectation	dissonant	open	dissonant	consonant	open	open	open
Conduciveness	obstruct	obstruct	obstruct	open	open	high	high
Urgency	very high	medium	high	low	high	medium	low
Coping potential							
Control	open	high	high	medium	open	open	open
Power	very low	medium	high	medium	open	open	open
Adjustment	low	high	high	high	medium	medium	high
Normative significance							
Internal standards	open	open	open	open	very low	very low	very high
External standards	open	low	low	open	open	very low	high

ENJ/HAP: enjoyment/happiness; ELA/JOY: elation/joy; DISP/DISG: displeasure/disgust; CON/SCO: contempt/scorn; SAD/DEJ: sadness/dejection; ANX/WORR: anxiety/worry; IRR/COA: irritation/cold anger; RAG/HOA: rage/hot anger; BOR/IND: boredom/indifference; med: medium; oth: other; nat: natural; int: intentional; cha: chance; neg: negligence

Figure 4-3: Predicted appraisal patterns for other major modal emotions, adapted from [35], Table 5.4

be either consonant or dissonant. Conduciveness represents how much an event helps progress toward current goals and ranges from very low to very high if conducive. If the events is not conducive, it is simply referred to as obstructive. Finally, urgency evaluates in how far high priority goals and needs are endangered, while also considering temporal limitations. As a SEC, urgency can be anywhere from very low to very high.

Coping potential

Coping potential is divided into three SECs; control, power and adjustment. Control is the amount of control an agent has over an event or its consequences. For example, stochastic events have a very low control value as the agent can do nothing to change the outcome. Control can range from very low to very high. Power represents the amount of resources available for the agent to exert control and is therefore only defined when control is present. It can also be anywhere from very low to very high. Adjustment is the capability of an agent to deal with the consequences of an event, adapting to them if necessary. As before, it ranges from very low to very high.

Normative significance

Normative signifance consists of two SECs, one checking the internal standards and the other checking external standards. Internal standard checks are a reflection of the compatability of an event with one's own values and norms and can be anywhere from very low to very high. External standard checks do the same for the values and norms of other agents and are defined similarly from very low to very high.

4-2-2 Uses of the model by Scherer

The Scherer model has not been used often in computational modeling of emotions. Because of the wide range of inputs required (all SECs need to be evaluated), programs based on the model by Scherer typically require input by the user. Wehrle and Scherer have made an attempt at computationally modelling his own model by calculating the Euclidean distance for each column of the given table [36], which is later used in an actual implementation based on neural networks [37]. By using a formal approach to the model, Broekens is able to use Prolog to evaluate emotions in a generic PacMan game [38]. Research that attempts to use the model in a more general setting is however hard to find.

4-2-3 Summary

The model created by Scherer relies completely on tabular lookup. Although many different emotions can be found by using this model, the need to use the table decreases the scientific value of results for three reasons. First of all, if not all values of the SECs found correspond to a specific emotion, we can only assume the emotion that comes closest to be the correct one. Secondly, not all theorized SECs have scientific grounds, meaning that a part of the table is purely based on assumptions and can be downright wrong. This becomes clear from the conduciveness entry for the emotion guilt, which is set to high, an obvious error. Finally, the

use of terminology such as 'high' and 'very high' is subject to free interpretation and requires some reference value to be utilised appropriately. Such a reference value may prove to be different accross all simulations and where the line is drawn between 'high' and 'very high' can affect the outcome for specific emotions. All of these problems result in the fact that an emotion derived from the table may not be correct at all, which is something we definitely have to keep in mind when using this model.

Chapter 5

Mapping situations in Reinforcement Learning to emotions

In previous chapters, we have treated the theory on human learning, Reinforcement Learning (RL) and appraisal theory. We are now able to combine the knowledge gained through these theories into an appraisal model that uses elements of RL to find the emotion experienced in any situation. Appraisal models are based on the notion of an event, something that is not present in RL. First, we discuss how we deal with this. We follow this by defining a number of restrictions to our mapping in order to ensure that it is usable in as many situations as possible. We then build appraisal models based on RL elements for both the OCC model and the model by Scherer, after which we compare the two to find which is more suited for the simulation.

5-1 Representing events in Reinforcement Learning

Both the OCC and Scherer model rely heavily on the notion of events as the cause of an emotion. Creating a mapping between RL and these appraisal theories therefore requires a sensible definition of an event in terms of RL. Events in appraisal theory are attributable, time-specific and usually have some consequence. The attributability is usually toward either the agent itself, some other agent, or natural causes. Time-specific means an event takes place at a certain moment in time. The consequences of an event might be direct or indirect, but a clear link should exist between the event and its consequences. If that link does not exist, emotions resulting from consequences can not be attributed to a specific event, which collides with the concept of appraisal. A state transition in RL satisfies these constraints. State transitions are always the result of an action, since doing nothing also counts as an action. It is clearly time specific; any state transition is bound to two time instances. The consequences of a state transition, either a reward or a change in value or both, are always specific to that state-transition. Events with direct consequences can therefore be accuractely represented by a state-transition, or the basic state-action-state (s, a, s')-triple. A problem arises when an

Table 5-1: List of available RL variables in the (s, a, s')-triple

RL Variable	Meaning
$V(s)$	Value of preceding state
$V(s')$	Value of resulting state
$P^a_{ss'}$	Probability of resulting state-change given action a
$R^a_{ss'}$	Expected reward given resulting state-change and action a
r	Actual reward received in state s'

action has consequences in the future, while other actions are executed in between. Imagine leaving the gas on, driving toward work and then finding a burning house after driving back. A classical RL agent may attribute the burning house to the action of driving back from work unless the "gas being turned on" is part of the state. This is a general problem in RL and a direct result of the assumption of satisfaction of the Markov Property. As a reminder, the Markov Property states that a specific action executed in a specific state always has the same probability distribution, regardless of any previous states. Using state-transitions as events is about as close as we can get, with the accuracy of that representation dependent on the completeness of the state. Using the (s, a, s')-triple does allow for prospect and retrospect reflections, if we consider s' as the state we just arrived in. The previous action and state are known and expectations $(P^a_{ss'}, R^a_{ss'})$ can be derived from the transition model, provided that one exists. Without such a model, no transition predictions can be made and any prediction will be completely based on the values. Assuming we do have a model, the (s, a, s')-triple by itself gives us a number of variables to work with, summarized in Table 5-1.

Apart from the given variables, an RL agent in a state also knows the actions that are available to it, while a transition model can provide an indication about states that may result from those actions. This summarizes all knowledge available based on the (s, a, s')-triple, which is all we can use to map emotions.

5-2 Mapping appraisal models to reinforcement learning

The formal definition of the event concept allows us to start on the actual mapping. However, there are some restrictions on this mapping. We start out by discussing these restrictions, after which we start on the actual mappings. The mappings are discussed per appraisal variable, which in the case of the OCC model also leads to a compound mapping as the emotion intensity is built up of a combination of appraisal variables.

5-2-1 Restrictions on the mapping

A mapping from appraisal to RL has several properties that must be satisfied if we want it to be applicable in a general RL setting. Being able to use the mapping in a wide range of scenarios increases the credibility of our statement that emotions are inherent to RL. An important requirement for the mapping is independence on the type of scenario. No matter in what environment the agent is learning, the mapping should be usable without any adaptations. We can meet this requirement by using no environment-specific variables in the

mapping. The same goes for the learning method, as the mapping should not directly depend on how the agent is learning. For instance, using the update signal of Temporal Difference Learning in our mapping could make the mapping invalid for Monte Carlo methods, that may have an entirely different update signal. The mapping can also not depend on the type of action-selection, meaning neither ϵ nor β may be part of it.

In this thesis we restrict ourselves to single-agent systems. This removes the need to deal with the effects of other agents and to define a social context in the RL-setting. It also limits the number of emotions we may represent. Any emotion that requires some other agent in the broadest sense is not mapped. In order to test the development and occurence of mapped emotions against expectations from reality, the environment should give rise to all other types of possible emotions, depending on the specific appraisal model. The behaviour required for attaining a reward should be non-trivial and easibly altered, as different behaviour may represent different emotions. A reward, either positive or negative, should be available somewhere and convergence to the optimal policy should not take too long.

These restrictions provide a guideline on both our mapping and the experimental setup. Details of the experimental setup are discussed later. Following these guidelines, we can now define our mappings for both the OCC model and that of Scherer.

5-2-2 OCC to Reinforcement learning

The OCC model has three emotion categories that are applicable to the single-agent system. The first of these categories is well-being, which comprises the joy and distress emotions. The second consists of the prospect emotions hope and fear. The final categorie contains the confirmation emotions fears-confirmed, relief, disappointment and satisfaction. The OCC model provides the internal variables that influence the intensity of each emotion group. We use these variables for our mapping, expressing them in RL variables where possible. We describe the reasoning behind the mapping for each group separately. Apart from internal variables, global variables affect the intensities of most emotions, so we have to map these as well.

Global variables

There are four global variables in the OCC model, which affect the intensity of all emotions. Sense of reality is the first of these and represents how real an event feels. Events that feel real have a much higher intensity, explaining the strong emotions resulting from dreams and the weaker ones resulting from imagining events. In RL, sense of reality does not play a role at all, as every event feels real to the agent. This variable can therefore not be mapped. The second global variable is called proximity and relates to the closeness of an event. Events happening to people closer to you, such as family, typically result in more intense emotions. This also goes for experiences closer in time or in space. By default, an RL agent lacks the social context to consider certain some people closer than other. Temporal proximity is, via the discount factor, incorporated in state or action-values, since further states have less effect on those values. The same can be said for spatial proximity, since the effect of events that are spatially further away typically take a longer time to have effect on the update. The next variable is unexpectedness. Unexpectedness reflects back on an event that has happened,

checking to what extent the event was expected. Events that are less expected usually are more intense. In RL, the expectedness of a state transition is represented by the probability of that transition given the previous action and calculated prior to executing the actual action. Computing the unexpectedness therefore requires a transition model of the environment. If that exists, the unexpectedness $U(s_{t-1}, a_{t-1}, s_t)$ of a state-transition is defined by the converse of the expectedness.

$$U(s_{t-1}, a_{t-1}, s_t) = 1 - P^{a_{t-1}}_{s_{t-1}s_t} \tag{5-1}$$

By definition, this unexpectedness is a result of the transition, not a value prior to it. The unexpectedness can only be experienced once state s' is reached. The final global variable is the arousal of the agent. It is about the level of physiological arousal, something that is not applicable for the basic RL agent and which we also do not map. We can therefore only map unexpectedness explicitly, while proximity is contained in the values. Since we were not planning on mapping emotions requiring a social context, the loss of social proximity is not a big problem. The arousal effect can not be used, so we assume that the agent is in a state of arousal similar to the average human to which we compare the results of the mapping.

Well-being emotion mapping

Well-being emotions depend on only one internal variable in the OCC model, which is the desirability of an event. There are a number of possible intepretations for desirability. The first is simply using the reward. Any transition that yields a reward has a desirability equal to that reward. In that case, the only desirable transitions are those that actually give a reward. However, we know from the human world that both anticipation and unexpected improvement can also result in joy [39]. A more accurate alternative is using both values and rewards. At first glance, high-valued states can be considered as desirable. However, we have to consider the desirability of a state-transition rather than that of a state. Taking the change in value as a measure for desirability fits this approach. A state-transition that has no value change or reward, such as staying in the same state without anything happening, then gives a desirability of 0. The previous example represents no change at all, so neutral desirability is a logical outcome. A third alternative is the use of the update signal as a measure for desirability. The update signal represents the change of value for the previous state and therefore may be seen as representation for the quality of the state-change; if the update signal is positive, things have gone better than expected and vice versa. However, this would make the emotion specific to the learning method as the value of the update signal depends on that method. We require our emotion mapping to yield the same results under any learning method, so we cannot use this alternative.

Unexpectedness also has a large influence on happiness. Events that can be precisely predicted typically yield little if any emotion, a concept also known as habituation [7]. Emotions get stronger if an event is more unexpected. Such a relationship can be quantified by multiplying unexpectedness and desirability. Given our previous choice for desirability, this leads to the following compound formula for the well-being $WB(s_{t-1}, a_{t-1}, s_t)$ of the state-transition that just happened:

$$WB(s_{t-1}, a_{t-1}, s_t) = U(s_{t-1}, a_{t-1}, s_t)D(s_{t-1}, a_{t-1}, s_t) \tag{5-2}$$

Where the desirability $D(s_{t-1}, a_{t-1}, s_t)$ is defined by:

$$D_V(s_{t-1}, a_{t-1}, s_t) = V(s_t) - V(s_{t-1}) + r(s_{t-1}, a_{t-1}, s_t) \tag{5-3}$$

or, if we use action values:

$$D_Q(s_{t-1}, a_{t-1}, s_t) = \max_{a_t} Q(s_t, a_t) - Q(s_{t-1}, a_{t-1}) + r(s_{t-1}, a_{t-1}, s_t) \tag{5-4}$$

Note that the emotion occurs after the transition, analogue to unexpectedness which is determined only after the transition. We can see that this mapping habituates by noting that values converge toward the actual outcome, such that the desirability converges toward a low number depending on the discount factor γ.

Prospect emotion mapping

Prospect emotions consider future events, or possible state-changes given the current state. In the OCC model, the prospect emotion depends on both the desirability of the prospected event as well as its likelihood. Unexpectedness is undefined in the prospect case, since it is looking forward. For prospect desirability, we can follow the same reasoning as desirability in the well-being case. However, since it is a prospect quantity, prospect desirability should be defined per transition. The reward is now dependent on expectation since it can not be observed until after the state transition. The prospect desirability becomes $D_p(s, a, s')$:

$$D_{p,V}(s, a, s') = V(s') - V(s) + R^a_{ss'} \tag{5-5}$$

and in case of action values:

$$D_{p,Q}(s, a, s') = \max_{a'} Q(s', a') - Q(s, a) + R^a_{ss'} \tag{5-6}$$

This equation gives a set of prospect desirabilities for every defined state-action-state combination. The likelihood $L(s, a, s')$ is also defined per combination and is simply the transition probability:

$$L(s, a, s') = P^a_{ss'} \tag{5-7}$$

Creating a compound emotion from these variables is more complicated. We can choose to let hope and fear occur simultaneously or define the strongest of both to be the leading emotion. The OCC model generally assumes this last case, meaning we take the maximum of the hope and fear values. On the other hand, several references [40, 41, 42] suggest that both emotions occur at the same time, although it also seems the strongest determines the decision made. We start out by taking this approach and therefore need to define two compound prospect emotions: hope and fear.

Hope and fear are a compound of likelihood and desirability, meaning the simplest approach to calculating hope $H(s)$ and fear $F(s)$ is taking the maximum and minimum value of the

multiplication of both these. We can choose to take the maximum and minimum over all possible actions:

$$H(s) = \max_a \sum_{s'} L(s,a,s') D_p(s,a,s') \tag{5-8}$$

$$F(s) = \min_a \sum_{s'} L(s,a,s') D_p(s,a,s') \tag{5-9}$$

Alternatively, we can base hope and fear on possible future states, in which case we have to find the probability to get in the specific state given any action. If we assume equal probability of chosing any action, this leads to:

$$H(s) = \max_{s'} \sum_{a} L(s,a,s') D_p(s,a,s')/n(a) \tag{5-10}$$

$$F(s) = \min_{s'} \sum_{a} L(s,a,s') D_p(s,a,s')/n(a) \tag{5-11}$$

where $n(...)$ is the number of the entity between the parentheses. We can also skip the equal probability assumption and use the known policy instead:

$$H(s) = \max_{s'} \sum_{a} \pi(s,a) L(s,a,s') D_p(s,a,s') \tag{5-12}$$

$$F(s) = \min_{s'} \sum_{a} \pi(s,a) L(s,a,s') D_p(s,a,s') \tag{5-13}$$

We now note that Eq. (5-8) is similar to Eq. (3-9) while Eq. (5-12) is similar to Eq. (3-7). The Bellman equations can be used as an update function for values similar to Eq. (3-12), making it a measure for the change in value from one state to the next as well. This suggests the possibility of simply using state- or action-values as a measure for hope and fear instead, such that $HF(s)$, the hope or fear value becomes:

$$HF_V(s) = V(s) \tag{5-14}$$

in case of state-values. For action-values, we get:

$$HF_Q(s) = \begin{cases} \max\limits_a Q(s,a) & \text{if } \max\limits_a Q(s,a) > |\min\limits_a Q(s,a)| \\ \min\limits_a Q(s,a) & \text{otherwise} \end{cases} \tag{5-15}$$

Both value types implicitly represent an expected return of future states weighted by the probability to get in those states, making them a valid representation for hope and fear. We should note that using state-values results in either hope or fear as a resulting emotion, as it yields only one number. This makes it a very different approach from the first three, meaning we have to test its results against one or a few of those. This approach assumes the strongest

emotion as the one felt at the time, meaning the separate hope and fear emotions are not available anymore in the end result.

For our simulations, we have to pick one of the first three alternatives. The choice between having the prospect emotion based on states or actions is similar to the choice between state- or action-values in the learning mechanism. So, we could use the first option if we have action-values and the second or third if we use state-values. Upon close study, the use of equations 5-11 or 5-13 results in constant fear if a loss of value is available, which is always the case unless all states have the same value. This means that in a very positive world, where just about each action yields some reward, the agent could still be afraid to make a wrong move. We do not consider this realistic and therefore choose to use values to map prospect emotions. Considering the habituation of fear, we note that the policy of an agent converges toward picking the best action. This means that values converge upward, resulting in the prospec-emotion moving toward its maximum value. If a penalty results from a negative action, the agent will typically avoid that action, increasing the value of the state and thus habituating fear; the agent knows it controls the outcome of a situation and therefore loses initial fear as a result of choosing something else to do.

In conclusion, we will use $HF(s)$ for both state- and action-values.

Confirmation emotion mapping

Confirmation emotions are a retrospective view on earlier prospect emotions. Unexpectedness has a large influence, as high values of unexpectedness decrease the feeling of confirmation, up to the point that well-being emotions take over if the unexpectedness is large enough. For example, strongly fearing something that is very likely to happen is more likely to result in joy than in relief if it does not happen [28]. Internal variables influencing confirmation are the intensity of the attendant prospect emotion, the effort expended into attaining or preventing the event and the degree of realization of the event. The attendant prospect emotion is found by storing the hope or fear in the previous state. If both were present, we assume the strongest of hope and fear as the leading prospect emotion in that state. Since the prospect emotion forms a prediction, it can also be used to determine the degree of realization. Confirmation is typically a percentage, so we use a ratio to determine the degree of realization. Since prospect emotions are a prediction about the desirability of the future state, the degree of realization is the ratio between actual outcome and expected outcome. The expectedness of the event than acts as a correction on the intensity of the confirmation emotion via multiplication. In RL, there is no measurement of effort. Some states may have rewards far in the future, but because of the Markov property rewards can not be attributed to actions in a specific state, since the outcome of the last state transition would be the same regardless of earlier actions. This renders the only possible representation of effort in a basic RL-setting, the number of time steps, obsolete. Alternatively, the energy used or something similar may be represented in the state itself, but since this is an adaptation of the basic setting, we do not consider this option. Therefore, effort is not mapped into any RL-variable. Our formula for the confirmation emotion $C(s_{t-1}, a_{t-1}, s_t)$ becomes:

$$C_V(s_{t-1}, a_{t-1}, s_t) = \frac{V(s_t) + r(s_{t-1}, a_{t-1}, s_t)}{HF_V(s_{t-1})}(1 - U(s_{t-1}, a_{t-1}, s_t)) \tag{5-16}$$

Table 5-2: Suggested mapping of the OCC model to Reinforcement Learning when using state-values

Intensity variable	RL mapping
Global - Unexpectedness $U(s_{t-1}, a_{t-1}, s_t)$	$U(s_{t-1}, a_{t-1}, s_t) = 1 - P^{a_{t-1}}_{s_{t-1}s_t}$
Well-being $WB(s_{t-1}, a_{t-1}, s_t)$ - Desirability $D(s_{t-1}, a_{t-1}, s_t)$	$WB_V(s_{t-1}, a_{t-1}, s_t) = U(s_{t-1}, a_{t-1}, s_t) D_V(s_{t-1}, a_{t-1}, s_t)$ $D_V(s_{t-1}, a_{t-1}, s_t) = V(s_t) - V(s_{t-1}) + r(s_{t-1}, a_{t-1}, s_t)$
Prospect $HF(s)$	$HF_V(s) = V(s)$
Confirmation $C(s_{t-1}, a_{t-1}, s_t)$	$C_V(s_t) = \frac{V(s_t)+r(s_{t-1},a_{t-1},s_t)}{HF_V(s_{t-1})}(1 - U(s_{t-1}, a_{t-1}, s_t))$

Table 5-3: Suggested mapping of the OCC model to Reinforcement Learning when using action-values

Intensity variable	RL mapping
Global - Unexpectedness $U(s_{t-1}, a_{t-1}, s_t)$	$U(s_{t-1}, a_{t-1}, s_t) = 1 - P^{a_{t-1}}_{s_{t-1}s_t}$
Well-being $WB(s_{t-1}, a_{t-1}, s_t)$ - Desirability $D(s_{t-1}, a_{t-1}, s_t)$	$WB_Q(s_{t-1}, a_{t-1}, s_t) = U(s_{t-1}, a_{t-1}, s_t) D_Q(s_{t-1}, a_{t-1}, s_t)$ $D_Q(s_{t-1}, a_{t-1}, s_t) =$ $\max_{a_t} Q(s_t, a_t) - Q(s_{t-1}, a_{t-1}) + r(s_{t-1}, a_{t-1}, s_t)$
Prospect $HF(s)$	$HF_Q(s) = \max_a Q(s, a)$ or $\min_a Q(s, a)$
Confirmation $C(s_{t-1}, a_{t-1}, s_t)$	$C_Q(s) = \frac{\max_{a_t} Q(s_t,a_t)+r(s_{t-1},a_{t-1},s_t)}{HF_Q(s)}(1 - U(s_{t-1}, a_{t-1}, s_t))$

and for action values:

$$C_Q(s_{t-1}, a_{t-1}, s_t) = \frac{\max_{a_t} Q(s_t, a_t) + r(s_{t-1}, a_{t-1}, s_t)}{HF_Q(s)}(1 - U(s_{t-1}, a_{t-1}, s_t)) \tag{5-17}$$

Conclusion

We were able to create mappings for well-being, prospect and confirmation emotion through the use of internal and global variables in the OCC model. By considering the notion of an event as an (s, a, s')-triple, it is possible to determine the intensities of each of these emotions. It should be noted that Eq. (5-17) requires knowledge of the set of actions a_t in state s_t, but this set is completely determined by that state, so we do not consider it an added variable. The mappings are listed in Table 5-2 and Table 5-3.

5-2-3 Scherer to Reinforcement Learning

In the Scherer model, a large number of properties of the event is evaluated, where each property is represented by a Stimulus Evaluation Check (SEC). A certain combination of evaluations of these SECs constitutes an emotion. Contrary to the OCC model, there are no specific influences of stimuli to some emotion. As a result, we cannot perform the mapping

emotion by emotion, but have to do it criteriom by criterion. This may result in a lot of discrepancies between emotions, so we have to be careful in defining that mapping. Scherer has defined four general groups in which the different criteria fall; relevance, implication, coping potential and normative significance. Our mapping considers the SECs per group.

Relevance

The relevance containes the novelty, intrinsic pleasantness, goal / need relevance and implication SECs. In the novelty group, Scherer considers three sensory-motor processing levels of novelty detection. At the lowest level, suddenness is considered, characterized by a stimulus that has an abrupt onset and relatively high intensity. On the medium level we find familiarity and on the highest level Scherer mentions predictability. Suddenness and predictability are practically opposites of each other and familiarity is hard to define in RL. We therefore simplify the novelty group to contain only predictability. This is a retrospect evaluation, as an event can only be assessed after it has taken place. The focus lies on the previous state and action and the current state, which leads to the transition probability of the state-transition as a representation for predictability $PR(s_{t-1}, a_{t-1}, s_t)$.

$$PR(s_{t-1}, a_{t-1}, s_t) = P^{a_{t-1}}_{s_{t-1}s_t} \tag{5-18}$$

The intrinsic pleasantness of an event is the amount of happiness that can be directly attributed to that event. A state transition has only one intrinsic measure for pleasantness; the reward of that transition. Values are a measure for the worth of the state when taking future results into account and never give a direct return. This clearly shows in terminating states where the value is always zero while they surely can be very pleasant if they yield large rewards, meaning intrinsic pleasantness $IP(s_{t-1}, a_{t-1}, s_t)$ can be mapped by using the reward.

$$IP(s_{t-1}, a_{t-1}, s_t) = r(s_{t-1}, a_{t-1}, s_t) \tag{5-19}$$

Goal or need relevance indicates how much an event is relevant to the agent and its goals. For a single-agent system that focuses on maximizing cumulative reward, each state-transition may be a step toward increasing (or decreasing) that cumulative reward. Actually, a state-transition is meant to do exactly that. A positive change in value constitutes a step forward, while a negative change is a step backward. Since reward is taken care of in the intrinsic pleasantness, the magnitude value change is the most accurate representation of relevance. If the value changes strongly, the state-transition is more relevant to the goal of maximizing return, whether the change is negative or positive. We represent relevance $RE(s_{t-1}, s_t)$ by the magnitude of the change in value for the transition.

$$RE_V(s_{t-1}, s_t) = |V(s_t) - V(s_{t-1})| \tag{5-20}$$

And similarly when using action values:

$$RE_Q(s_{t-1}, s_t) = |\max_{a_t} Q(s_t, a_t) - \max_{a_{t-1}} Q(s_{t-1}, a_{t-1})| \tag{5-21}$$

Implication

The implication group contains two cause SECs, for the agency and the motive of the event. It also comprises outcome probability, discrepancy from expectation, conduciveness and urgency. The cause element attempts to attribute results of an action to an agent and the motives behind the action. This has strong effects on the appraisal, as attributing something to yourself or someone else matters greatly, as does the fact if an action was intentional or not. The RL agent has no concept of such a cause attribution. Any state-transition that influences the agent is used to update its own model, whether the agent contributed to it or not. We therefore cannot map these SECs.

Outcome probability is a measure of the probability of the consequences of an event. In other words, it is a prospect variable that looks ahead to the possible state-transitions after the one that just happened. We consider all possible consequences and the outcome probability may be different for all of them. The outcome probability should therefore be defined for each possible future state-transition separately, evaluating the probability of that transition given some action. We need to use the probability for picking any action, which we assume equal for all actions in this case since outcome probability in this case is evaluated without considering ones own actions. This makes sense as the emotion attributed to a consequence is used as a means of selecting the proper action in the current state ; the action that has the highest chance of reaching a pleasant state is much more likely to be selected, meaning evaluation of consequences happens prior to the action-selection. Defining outcome probability $OP(s_t, s_{t+1})$ in this way allows an agent to be afraid in a situation where it still is free to pick an action that does not have any negative consequence.

$$OP(s_t, s_{t+1}) = \sum_{a_t} P^{a_t}_{s_t s_{t+1}} / n(a_t) \tag{5-22}$$

Discrepancy from expectation is a comparison between the expected and actual outcome. It may be consonant or dissonant, that is, meeting expectation or not. Expectations are represented by the value of a state. Discrepancy from expectation can then be defined as either the ratio or the difference between actual outcome and expectation. In the case of difference, an expectation of 10 with an outcome of 1 would give the same discrepancy as an expectation of 10000 with an outcome of 9991, where one would expect a lot more disappointment in the first case. Therefore, we choose to use the ration. Since the expectation is about both future values and rewards, the measure for discrepancy $DI(s_{t-1}, a_{t-1}, s_t)$becomes:

$$DI_V(s_{t-1}, a_{t-1}, s_t) = (V(s_t) + r(s_{t-1}, a_{t-1}, s_t))/V(s_{t-1}) \tag{5-23}$$

With action-values, we find:

$$DI_Q(s_{t-1}, a_{t-1}, s_t) = (\max_{a_t} Q(s_t, a_t) + r(s_{t-1}, a_{t-1}, s_t))/\max_{a_{t-1}} Q(s_{t-1}, a_{t-1}) \tag{5-24}$$

Both are similar to the confirmation in the OCC mapping.

Conduciveness is a measure that evaluates the event in terms of the degree to which it helps to attain some or several of the current goals and needs. The sole purpose of the agent is

maximizing its reward over time, so conduciveness means how much the current state helps in increasing the total reward. The value of a state indicates precisely that quantity, as it symbolizes what rewards may be expected from that state in the future. The change in value effected by the state-transition represents if the agent is taking a step forward or back, a measure for conduciveness. Actually receiving a reward is of course conducive as well, meaning the formula for conduciveness $CO(s_{t-1}, a_{t-1}, s_t)$ becomes:

$$CO_V(s_{t-1}, a_{t-1}, s_t) = V(s_t) + r(s_{t-1}, a_{t-1}, s_t) - V(s_{t-1}) \tag{5-25}$$

Or, using action-values:

$$CO_Q(s_{t-1}, a_{t-1}, s_t) = \max_{a_t} Q(s_t, a_t) + r(s_{t-1}, a_{t-1}, s_t) - \max_{a_{t-1}} Q(s_{t-1}, a_{t-1}) \tag{5-26}$$

This equals the desirability in the OCC mapping, so we are actually saying that the conduciveness of a state change toward the goal of increasing the total reward equals the desirability of that state change in terms of maximization of utility. Since in RL the maximization of utility comes down to maximization of the total reward this is a logical statement.

The urgency of an event denotes the importance of dealing with that event. Urgency should be high if the event is likely to take place in the near future, or if it is a high threat to current goals and needs. In a setting where maximizing reward is the ultimate goal, urgent states are states with a high negative conduciveness, which are states that obstruct the progress of the agent towards maximising its reward. Of course, only relevant states can be urgent, and states with positive conduciveness are never urgent, disregarding the erroneous entry for guilt. We measure urgency in terms of the magnitude of the negativity of the conduciveness. This presents us with a problem in case of quantification; a negative conduciveness has the same value as its urgency, albeit with a different sign. However, in the table made by Scherer, all negatively conducive events are called obstructive rather than being quantified. If negative conduciveness constitutes an urgency, this actually works out similar. The urgency $UR(s_{t-1}, a_{t-1}, s_t)$ created by the transition to state s' becomes:

$$UR(s_{t-1}, a_{t-1}, s_t) = \begin{cases} |CO(s_{t-1}, a_{t-1}, s_t)| & \text{if } CO(s_{t-1}, a_{t-1}, s_t) < 0 \\ 0 & \text{if } CO(s_{t-1}, a_{t-1}, s_t) \geq 0 \end{cases} \tag{5-27}$$

Coping potential

The coping potential group contains the control, power and adjustment SECs, which consider how well the agent can deal with the event that just took place. Control is a measure for the probability of preventing or bringing about an event by a natural agent. That is, whether the occurrence or consequences of some event may be influenced by someone. Consequences of an event in RL are rewards and value-changes as well as the resulting set of possible next states. In order to determine control in an RL setting, we need to quantify in how far the future state-transitions may be influenced by the agent, as the value-change and reward are fixed. If any action available to an agent results in a negative conduciveness, the situation deteriorates no matter what the agent does. This represents a form of uncontrollability. On the other hand, if every action results in a positive conduciveness, the situation is under

control. Therefore, we define control $CT(s_t)$ as the percentage of available actions that results in a positive conduciveness:

$$CT_V(s_t) = \frac{n(a_t)|\sum_{s_{t+1}} P^{a_t}_{s_t s_{t+1}}(V(s_{t+1}) + R^{a_t}_{s_t s_{t+1}}) > V(s_t)}{n(a_t)} \tag{5-28}$$

With action-values we find:

$$CT_Q(s_t) = \frac{n(a_t)|\sum_{s_{t+1}} P^{a_t}_{s_t s_{t+1}}(\max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) + R^{a_t}_{s_t s_{t+1}}) > \max_{a_t} Q(s_t, a_t)}{n(a_t)} \tag{5-29}$$

Power measures the amount of control for the agent, in terms of improving its situation. However, a similar problem as for effort exists; there is no indication of the means an agents possesses in order to bring about a specific state change. We therefore do not map power using RL elements.

Adjustment measures the ability of an agent to cope with the consequences of an event, whether they are positive or negative. Negative effects have a high adjustment if their effect in the long run is positive, while positive effects have a high adjustment value by definition. The direct consequences of an event are reflected in the conduciveness of the event, so the adjustment should consider the possible conduciveness of future states with respect to the conduciveness realised by the state change that just took place.

$$AD_V(s_t) = \max_{a_t} \sum_{s_{t+1}} \left[P^{a_t}_{s_t s_{t+1}}(V(s_{t+1}) - V(s_t) + R^{a_t}_{s_t s_{t+1}}) \right] - CO_V(s_{t-1}, a_{t-1}, s_t) \tag{5-30}$$

Similarly, for action-values we find:

$$AD_Q(s_t) = \max_{a_t} \sum_{s_{t+1}} \left[P^{a_t}_{s_t s_{t+1}}(\max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) + R^{a_t}_{s_t s_{t+1}}) \right] - CO_Q(s_{t-1}, a_{t-1}, s_t) \tag{5-31}$$

Normative significance

Normative significance considers the norms of the society an agent is living in. In a single-agent system, there is no concept of a society. We can therefore not map any SEC of the normative significance group to basic RL-variables.

Conclusion

We were able to create mappings for well-being, prospect and confirmation emotion through the use of internal and global variables in the OCC model. By considering the notion of an event as an (s, a, s')-triple, it is possible to determine the intensities of each of these emotions.

Table 5-4: Suggested mapping of the Scherer model to Reinforcement Learning when using state-values

SEC	RL variables
Relevance	
Predictability	$PR(s_{t-1}, a_{t-1}, s_t) = P^{a_{t-1}}_{s_{t-1}s_t}$
Intrinsic pleasantness	$IP(s_{t-1}, a_{t-1}, s_t) = r(s_{t-1}, a_{t-1}, s_t)$
Goal / need relevance	$RE_V(s_{t-1}, s_t) = \lvert V(s_t) - V(s_{t-1})\rvert$
Implication	
Cause: agent	Unmapped
Cause: motive	Unmapped
Outcome probability	$OP(s_t, s_{t+1}) = \sum_{a_t} P^{a_t}_{s_t s_{t+1}}/n(a_t)$
Discrepancy from expectation	$DI_V(s_{t-1}, a_{t-1}, s_t) = (V(s_t) + r(s_{t-1}, a_{t-1}, s_t))/V(s_{t-1})$
Conduciveness	$CO_V(s_{t-1}, a_{t-1}, s_t) = V(s_t) + r(s_{t-1}, a_{t-1}, s_t) - V(s_{t-1})$
Urgency	$UR(s_{t-1}, a_{t-1}, s_t) = \lvert CO(s_{t-1}, a_{t-1}, s_t)\rvert$ or 0
Coping potential	
Control	$CT_V(s_t) = \frac{n(a_t)\lvert \sum_{s_{t+1}} P^{a_t}_{s_t s_{t+1}}(V(s_{t+1})+R^{a_t}_{s_t s_{t+1}})>V(s_t)}{n(a_t)}$
Power	Unmapped
Adjustment	$AD_V(s_t) = \max_{a_t} \sum_{s_{t+1}} \left[P^{a_t}_{s_t s_{t+1}}(V(s_{t+1}) - V(s_t) + R^{a_t}_{s_t s_{t+1}})\right] - CO_V(s_{t-1}, a_{t-1}, s_t)$
Normative significance	
Internal standards	Unmapped
External standards	Unmapped

It should be noted that Eq. (5-17) requires knowledge of the set of actions a' in state s', but this set is completely determined by state s', so it is not an added variable.

We have created mappings for SECs in 3 of the 4 groups defined by Scherer. Outcome probability requires a choice for the next state s'', since the emotion is used as a means of action selection and not the other way around. This requirement goes beyond our original choice for the (s, a, s')-triple. Another result of using Scherer is that we have to find an emotion through the tabular lookup. Although we are able to quantify all the SECs using the given mapping, the table uses unspecific terms from "very low" to "very high". This provides an extra problem as defining that range in terms of numbers is no straightforward task.

The mappings are listed in Table 5-4 and Table 5-5.

5-2-4 Comparison between the OCC and Scherer mappings

Having defined the OCC and Scherer mappings, we compare the two here. The OCC mapping is simpler, as it requires no summation over future states. It also quantifies 3 different emotion types at the same time, while requiring no input of variables beyond the (s, a, s') triple. Using action-values does require knowledge of those values in state s', but that knowledge is not hard to come by. The mappings for confirmation consist of a ratio, which may give problems when the denominator is 0. In the case of confirmation, a denominator of 0 means there was no

Table 5-5: Suggested mapping of the Scherer model to Reinforcement Learning when using action-values

SEC	RL variables
Relevance	
Predictability	$PR(s_{t-1}, a_{t-1}, s_t) = P^{a_{t-1}}_{s_{t-1}s_t}$
Intrinsic pleasantness	$IP(s_{t-1}, a_{t-1}, s_t) = r(s_{t-1}, a_{t-1}, s_t)$
Goal / need relevance	$RE_Q(s_{t-1}, s_t) = \lvert \max_{a_t} Q(s_t, a_t) - \max_{a_{t-1}} Q(s_{t-1}, a_{t-1}) \rvert$
Implication	
Cause: agent	Unmapped
Cause: motive	Unmapped
Outcome probability	$OP(s_t, s_{t+1}) = \sum_{a_t} P^{a_t}_{s_t s_{t+1}} / n(a_t)$
Discrepancy from expectation	$DI_Q(s_{t-1}, a_{t-1}, s_t) = \frac{(\max_{a_t} Q(s_t, a_t) + r(s_{t-1}, a_{t-1}, s_t))}{\max_{a_{t-1}} Q(s_{t-1}, a_{t-1})}$
Conduciveness	$CO_Q(s_{t-1}, a_{t-1}, s_t) = \max_{a_t} Q(s_t, a_t) + r(s_{t-1}, a_{t-1}, s_t) - \max_{a_{t-1}} Q(s_{t-1}, a_{t-1})$
Urgency	$UR(s_{t-1}, a_{t-1}, s_t) = \lvert CO(s_{t-1}, a_{t-1}, s_t) \rvert$ or 0
Coping potential	
Control	$CT_Q(s_t) = \frac{n(a_t) \lvert \sum_{s_{t+1}} P^{a_t}_{s_t s_{t+1}} (\max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) + R^{a_t}_{s_t s_{t+1}}) > \max_{a_t} Q(s_t, a_t) \rvert}{n(a_t)}$
Power	Unmapped
Adjustment	$AD_Q(s_t) = \max_{a_t} \sum_{s_{t+1}} \left[P^{a_t}_{s_t s_{t+1}} (\max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) + R^{a_t}_{s_t s_{t+1}}) \right] - CO_Q(s_{t-1}, a_{t-1}, s_t)$
Normative significance	
Internal standards	Unmapped
External standards	Unmapped

prospect-emotion. Confirmation is therefore undefined without any prospect-emotion, which makes sense as there is nothing to confirm. The ratio therefore poses no real problem for this mapping as long as we make sure to leave the confirmation undefined if the prospect-emotion is 0. The Scherer mapping contains a number of more complicated expressions requiring summations over states beyond s'. It also contains ratios, three of which have a denominator that becomes zero if no actions are available, as in a terminating state. This goes for outcome probability, control and power. Another ratio is based on the number of possible state, which becomes 0 in a terminating state as well. This means outcome probability, control, power and adjustment are undefined in terminating states. This makes sense as all these SECs make use of expectations which are absent in a terminating state. Besides that, the Scherer mapping requires a choice for a specific future state when considering the outcome probability. Emotions resulting from this state are per future state, which makes it hard to find out exactly what the agent is feeling since a large number of different emotions may be avaialable. Using the Scherer model also requires creating and calculating some distance function to find that closest match, since a perfect match is unlikely to occur. The distance function would then give a number of best fitting emotions and somehow a choice has to be made between them. Finally, using the table of emotions requires another mapping from the actual values of the SECs to the expressions used in the table, such as the range from "very high" to "very low". This means the mapping of emotions may be specific to the RL setting, as what is a high value in one case may be low in another. Apart from these observations, both mappings appear similar in many ways. Both require retrospect on the past action, requiring a transition model. Both use change in values as a representation for the utility of the state-transition. The OCC model does have a less specific prospect-view, although in both cases summations are taken over future actions and states.

5-3 Conclusion

Having examined the concept of an event in RL, we have come up with two mappings based on the OCC and Scherer models. The Scherer mapping was found not to meet earlier defined restrictions, most importantly because the tabular lookup requires an environment-specific mapping of SEC-values to nominal expressions; that is, the nominal notions ranging from very low to very high that are required for using the table differ per simulation and per setting. The OCC mapping does meet these restrictions and therefore is preferred. This complies with the general view that the OCC model is better fit to be used for computational appraisal. The OCC mapping is straightforward and simple, giving a direct quantification of well-being, prospect and confirmation emotions. The Scherer mapping requires manipulation of SEC quantification to achieve only one specific emotion per state. In conclusion, the OCC mapping is the best fit for use in the experiments.

Chapter 6

Simulations with a Reinforcement Learning-based appraisal model

Comparing human learning and Reinforcement Learning (RL) has suggested the possibility for a representation of emotions using elements of RL. We have created an appraisal model based on the OCC model that maps some of these elements to emotions. Theory about human learning and the use of emotions as action-selection has provided us with an impression of the type of simulation that is suited for testing the development of the mapped emotions against the known development in humans. This chapter describes our setup for that simulation as well as the hypotheses that flow forth from human learning theory and ends with a presentation of the results.

6-1 Experimental setup

In our simulation, we must make sure that we create results that are relevant to our research. Therefore, we have to define the boundary conditions of the experiments clearly. Since we are simulating a learning agent, the solition method itself is an important choice. It influences the development of the values over time as well as the policy of the agent. Although the convergence of most discussed solution methods is proven under certain circumstances, there is much to say for keeping the scenario simple. We are after all interested in the developing emotions and not in solving a difficult RL problem. Both the solution method and the scenario are discussed in this chapter, as well as the parameters resulting from our particular scenario choice.

6-1-1 Solution method

The most important choice for our scenario is the learning method. Temporal Difference Learning (TD) was earlier shown to be closely connected to human learning, so it seems a fitting choice. However, since TD is a model-free learning method, it has no available

transition model by default. Such a model is required to correctly represent several of the mapped emotions. It is quite simple to sustain such a model without actually using it, but we prefer a learning method similar to TD that does make use of such a model. Value-iteration (VI) is a model-based Dynamic Programming (DP) method that, given the complete model, continuously updates the complete state space until the difference between two iteration falls below a certain threshold. We can adapt this method to update only a visited state, based on the model at that time, such that, in the off-policy case:

$$V(s) = \max_a \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V_k(s') \right]. \tag{6-1}$$

and in the on-policy case:

$$V(s) = \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V_k(s') \right]. \tag{6-2}$$

where a is chosen by the action-selection method. This is a very simple method that nevertheless converges to the correct values under the same circumstances as any Monte Carlo (MC) method. In other words, the value of each state that is visited often enough to have an accurate transition model converges, yielding the optimal policy if all states meet this condition. If we look at TD with the learning rate α set to 1, we find:

$$V(s_t) \leftarrow r_{t+1} + \gamma V(s_{t+1}) \tag{6-3}$$

This expression is similar to VI; the difference is that the TD updates the previous value directly after the transition and value-iteration updates the transition model after the transition and the value upon entering the state the next time. The learning rate α in TD is similar to a sampling parameter that makes sure that the updates are not too big since they are all based on a single state transition. In that sense, the VI approach is more accurate, although all value-updates occur one step later. The advantage is that we make specific use of the transition model, while keeping all options for manipulating paramaters open. We therefore have chosen to use this approach in our simulations.

6-1-2 Scenario

Events in the scenario must have the potential to give rise to all mapped emotions. That is, both well-being and prospect emotions should be expected. Confirmation emotions are a result of prospect emotions and therefore fall in the same category. A foraging task allows an agent to roam around in a world containing one or more rewards. Joy or distress may be found upon collecting such a reward and hope and fear can occur in anticipation of finding a positive or negative reward, with their respective confirmations about the height of the reward. We have noticed earlier that a deterministic world is unlikely to result in strong emotions because of the importance of unexpectedness, so our task should also have the possibility of manipulating the amount of stochasticity. Furthermore, the task should be relatively simple but not straightforward, meaning several paths should be possible, where only one or a few lead to a reward. A maze which the agent has to navigate in order to find

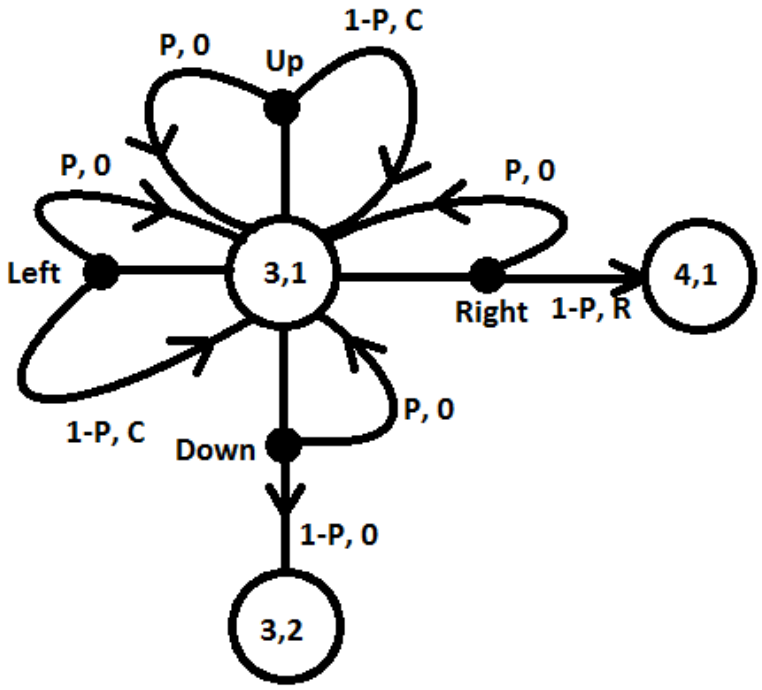

Figure 6-1: An example state transition graph in a state where the agent has a wall above and to the left of itself and a reward in the state to the right. P is the chance to stay in place, C is the collision penalty and R is the reward value.

a reward fulfills most of these requirements. In order to deal with stochasticity, we allow the agent to move toward the four cardinal directions while adding the possibility that it does not move after picking one of these. In this way, the effect of an action remains partially unpredictable. By adding a collision penalty when moving into a wall we motivate the agent to not make the wrong choice, which may have effect on the emotions. This setting is flexible enough to deal with a broad range of different scenarions, but simple enough to program effectively. An example state transition graph is given in Figure 6-1.

6-1-3 Scenario parameters

Given the maze setting, we have a number of parameters in the simulation that can be varied. The first of these that we have to test is the mapping of the prospect-emotions. Since we have determined two alternatives earlier, our first objective is finding out which of these mappings results in emotional development befitting theory and common sense. VI can be done both on- and off-policy, while the action-selection can also play an important role in the progress of the agent, especially during off-policy learning. Both of these are tested as well. The discount factor is the only RL-parameter present in VI and has to be manipulated to examine its effects.

By creating a random maze, we can test different path lengths to the reward. As discussed

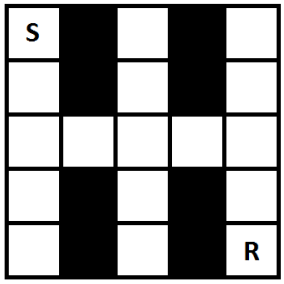

Figure 6-2: The maze used in the experiments. Each square is a state. The agent starts at S and the reward can (initially) be found at R.

before, both the collision penalty and the probability to remain in place after picking an action can influence emotions. To simplify the setting, we place only one reward in the maze. We let this reward be either positive or negative to test the behaviour and emotions resulting from both. Finally, collecting a reward may have one of two consequences. The first is returning to the starting location, while the reward remains in place. This allows the agent to learn the optimal behaviour quickly. The alternative is relocating the reward while leaving the agent in place, forcing it to explore while continously changing its values to adapt to the new circumstances. This is likely to result in very different emotions and both options have to be tested.

6-1-4 Experimental specifics

Using a Java program (pseudocode given in appendix A), we can now simulate the different settings. In order to keep the solution as simple as possible, we generate a single maze on a 5 by 5 grid, depicted in Figure .

We test hypotheses either in the control setting or by setting one of the parameters to the test value. In each experiment, we test 50 agents. For these agents, the values of β, γ, the collision penalty and the probability to fail an action are varied using a normal distribution such that 95% of the experimental values is within 5% of the given mean. Agents move through the maze for 10000 steps, which is sufficient for conversion, while the initial values of all states are set to 0 at the beginning of the experiment. In the plot, the setting using the control set of parameters is denoted as the first run, with the test set denoted as the second run. Table 6-1 shows the control and test values for each of the parameters.

6-2 Hypotheses

Before simulating, we define a number of hypotheses that flow forth from our main research question and theory on emotion. Our main research question concerns the possibility of representing human emotions using elements of Reinforcement Learning such that the development and occurence of those emotions is similar to that in humans. Theory on human emotions has presented us with an order in the developent of emotions from simple to complex [2]. We have

Table 6-1: Control and test values of different parameters used in the experiments

	Control value	Test value
Learning method	Off-policy VI	On-policy VI
Inverse temperature (mean)	10	2
Discount factor (mean)	0.9	1.0
Collision penalty (mean)	0.0	-0.01
P(remain in place) (mean)	0.1	0.0
Reward action	Return to start	Relocate reward
Reward value	1.0	-1.0

mapped three emotions, which are in theoretical order of occurence: well-being, prospect and confirmation. For simplicity, we refer to well-being as joy/distress and prospect emotions as hope/fear. In simulations, we expect to observe the specified order of occurence, hence:

Hypothesis 1. *In all simulations, joy/distress is the first emotion to be observed. It is followed by hope, which is followed by confirmation.*

Any other observation would go against human emotion development and thus prove our mapping incorrect. Habituation is a mechanic of emotion that is known to occur for both the joy and fear emotions [7, 8, 9, 14]. Like the development of emotions, habituation should also be observed in all simulations.

Hypothesis 2. *In all simulations where they exist, joy and fear habituate over time.*

The change from off-policy to on-policy learning represents the assumed amount of control over the action selection. An off-policy learning agent uses the maximum return of the possible actions as an update for learning, thus assuming it is always capable of making that choice. On-policy learning agents update based on the actual choice, meaning they take into account the action selection. In human terms, this is comparable to the difference between optimism and realism; the off-policy agent learns by expecting the best all the time while the on-policy agent learns on a realistic basis, knowing what choices it makes beforehand. Optimistic humans in general have a higher amount of hope [43], though the happiness in comparison to realistic humans may depend strongly on the circumstances. In a positive world such as the one we use for the basic tests, the expectations of the realistic agent are likely to be too low, possibly resulting in more happiness. We also expect the realistic agent to have a higher satisfaction value.

Hypothesis 3. *Optimism increase the intensity of hope, while realistic agents have more happiness and satisfaction in the given scenario.*

The action-selection itself is determined by the value of the inverse temperature β. In Subsection 3-2-2 it was determined to be a measure for the balance between exploration and exploitation. Our scenario is not build such that there is a need for a lot of exploration, so the impact of the inverse temperature should not be that large, though being too explorative may result in somewhat erratic behaviour.

Hypothesis 4. *Explorative agents in the given scenario behave less optimal, resulting in lower happiness.*

The discount factor is a measure for the impact of expected future reward to the evaluation. In other words, it determines whether the agent is a short- or long-term thinker. Long-term thinking in a positive setting increases the intensity of expectations; in any state, the agent knowns it can collect a reward somewhere in the future. However, a decrease in expectation increases the intensity of well-being emotions [9, 28], so we expect the optimistic agent to have a lower intensity of joy/distress.

Hypothesis 5. *Long-term thinking agents in the control setting have a higher amount of hope, while short-term thinking agents have a higher intensity of joy and distress.*

By varying several elements of the situation and looking at more theory on human learning, we can define several other hypotheses specific to the environment. First of all, we can add a penalty for making a false move. This lowers the expectations in any state, since a chance exists for making a false move. The advantage of a decrease in expectation is an increase in the intensity of joy [9, 28].

Hypothesis 6. *The presence of a penalty for making a wrong move decreases expectation, increasing fear or decreasing hope. The resulting lower expectation results in a higher intensity for joy/distress.*

This hypothesis assumes the agent does not always make the perfect move, since in that case, it would have nothing to fear. Predictability has been mentioned several times in the mapping of emotions. There are three ways to vary predictability. First of all, we can let the agent selects its actions at random. This results in an inefficient policy, but does not change the predictability of the effects of an action once it is chose. Secondly, we can make the results of an action stochastic, for example by letting the action fail completely every once in a while. A third method is handing out rewards at random points rather than at the same transition all the time. This effectively randomizes the reinforcing part of an experiment. The last two options increase the unexpectedness of the result of action and we test both of them. The increase in unexpectedness should in both cases increase the intensity of joy/distress [28, 35].

Hypothesis 7. *Unpredictability of the results of actions increases the intensity of the joy/distress emotion.*

The reward in the maze may be either positive or negative, representing a good or bad world; since their is only one reward in the maze, a negative reward means the maximum total reward achievable is 0. If the reward is negative, it can be expected that the agent learns to avoid the negative reward, resulting in it receiving no rewards at all.

Hypothesis 8. *If the reward is negative, this cause high fear and unhappiness in the beginning until the agent learns to avoid the reward.*

6-3 Results

Here, we present the results of our simulations and discuss them with respect to the hypotheses.

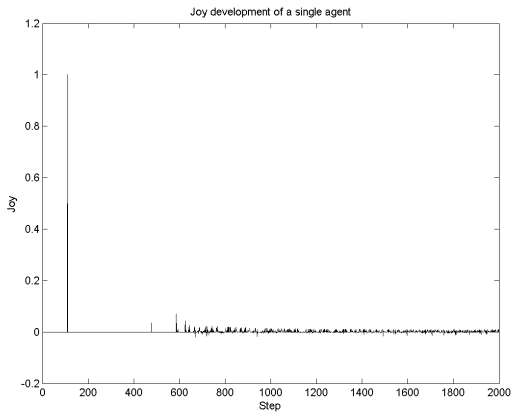

Figure 6-3: Intensity of joy/distress for a single agent, observed in the first 2000 steps

6-3-1 Results using control values

In the control case, we tested the development of emotions and habituation of joy/distress. The intensity of joy/distress for a single agent in the control setting during the first 2000 steps of the experiment is given in Figure 6-3.

We see that, for a number of steps, the agent feels nothing at all. A sudden spike then signals the first time the reward is collected. This updates the transition model for the state above the reward. The next time that state is visited, the value is updated based on the transition model. The reward is collected shortly afterward, but its intensity is 0, represented by the long window where no joy/distress is felt. This is caused by the unexpectedness of 0 in the transition model, which is based on earlier observation. The transition downward from the state above the reward has a 90% chance of succes, so the model is likely to be updated only around the 10th time the transition is made, although values are still updated in the meantime. Only once an action fails, the first spike starts appearing again. Still, this spike is much lower, since it is influenced by two factors. If the model converges toward the real values, the unexpectedness goes to 0.1 and the previous state value converges toward the reward based on Equation 6-1. Both factors strongly decrease the term in 5-2. Individual positive spikes are also caused by succesful transitions toward higher valued states, while the negative spikes are transitions toward lower valued states, both with low intensities caused by high expectedness. Moving toward and away from the goal causes joy and distress respectively, as intended by our mapping. The habituation of joy/distress becomes clear from this picture, since the intensity quickly goes down after the first time, validating the first part of Hypothesis 2.

From the same simulation, we plot the mean joy/distress and hope over all agents for the first 2000 steps. See Figures 6-4 and 6-5.

We can clearly see that joy appears some time before hope. We explained earlier that the values can only be updated once a reward is gained, which now also becomes clear from the simulation. Confirmation is directly dependent on hope, so it follows directly after. This validates Hypothesis 1.

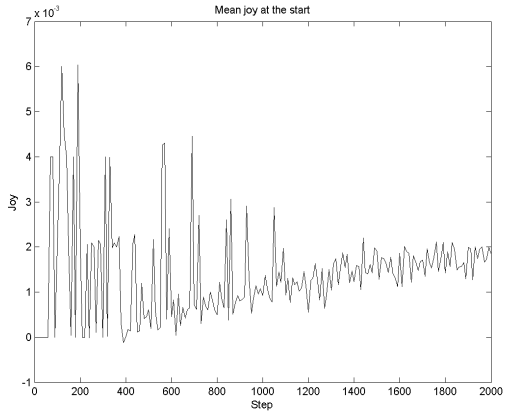

Figure 6-4: Intensity of joy/distress, mean over 50 agents, observed in the first 2000 steps

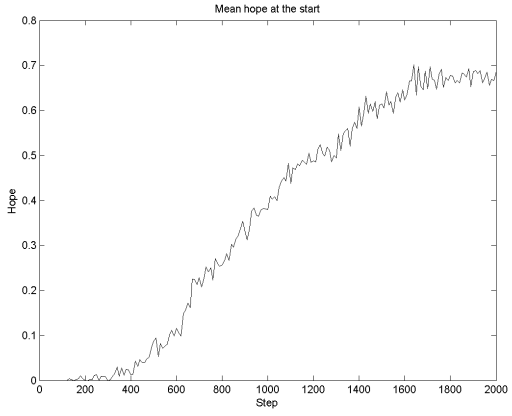

Figure 6-5: Intensity of hope, mean over 50 agents, observed in the first 2000 steps

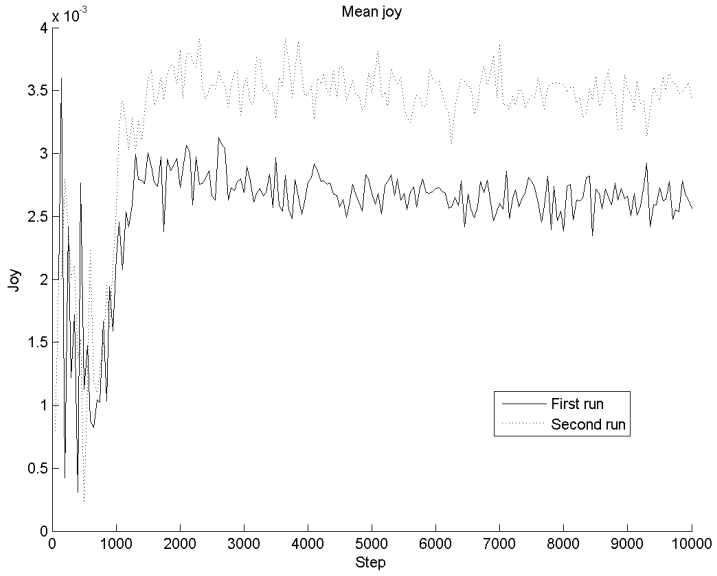

Figure 6-6: Intensity of happiness, mean over 50 agents, using off-policy (first run) and on-policy (second run) learning

6-3-2 Off-policy vs on-policy learning

In order to test the effect of positivism versus realism, we simulate the foraging task with both off-policy and on-policy VI. Results are shown in figures 6-6 up to 6-8.

Both options collect about an even number of rewards, but the happiness and hope show different values. The agent we referred to as the optimistic agent has a higher hope value, while the realistic agent has a higher happiness value. The increased hope for optimistic agents follows intuition and results in expectations that are too high. As a consequence, the happiness is lower. Mathematically, the state-values converge to a lower value for the on-policy agent, making the difference between states higher. Since happiness is based on the change in state-value, the realistic agent is happier. Both observations agree with Hypothesis 3, but contrary to expectations the satisfaction for the realistic agent is lower than for the optimistic agent. The realistic agent has lower values, meaning disappointment intensity is relatively higher when values decrease with the same amount.

6-3-3 Manipulating action-selection

Since the action-selection may have some small effects on the learning and thus development of emotions, we test two different values for the inverse temperature. Results are shown in figures Figure 6-9 to Figure 6-11.

The main effect of changing the temperature is a worse behaviour of the agent, making it collect less rewards, resulting in a lower value for happiness. Hope remains about equal, because the behaviour is not random enough to make the convergence slower in such a simple scenario. Satisfaction actually stays equals because of these similar expectations. This validates Hypothesis 4.

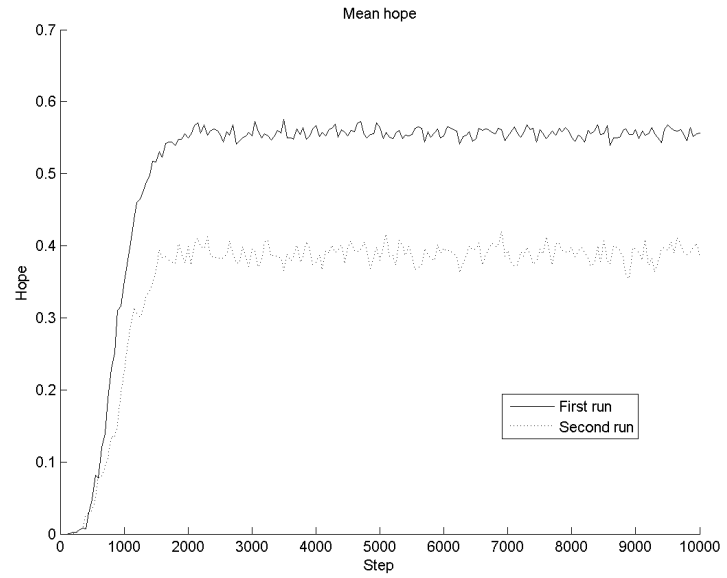

Figure 6-7: Intensity of hope, mean over 50 agents, using off-policy (first run) and on-policy (second run) learning

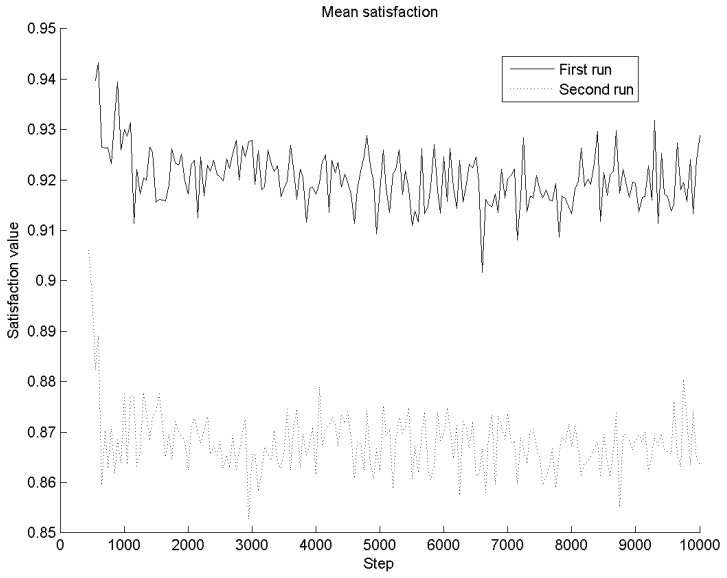

Figure 6-8: Intensity of satisfaction, mean over 50 agents, using off-policy (first run) and on-policy (second run) learning

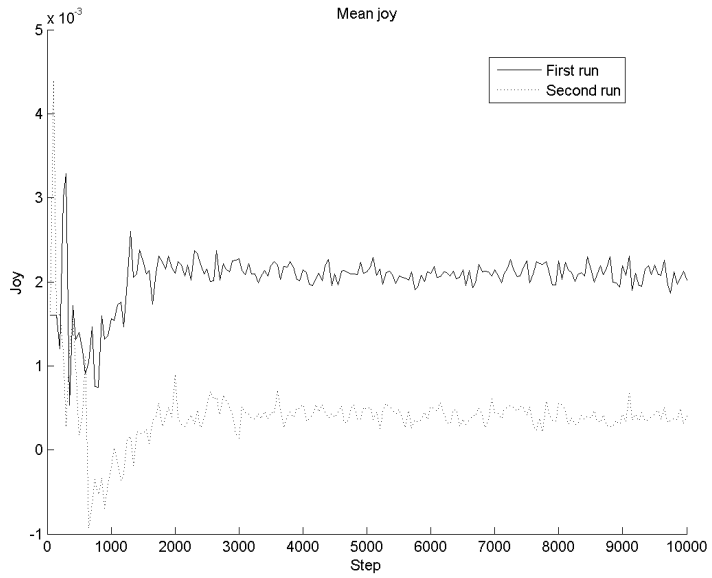

Figure 6-9: Intensity of happiness, mean over 50 agents, with $\beta = 10$ (first run) and $\beta = 2$ (second run)

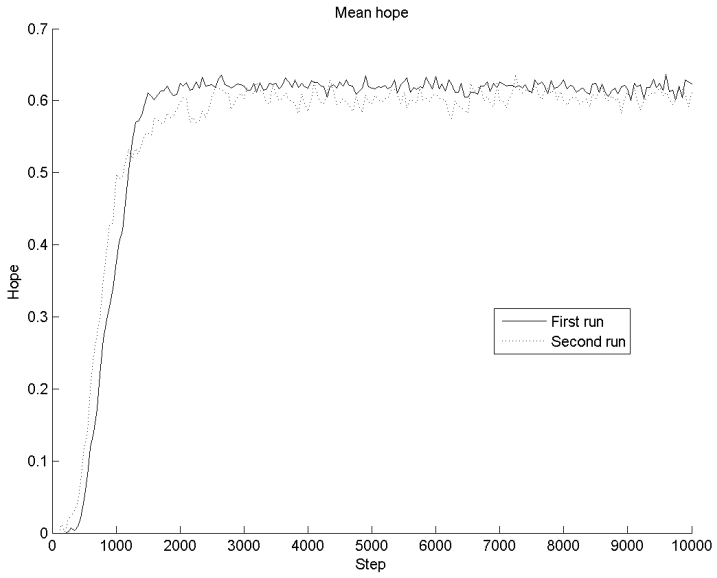

Figure 6-10: Intensity of hope, mean over 50 agents, with $\beta = 10$ (first run) and $\beta = 2$ (second run)

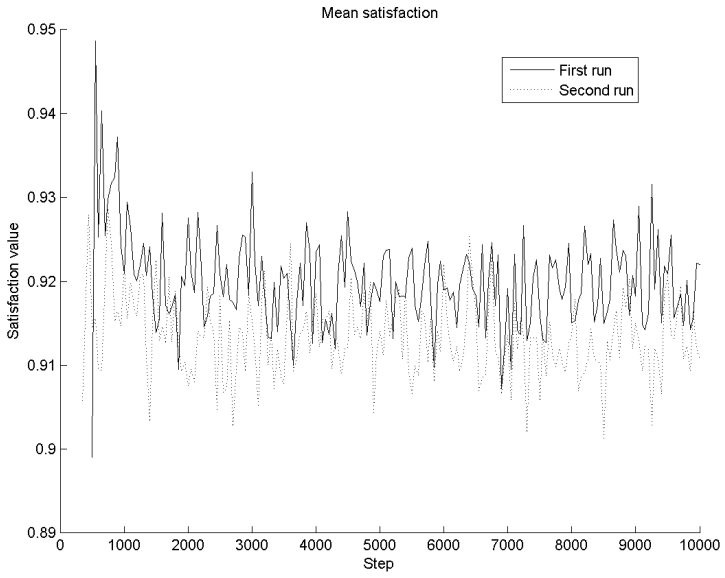

Figure 6-11: Intensity of satisfaction, mean over 50 agents, with $\beta = 10$ (first run) and $\beta = 2$ (second run)

6-3-4 Changing the discount factor

Long-term and short-term thinking agents may have very different emotions, even in such a simple scenario as the foraging task. We test a factor of 0.9 vs a factor of 0.5. See figures 6-12 to 6-15.

With a discount value of 0.5, hope decreases strongly as all states have lower values. The decrease in expectation results in higher intensity for joy/distress as is represented by the larger spikes in the corresponding figure. Combined with the knowledge that the agent actually collects less rewards, it is clear that the benefit of short-term thinking is a higher amount of happiness per collected reward. The benefit of long-term thinking appears to be a higher satisfaction and the collection of more rewards in total. These results validate Hypothesis 5.

6-3-5 Results with a collision penalty

We tested the effect of a constant possibility to receive a penalty on the emotions of the agent by adding a collision penalty in case the agent moves into a wall. We first show the effects of this setting on the fear of the agent in Figure 6-16

Fear goes to 0 after a small number of steps. The starting fear is caused by running into the wall a few times at the beginning of the simulation. However, the agent always has options available that never cause him to collide. Because of the max function in Equation 6-1, these types of actions take precendence in the value updates. Using this update function, if any action is available that causes no penalty whatsoever, the value is never negative. It represents an agent that assumes complete control over its actions and is therefore not afraid anymore once it knows how to avoid penalties. The agent updates its utility of the state and the initial

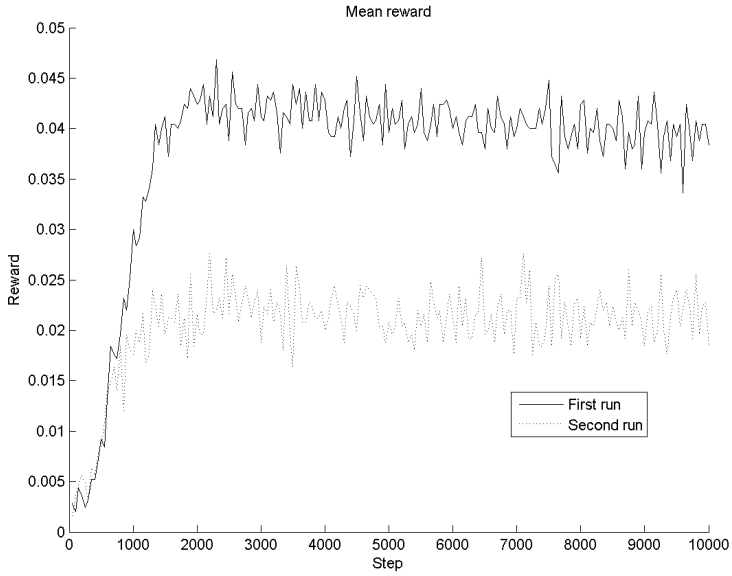

Figure 6-12: Collected rewards, mean over 50 agents, with $\gamma = 0.9$ (first run) and $\gamma = 0.5$ (second run)

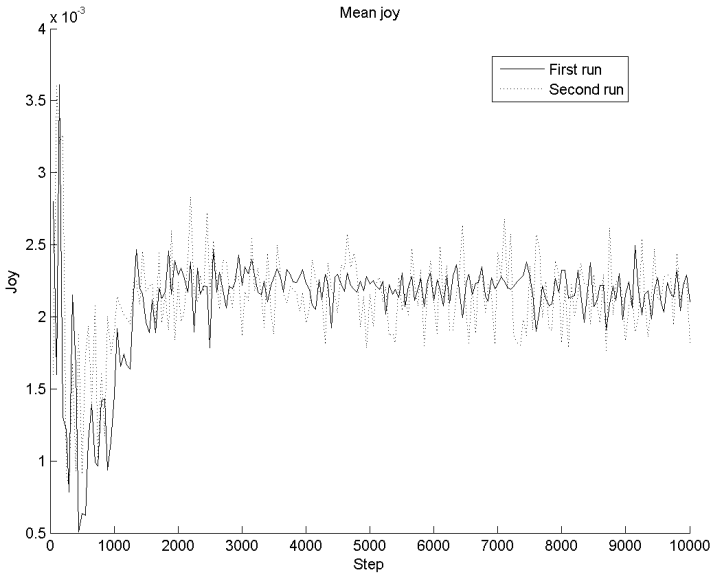

Figure 6-13: Intensity of happiness, mean over 50 agents, with $\gamma = 0.9$ (first run) and $\gamma = 0.5$ (second run)

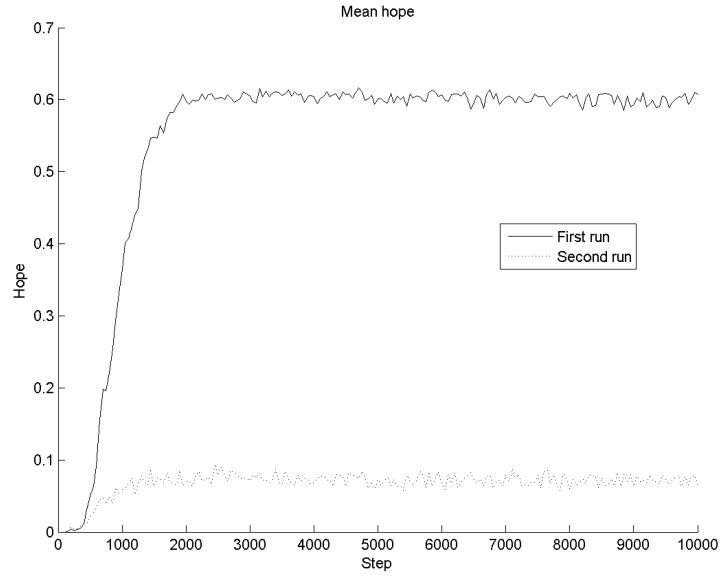

Figure 6-14: Intensity of hope, mean over 50 agents, with $\gamma = 0.9$ (first run) and $\gamma = 0.5$ (second run)

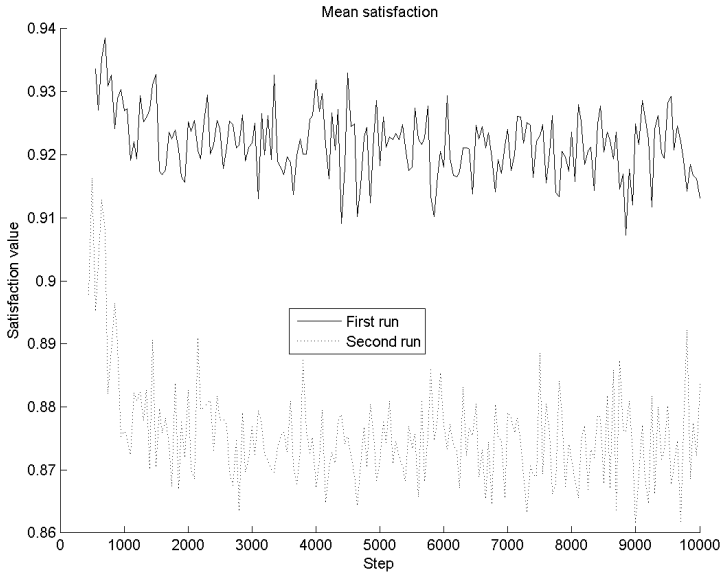

Figure 6-15: Intensity of satisfaction, mean over 50 agents, with $\gamma = 0.9$ (first run) and $\gamma = 0.5$ (second run)

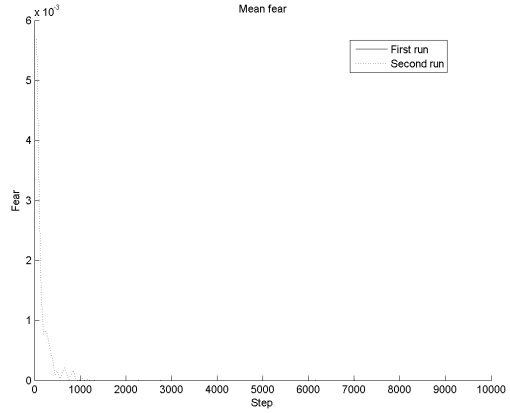

Figure 6-16: Intensity of fear, mean over 50 agents, in the presence of a collision penalty

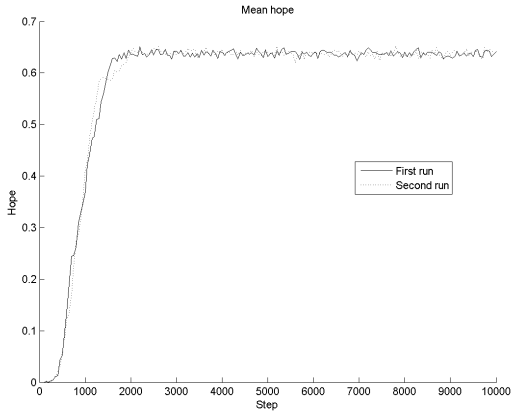

Figure 6-17: Intensity of hope, mean over 50 agents, without (first run) and with (second run) a collision penalty

fear habituates to 0. This demonstrates the mechanism of habituation in fear, validating the second part of Hypothesis 2.

The same experiment was used to investigate Hypothesis 6. The mean hope and joy over 50 agents with and without collison penalty are shown in Figures 6-17 and 6-18.

The intensity of hope is almost unaffected by the collision penalty, while joy decreases a bit. This falsifies our initial assumption from Hypothesis 6, which stated that the collision penalty would result in a decrease in expectation. Upon inspection, we find that the collision penalty actually causes a faster convergence of values. This has two reasons. First of all, as we discussed before, the use of max in Equation 6-1 results in the agent not considering the possibility for a penalty in the updates of values. This alone causes convergence that is just as fast as in the case without collision. Secondly, the Boltzmann action selection method ensures actions with potentially negative reward are chosen less often. The collision penalty actually causes the correct action to be chosen more often, resulting in faster convergence toward the optimal policy. As a result, values are slightly higher in the experiment with a collision penalty, decreasing the intensity of joy/distress since expectations are higher as well.

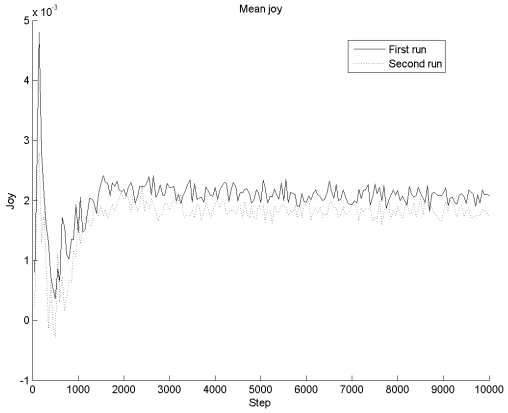

Figure 6-18: Intensity of joy/distress, mean over 50 agents, without (first run) and with (second run) a collision penalty

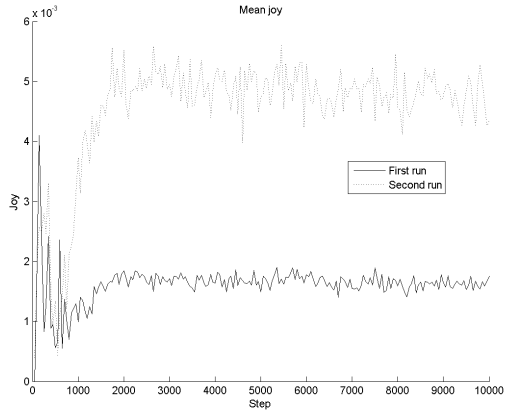

Figure 6-19: Intensity of joy/distress, mean over 50 agents, with a probability of 0.1 (first run) and 0.25 (second run) of failing an action

6-3-6 Results with increased unpredictability

We tested the effect of the predictability of the results of actions on the intensity of joy/distress in two separate experiments. First of all, we increased the probability of a chosen action failing, that is, remaining in place instead of moving in the desired direction. The resulting joy is shown in Figure 6-19. The second experiment consisted of randomly relocating the reward after each time it was collected, instead of returning the agent to the starting position. The joy intensity in these experiments is shown in Figure 6-20.

Both experiments increase the unexpectedness, in the first case for a succesful transition, in the second case for a rewarded transition. This causes an increase in intensity of the joy/distress emotions which can clearly be observed in both graphs. The effect of relocating the reward is much more prominent, since it reduces the predictability of a reward following a specific transition from close to 100% to about 6%, or 1 out of the 17 available states. This reduction is greater than making it 2.5 times more likely that an action fails, which is reflected in the larger intensity increase. Furthermore, the randomness of receiving rewards

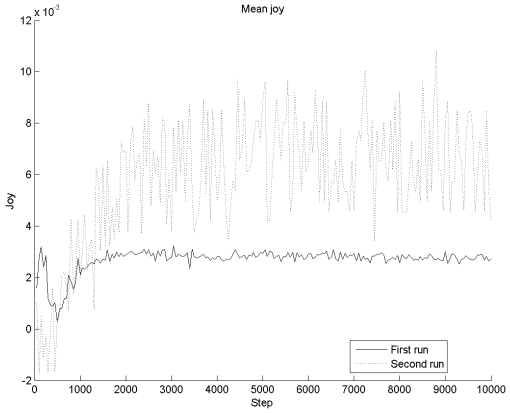

Figure 6-20: Intensity of joy/distress, mean over 50 agents, returning the agent (first run) or relocating the reward (second run)

also counteracts the habituation mechanism. Repeated rewards following the same action are sparse, so habituation does not really take place. Therefore, the intensity of the joy felt when receiving a reward does not decrease over time. The results of both experiments validate Hypothesis 7.

6-3-7 Negative vs positive reward

The value of the reward may be positive or negative, which can strongly influence emotions. Figures 6-21 to 6-24 show the results.

The agent quickly learns to avoid the negative reward, resulting in it keeping away from that reward. This also causes a small bit of fear when the agent gets close to the reward sometimes. No fears-confirmed emotions are experienced since the agent very quickly learns to avoid actions that lead close to the negative reward. This is also a form of habituation as the fear quickly goes to 0. Negative-valued states after a while are only reached by accident, as the result of the action-selecion. Results of this simulation comply with Hypothesis 8.

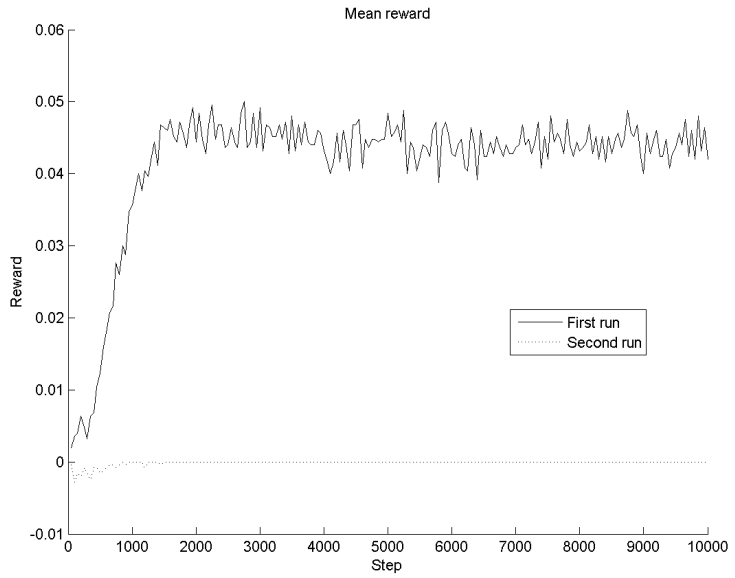

Figure 6-21: Collected rewards, mean over 50 agents, with a positive (first run) and negative (second run) reward

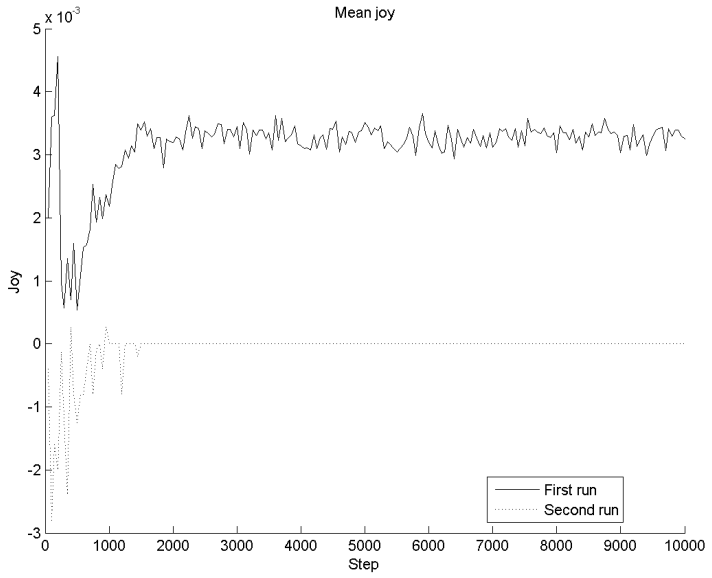

Figure 6-22: Intensity of joy/distress, mean over 50 agents, with a positive (first run) and negative (second run) reward

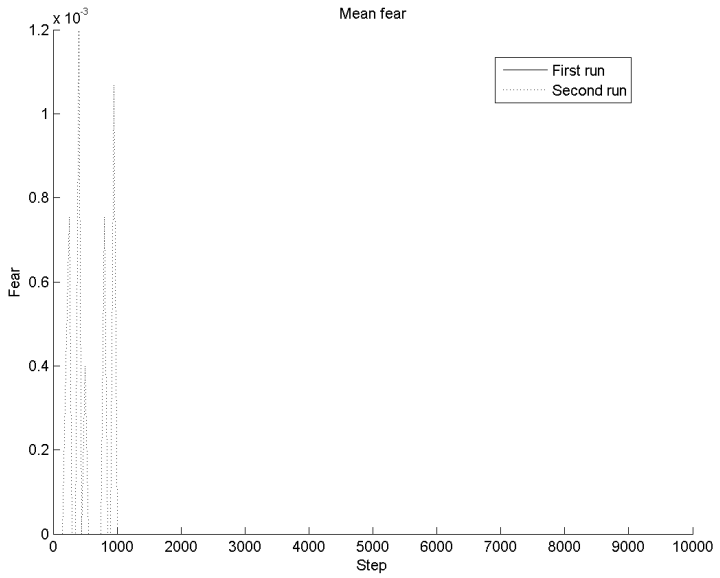

Figure 6-23: Intensity of fear, mean over 50 agents, with a positive (first run) and negative (second run) reward

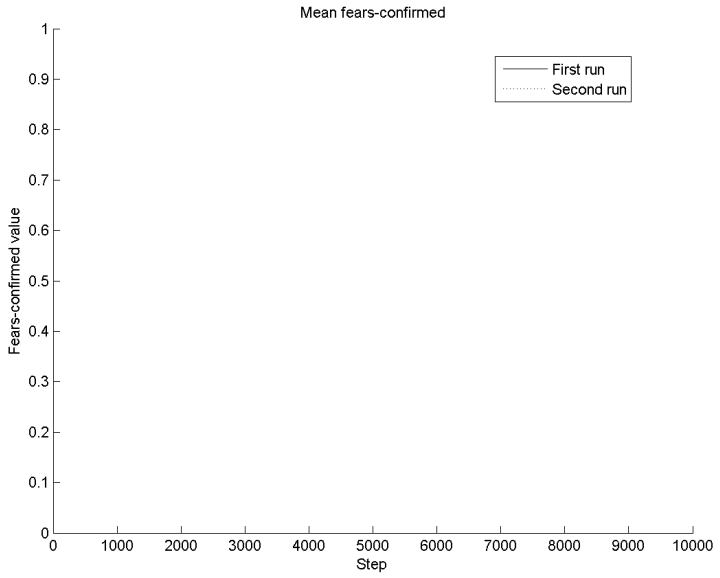

Figure 6-24: Intensity of fear-confirmation, mean over 50 agents, with a positive (first run) and negative (second run) reward

Chapter 7

Conclusion and discussion

We tested several hypotheses through a simulation in a maze. The development of emotions was shown to behave equal to humans, as expected based on our mapping. It has been constructed such that hope can not exist before having experienced joy/distress at some point, which is demonstrated in the simulations. At the same time, we showed the habituation of joy/distress in the control experiment, where the intensity of that emotion quickly diminished for a single agent. The speed at which the habituation takes place is very high as a result of the use of unexpectedness based on the model. Unexpectedness becomes very low at the beginning because the transition model is based on only a few experiences. This differs from humans, who in general have certain expectations about an event before ever having experienced it, mostly based on the experience of other people. We can simulate such experiences by starting with some model of the environment, which results in the initial intensity of joy/distress being much lower as the result of unexpectedness. Changes to the transition model are also less extreme, meaning the speed at which joy/distress habituates will be lower.

We also observed that the habituation of joy/distress does not go completely to 0. The reason for this is the discount factor, which ensures that expectations are always smaller than the actual outcome. This translates to humans that still experience joy, even when getting a reward that they have often received. Habituation mainly concerns a diminishing intensity and not a complete loss of emotion, so we consider this a valid result. Another mechanic for habituation was shown in the experiment where the reward was located. The lack of habituation observed in this simulation reflects the constant surprise experienced when getting rewards at unexpected momentes and fits theory on emotion.

By varying the learning type from off- to on-policy, which is similar to optimistic and realistic human, we were able to show that optimism leads to increase hope while realism increases happiness since expectations are lower. Contrary to our hypothesis, satisfaction was lower for realistic agents, which we attributed to the lower values as well.

Making an agent behave more randomly by manipulation action-selection decreased its happiness, but hope and satisfaction remained the same because of the simplicity of the setting. In a world that is harder to learn, we can expect a more striking results as inefficiency can be severely punished.

We also studied the effects of the discount factor by changing it from high to low, which we compared to long-term and short-term thinking humans. It showed that short-term thinkers have lower hope but reach a higher value of intensity for individual collected rewards. For the long-term thinkers, satisfaction is higher, a typical result which is indeed comparable to long- and short-term thinking humans.

The habituation of fear was demonstrated in the experiment with a collision penalty. Here, fear habituates to 0 once the agent learns that every state has some action that on average does not result in a penalty. The use of the maximum in the update function equates to a very optimistic human, who assumes complete control over his actions since the evaluation of a state is based on the best possible action. The lack of fear shows that, according to the simulation, these humans are not afraid even if an action exists with a much larger penalty than the possible gain. We could set the collision penalty to -10000 and find that the system habituates even faster because the Boltzmann action selection results in a strong bias toward not selecting an action leading to collision. However, a setting where negative values are that high may be expected to at least cause some fear. In the current simulation settings, the agent does not take into account the possible penalty because a better action exists. This allows the fear to habituate to 0.

In the same simulation, hope did not change when adding a collision penalty. This can also be attributed to the use of the maximum, since bad actions are simply not taken into account. We were not able to validate Hypothesis 6 because we wrongfully assumed that the solution method would take bad actions into account, which would result in lower hope and higher joy intensity. This is another argument for making the update function more humanlike and shows that the utmost care should be taken when constructing a hypothesis. All settings for the parameters of the simulation should be considered with respect to their counterparts in the human world to ensure a correct hypothesis.

The unpredictability of the world demonstrate the expected effects on the intensity of joy/distress. Both simulations introducing some form of unpredictability increased that intensity, showing that surprise is an important factor in the strength of this emotion, as theorized by many psychologists.

Changing the reward to negative did not create many interesting results, since the agent simply learned to walk around states away from that reward. As expected, this simulation showed a bit of fear and the habituation of it as a result of the agent learning to act so that no penalty resulted anymore.

We can conclude that the mapping we created from Reinforcement Learning elements to emotions shares, both mathematically and in simulations, many traits with theory on human emotions. Joy/distress, hope/fear and confirmation can be demonstrated to develop correctly. Habituation was shown for both joy and fear. Manipulating elements of the learning algorithm had the expected effects in a large number of cases. Expectations for a setting with a collision penalty were shown to be incorrect, but were based on the wrong assumption. The intensity of joy was coupled to unpredicatibility in two different settings. This makes our mapping both accurate and useful for studying the given emotions, thus providing a solid base for mapping a larger number of emotions.

7-1 Using emotions to drive action-selection

The development and occurence of emotions resulting from our appraisal model is comparable to human learning. This result is satisfying and an interesting application is the use of emotions as a drive for action-selection, as was done earlier [6, 4, 5]. Using the emotions as a means of action-selection may change the results, as the emotions become part of the feedback loop. If we want to argue that the development remains the same, we need to demonstrate that picking the action resulting in the preferred emotion is similar to the currently used action-selection method based on the Boltzmann distribution. This distribution can also be used to pick the action resulting in the best emotion:

$$\frac{e^{\beta E_t(a)}}{\sum_{b=1}^{n} e^{\beta E_t(b)}} \tag{7-1}$$

Where E is the relevant emotion attributed to an action (or state-action combination). The relevant emotion is the emotion with the strongest (absolute) intensity. The intensity of confirmation emotions does not play a role here, since they can only be determined after the result of an action is determined. Relevance is therefore determined by comparing the expected desirability, hope and fear values resulting from a certain action. For example, an action that has a very high chance of resulting in a state with desirability 3 and a fear value of -5 has fear as its relevant emotion.

Using state-values, the expected intensities of hope and fear for a specific action a are:

$$H_V(a_t) = \max_{s_{t+1}} P^{a_t}_{s_t s_{t+1}} V(s_{t+1}) \tag{7-2}$$

$$F_V(a_t) = \min_{s_{t+1}} P^{a_t}_{s_t s_{t+1}} V(s_{t+1}) \tag{7-3}$$

$$\tag{7-4}$$

The desirability attributed to a specific action is

$$D_V(a_t) = \sum_{s_{t+1}} P^{a_t}_{s_t s_{t+1}} (V(s_{t+1}) - V(s_t) + R^{a_t}_{s_t s_{t+1}}) \tag{7-5}$$

Since $V(s)$ is fixed, the distribution of $D_V(a_t)$ accross possible actions a_t depends solely on the term $V(s_{t+1}) + R^{a_t}_{s_t s_{t+1}}$, which is precisely used in the state-value determination of the Boltzmann distribution. Hope or fear are only relevant if they have a higher intensity than the desirability, meaning that the prospect emotions provide a (positive or negative) bias towards extreme valued states s_{t+1}. We expect this added information to be reflected in a less explorative behaviour caused by taking into account the extremes in that extra state. Using normal action-selection also takes into account these extremes, but one iteration later since they are processed in the update of the state-value $V(s)$.

The exact distribution resulting from using emotions as the action-selection method can not be predicted, but based on this discussion we expect it to be approximately equal to the

Boltzmann distribution using state-values. However, the distribution when using emotions is very specific for the type of simulation, as different environments can vary greatly in emotion intensity. This is an unfortunate consequence of the need to pick a relevant emotion out of the available emotions and suggests the need to look for a more appropriate action-selection method. We consider such a task to be beyond the scope of this thesis and leave it for future work.

7-2 Recommendations for future research

The research described in this thesis only contains the first steps toward a complete emotion mapping. As results were promising, this opens up the path to more complicated scenarios and research on other emotions.

Some discussion on using the emotions as part of the action-selection was already performed. Some mathematical operations showed that the Boltzmann distribution using relevant emotions should be similar, but no definitive conclusion can be drawn yet. The concept of relevant emotions is too scenario-specific and future research would benefit from finding a more broadly usable distribution.

A lot of emotions come relate to the presence of other agents. Simulations in a multi-agent scenario would certainly add to the range of possible emotions, allowing a broader range of them to be tested. In fact, emotions aimed at objects or other agents make up a large part of the complete set of emotions and determine social interaction. Such emotions may be used to simulate specific forms of group behaviour, allowing examination of group dynamics. Other emotions related to multi-agent systems include anger, gratitude and shame.

We mentioned the human hormone model with respect to learning earlier. With the availability of emotions through the mapping, we can find the resulting emission of specific hormones. If we connect these hormones to specific Reinforcement Learning (RL) variables, it would in theory be possible to create a dynamic learning system that determines its emotion in each step, which is then used to alter the learning parameters. This is an interesting option for further research, possibly making the learning agents more human. A comparison can then also be performed by doing the same experiments with live animals to verify hypotheses.

More scenarios may be thought of to examine specific emotions or settings. For instance, a bias for a given area may be defined by initiating its values negatively. Such a setup makes the difference between exploring and exploiting agents much more pronounced. Another possibility is an addictive scenario, where the benefits of an easily reachable reward decrease quickly, forcing the agent to collect the reward more and more often to pass a certain happiness threshold. This could provide a deeper insight on the mechanisms behind addiction.

Using the examined emotion model provides a lot of options for researching emotion development. It may prove to be a large step towards understanding more of human emotions, which is an interesting result indeed.

Appendix A

Software specifics

A-1 Pseudocode

A-1-1 Pseudocode for off-policy Value Iteration

Initialize V such that $V(s) = 0$ for all s
$s = (0, 0)$
repeat
  $a \leftarrow$ action given by π for s
  Take action a; find reward r and next state s'
  Find the unexpectedness of the transition (s, a, s')
  Update the transition model
  Set $s \leftarrow s'$
  Find emotion values
  $V(s) = \max_a \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V_k(s') \right].$
until all steps taken

A-1-2 Pseudocode for on-policy Value Iteration

Initialize V such that $V(s) = 0$ for all s
$s = (0, 0)$
$a \leftarrow$ action given by π for s
repeat
  Take action a; find reward r and next state s'
  Find the unexpectedness of the transition (s, a, s')
  Update the transition model
  Set $s \leftarrow s'$
  Find emotion values
  $a \leftarrow$ action given by π for s
  $V(s) = \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V_k(s') \right].$

until all steps taken

A-2 Class diagram

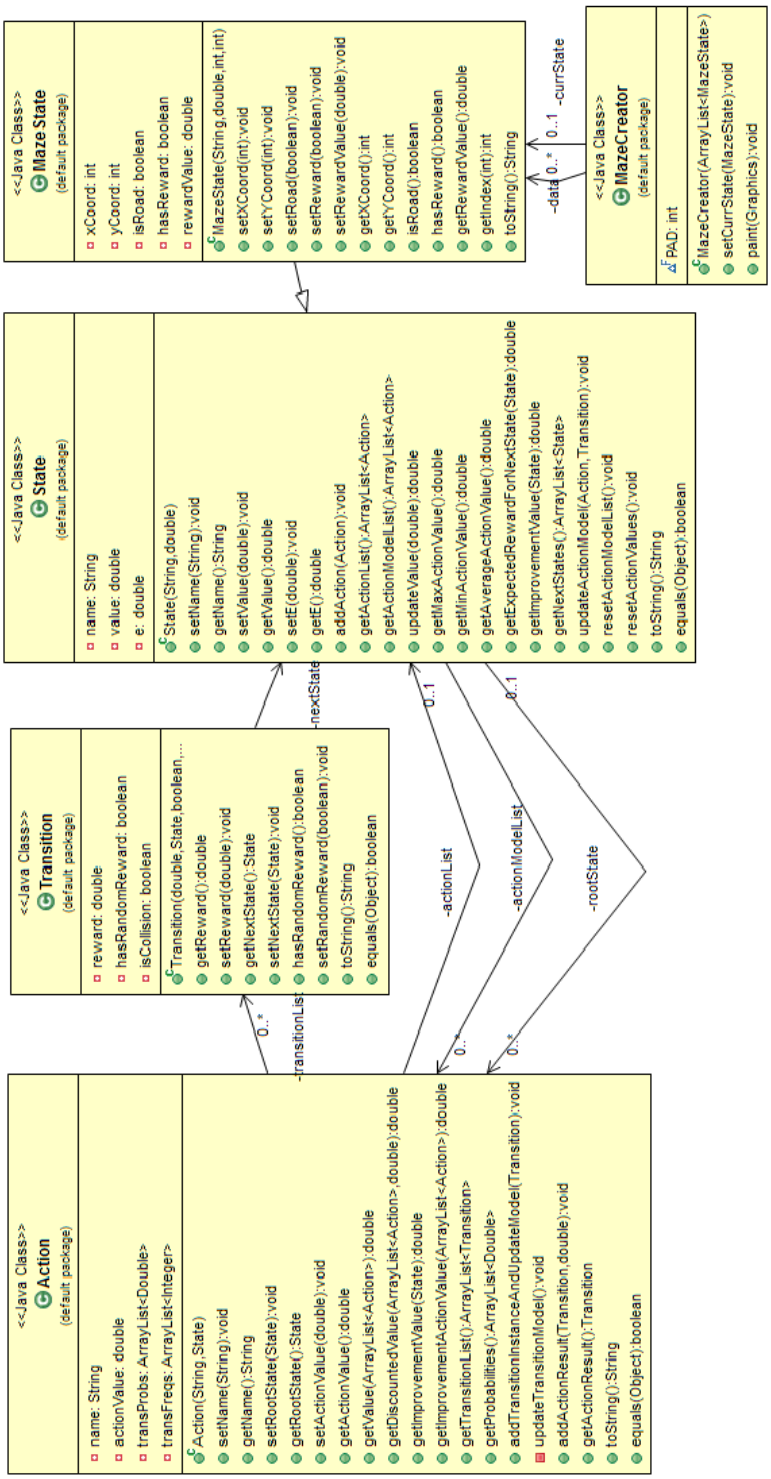

Figure A-1: Class diagram of the program, part 1

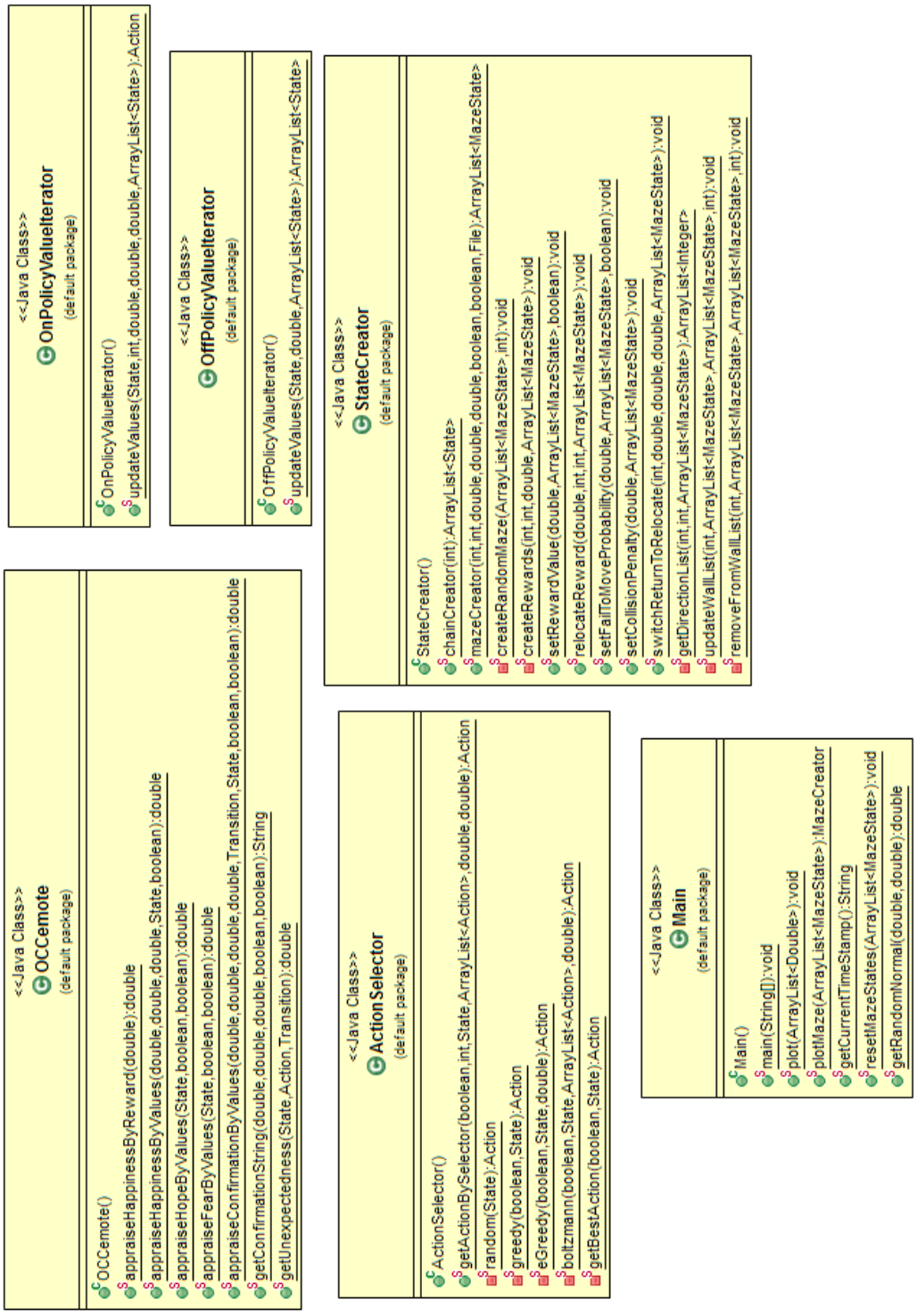

Figure A-2: Class diagram of the program, part 2

Appendix B

Complete experimental results

B-1 Off-policy vs on-policy learning

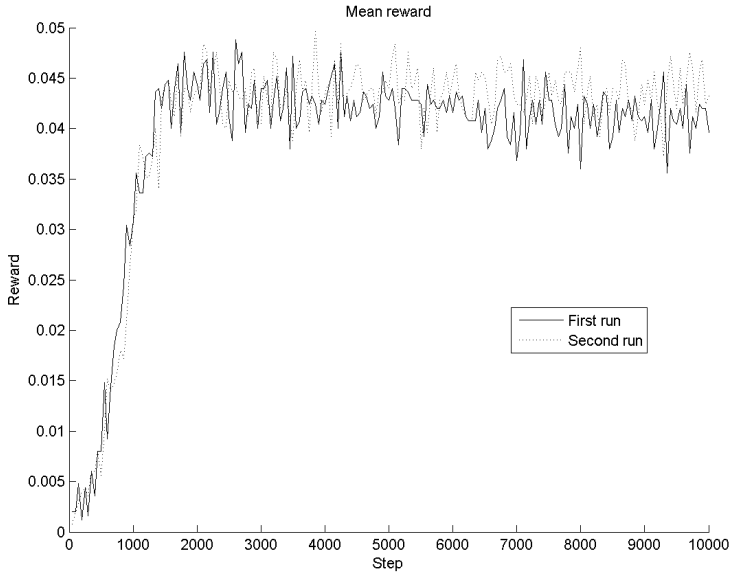

Figure B-1: Reward value for off-policy and on-policy learning

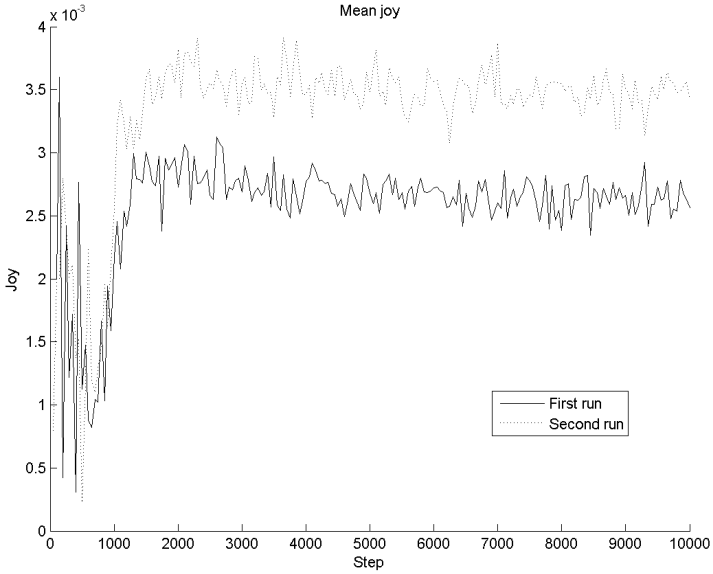

Figure B-2: Happiness value for off-policy and on-policy learning

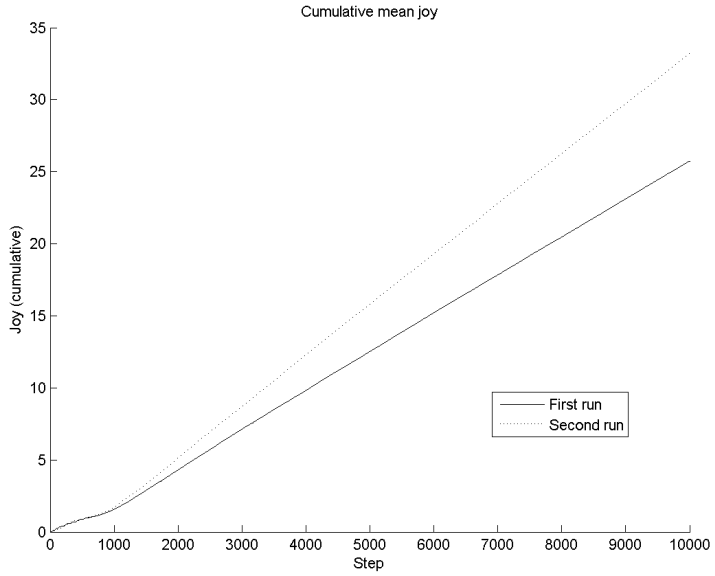

Figure B-3: Cumulative happiness value for off-policy and on-policy learning

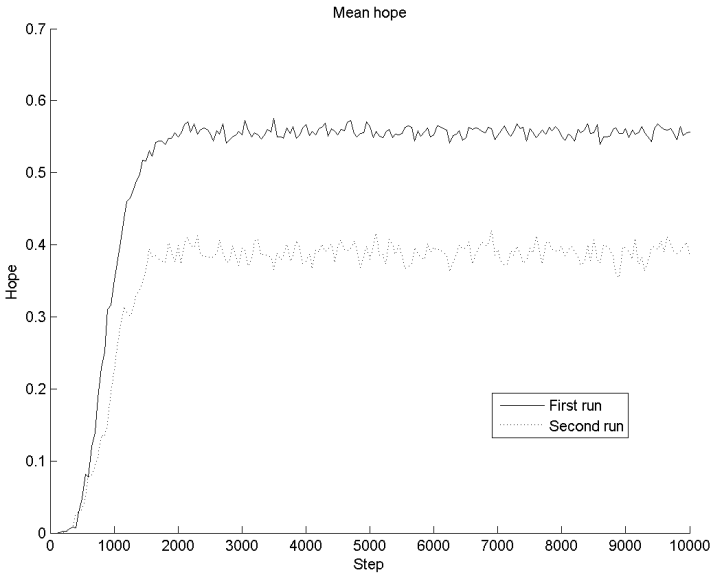

Figure B-4: Cumulative hope value for off-policy and on-policy learning

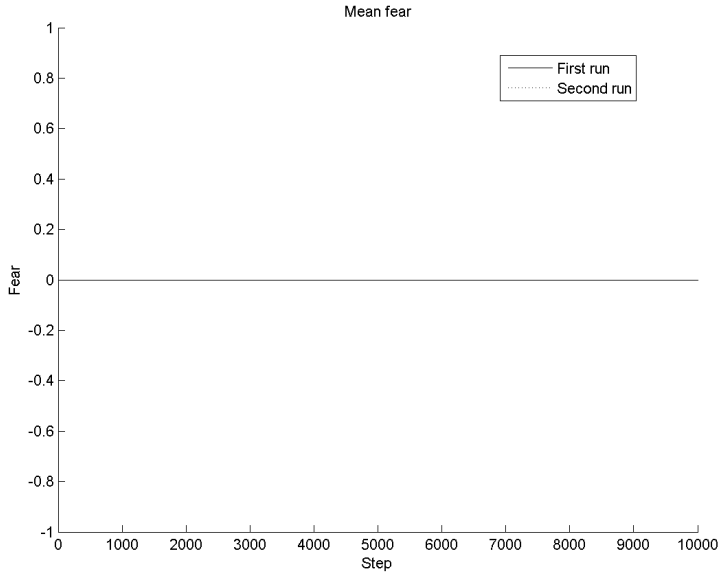

Figure B-5: Cumulative fear value for off-policy and on-policy learning

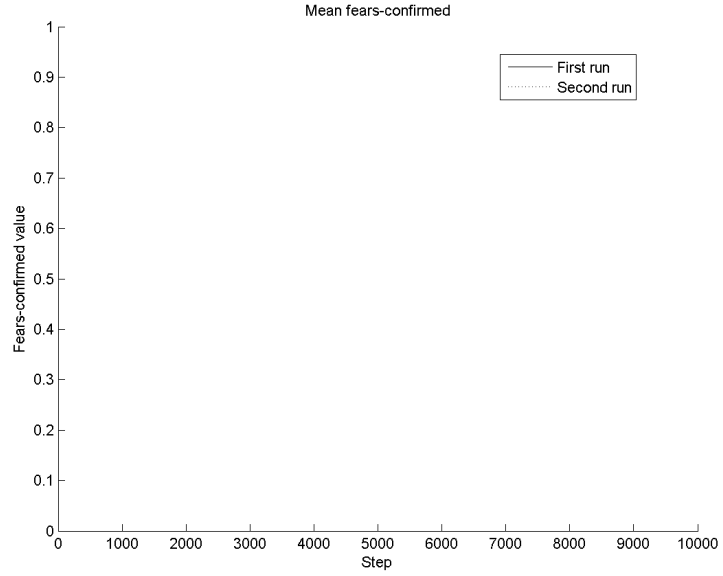

Figure B-6: Fears-confirmed value for off-policy and on-policy learning

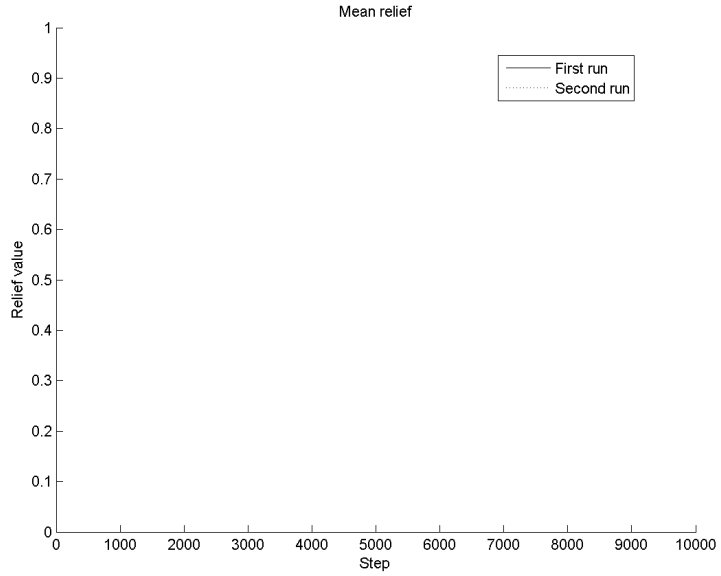

Figure B-7: Relief value for off-policy and on-policy learning

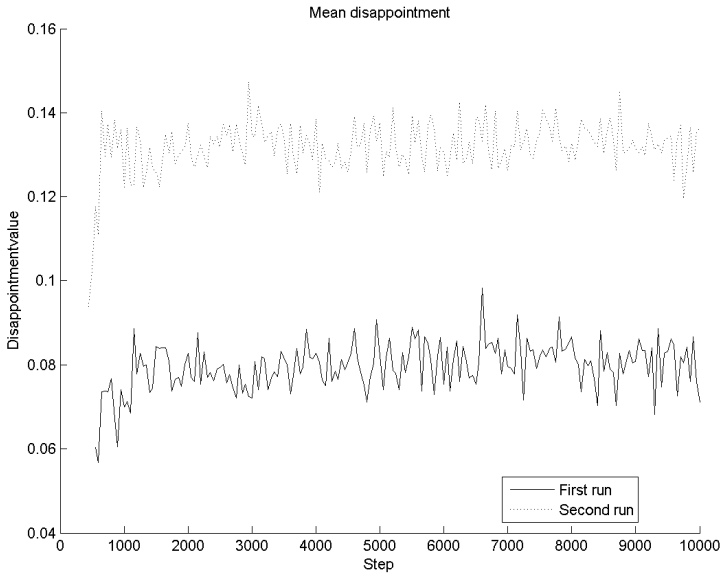

Figure B-8: Disappointment value for off-policy and on-policy learning

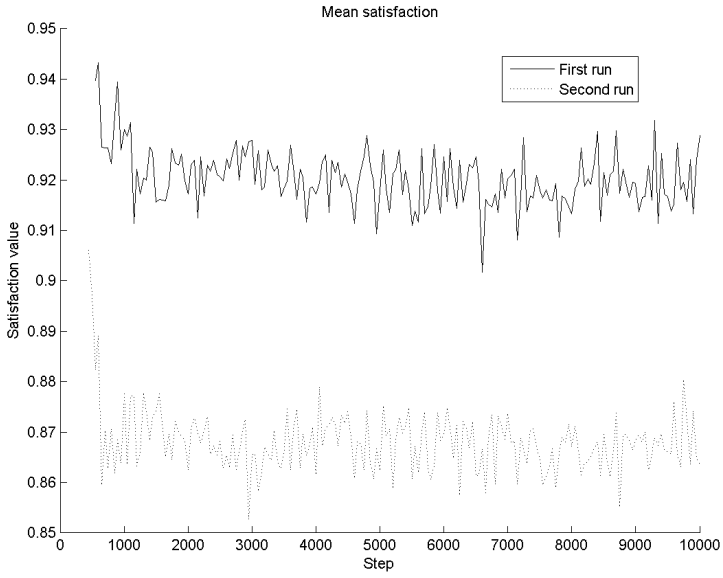

Figure B-9: Satisfaction value for off-policy and on-policy learning

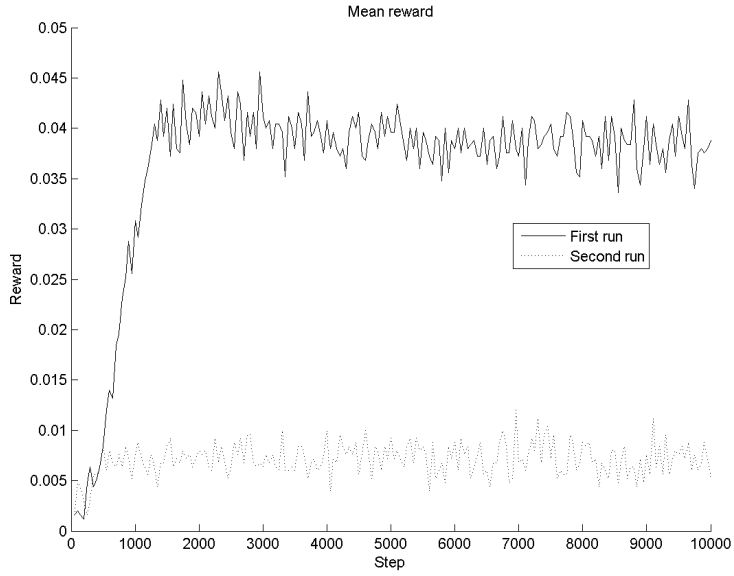

Figure B-10: Reward value for two different inverse temperatures

B-2 Manipulating action-selection

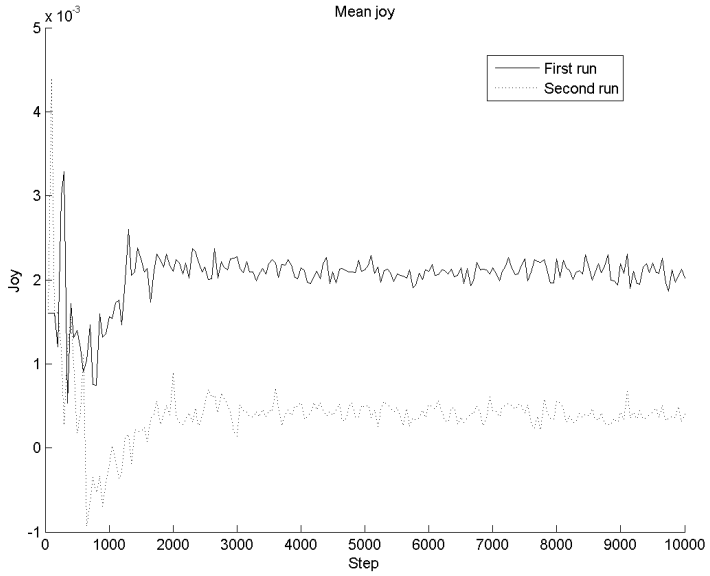

Figure B-11: Happiness value for two different inverse temperatures

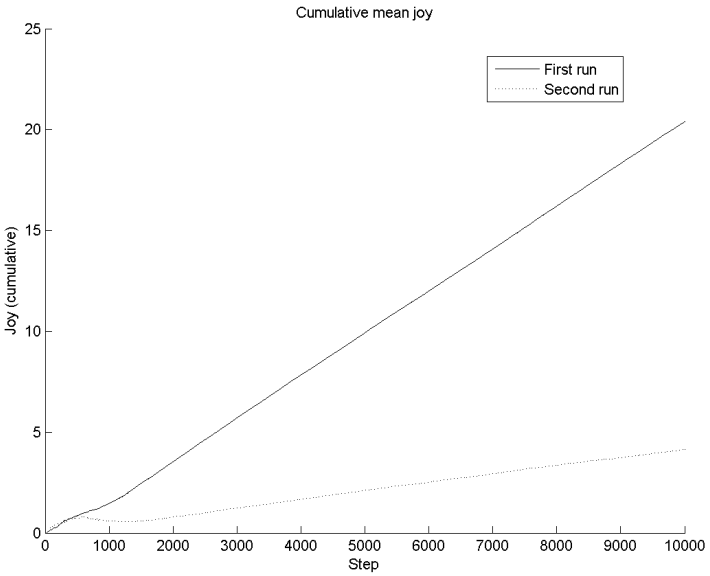

Figure B-12: Cumulative happiness value for two different inverse temperatures

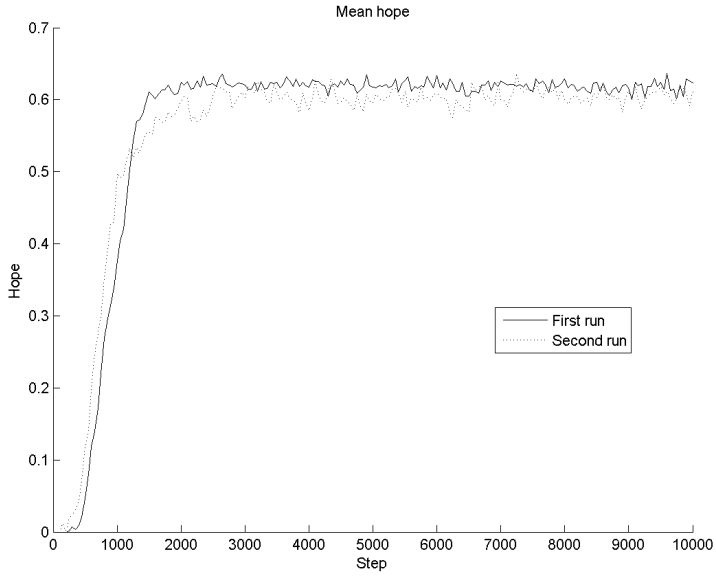

Figure B-13: Cumulative hope value for two different inverse temperatures

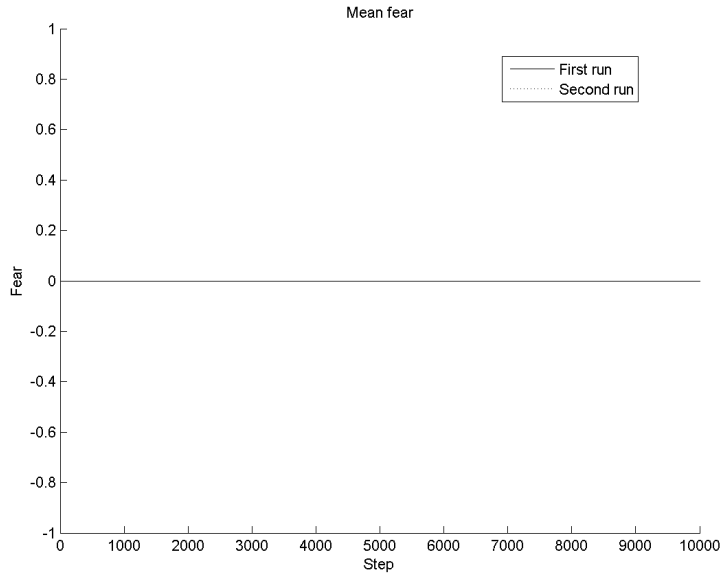

Figure B-14: Cumulative fear value for two different inverse temperatures

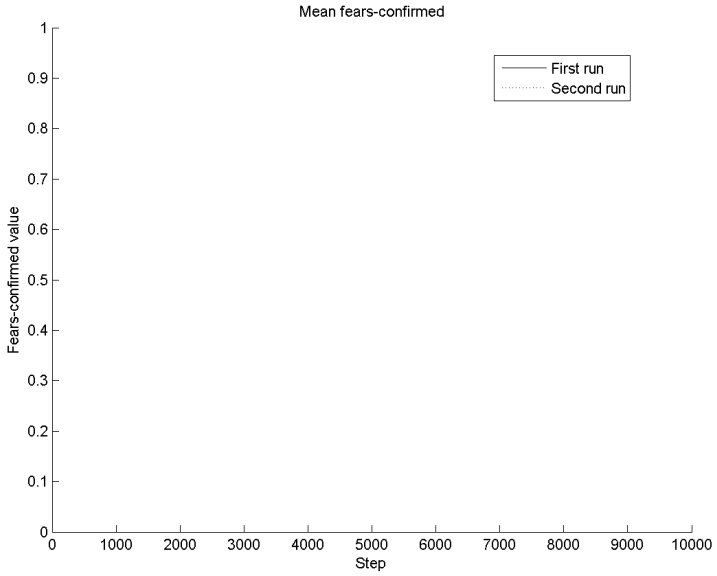

Figure B-15: Fears-confirmed value for two different inverse temperatures

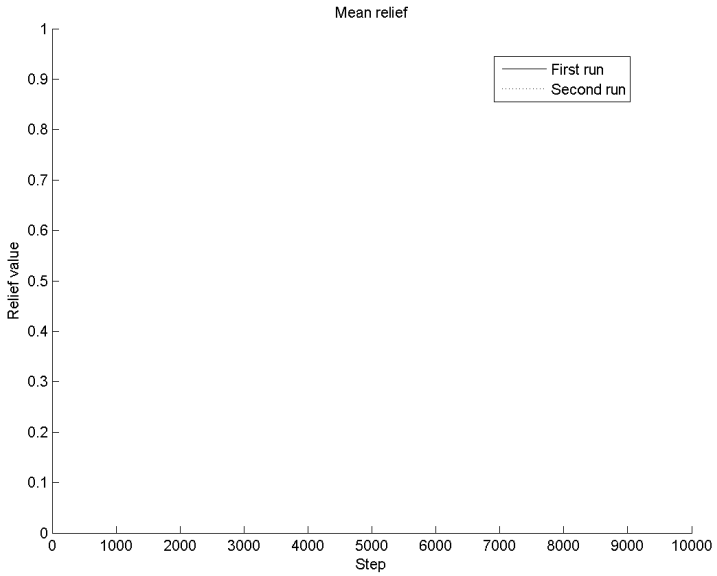

Figure B-16: Relief value for two different inverse temperatures

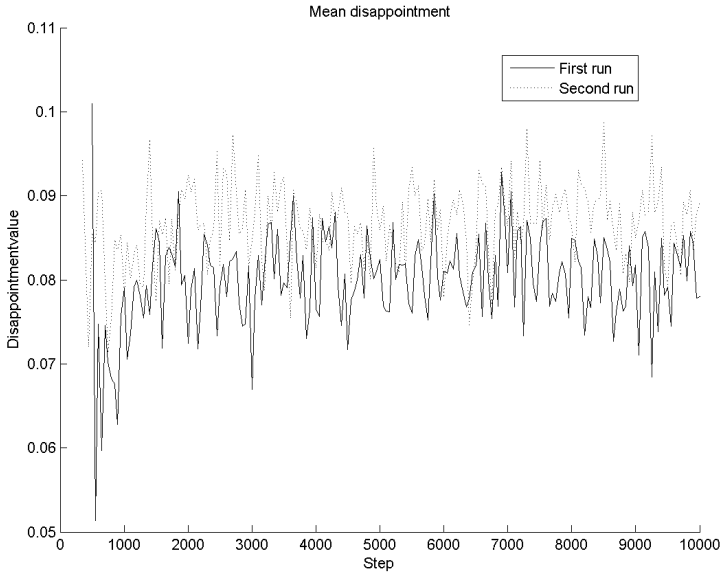

Figure B-17: Disappointment value for two different inverse temperatures

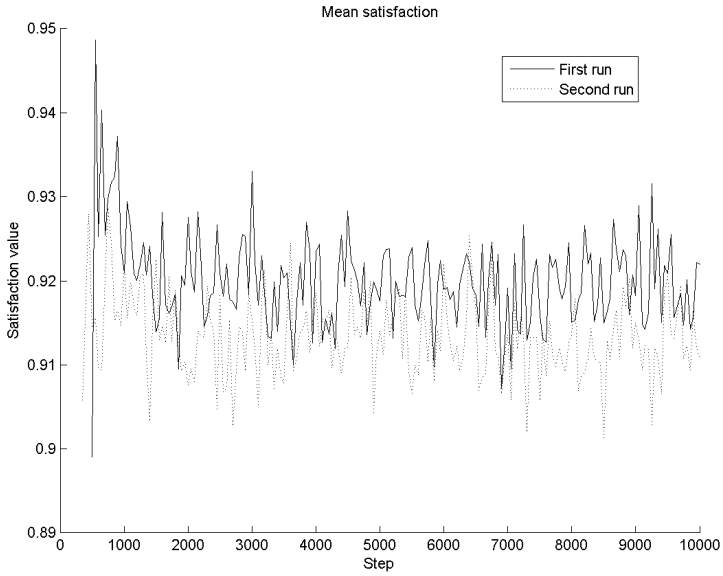

Figure B-18: Satisfaction value for two different inverse temperatures

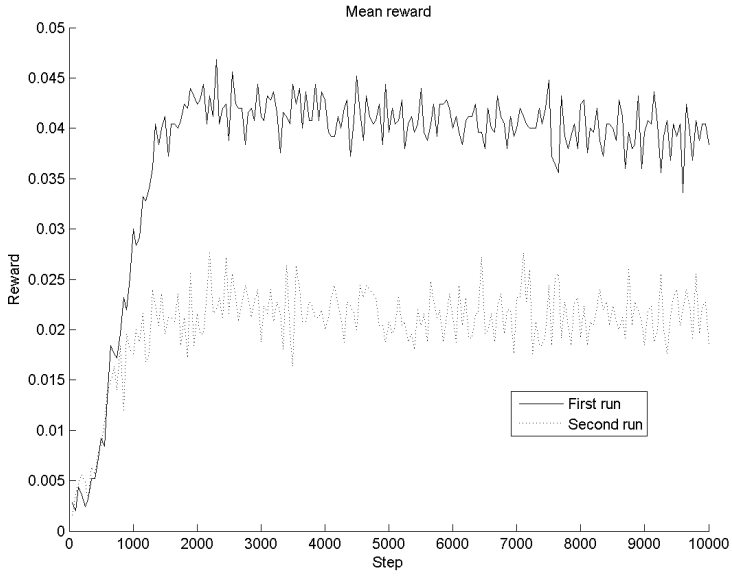

Figure B-19: Reward value for two different values of gamma

B-3 Changing the discount factor

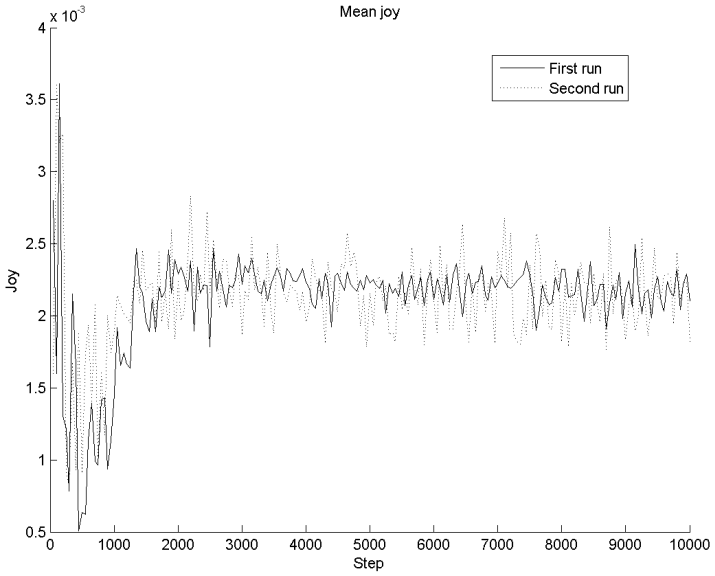

Figure B-20: Happiness value for two different values of gamma

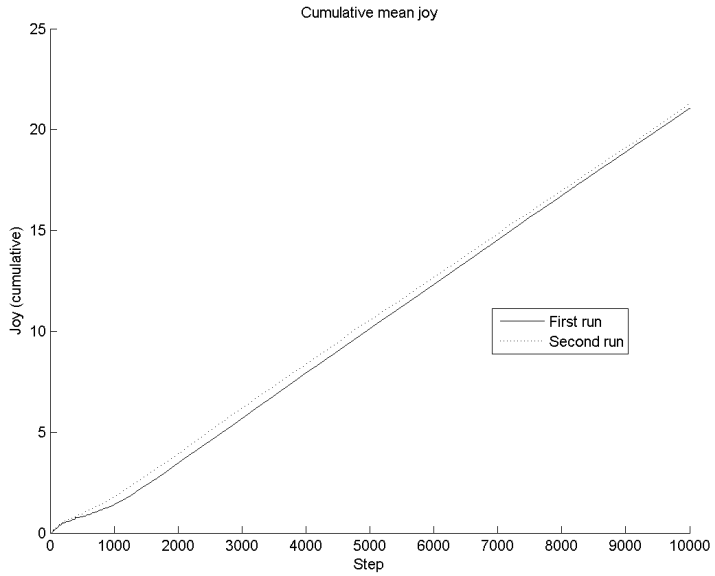

Figure B-21: Cumulative happiness value for two different values of gamma

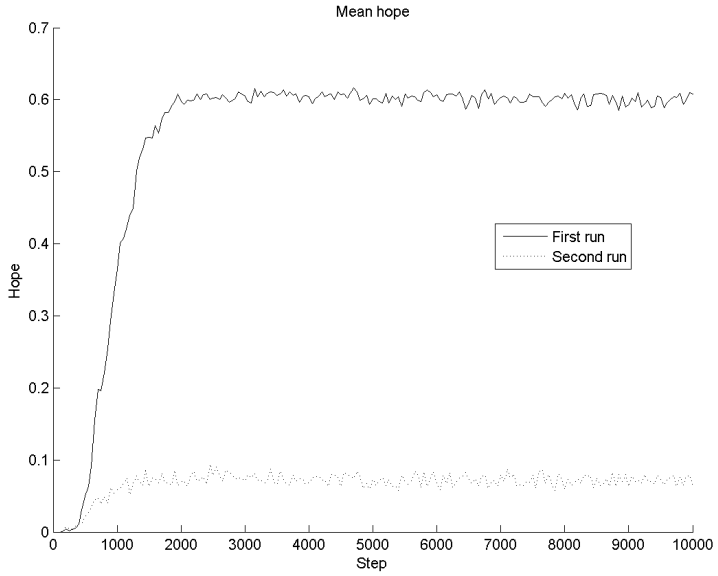

Figure B-22: Cumulative hope value for two different values of gamma

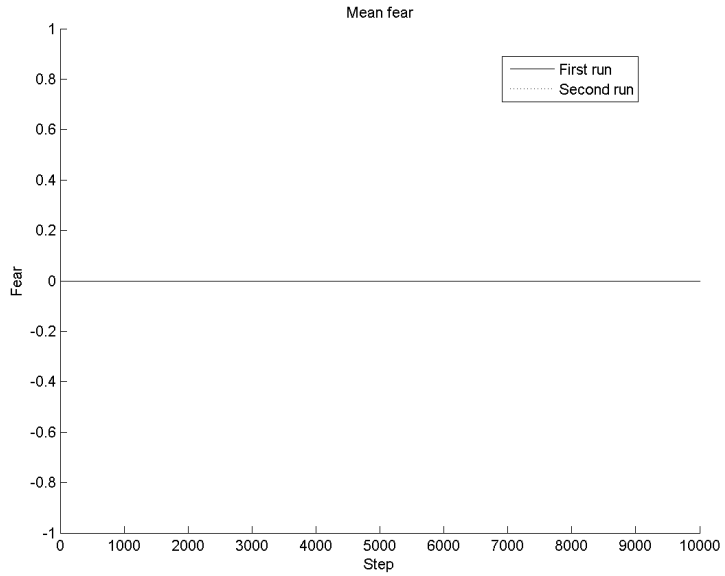

Figure B-23: Cumulative fear value for two different values of gamma

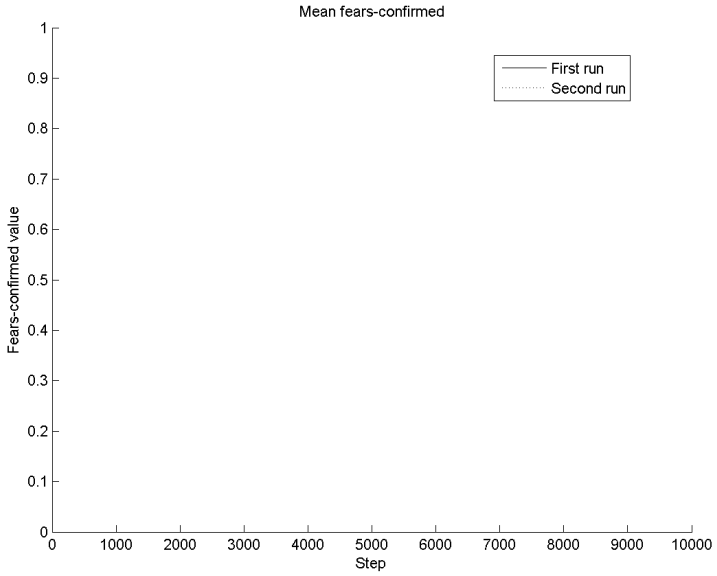

Figure B-24: Fears-confirmed value for two different values of gamma

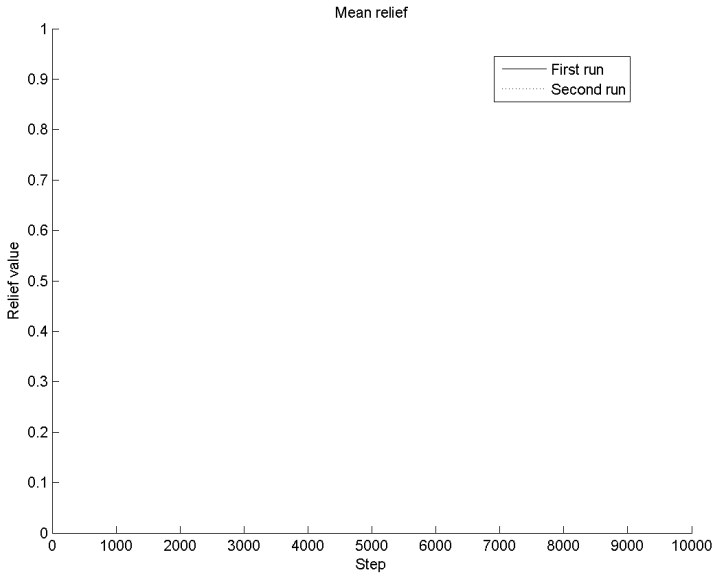

Figure B-25: Relief value for two different values of gamma

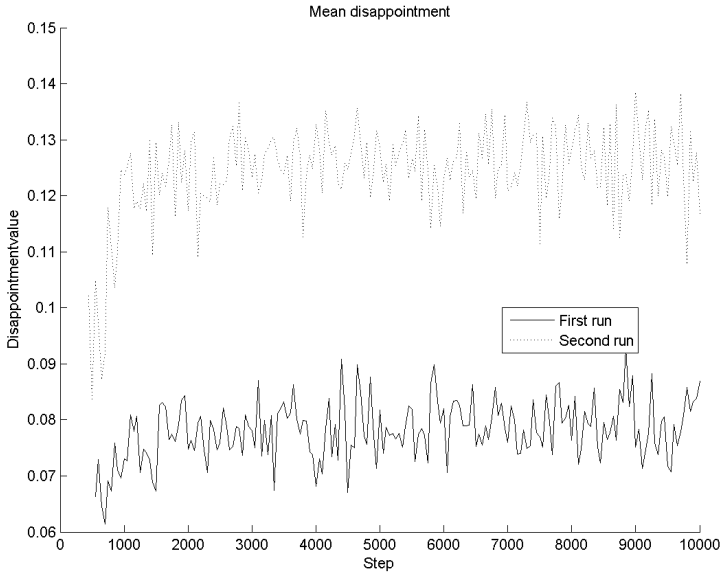

Figure B-26: Disappointment value for two different values of gamma

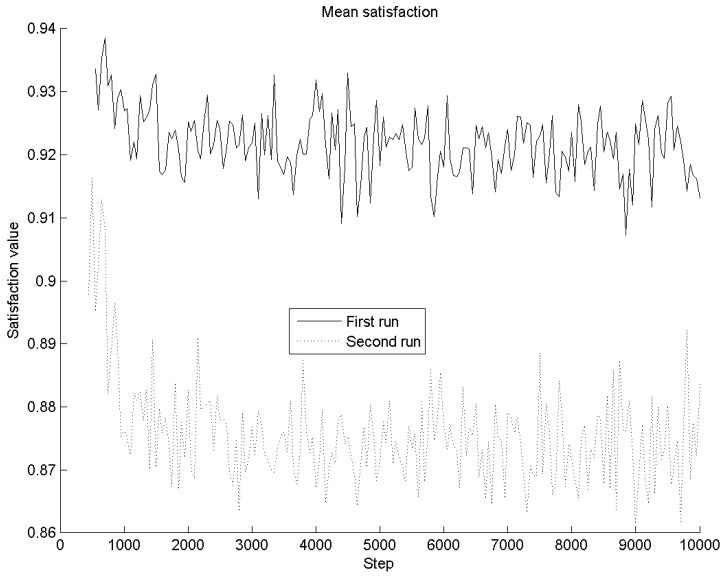

Figure B-27: Satisfaction value for two different values of gamma

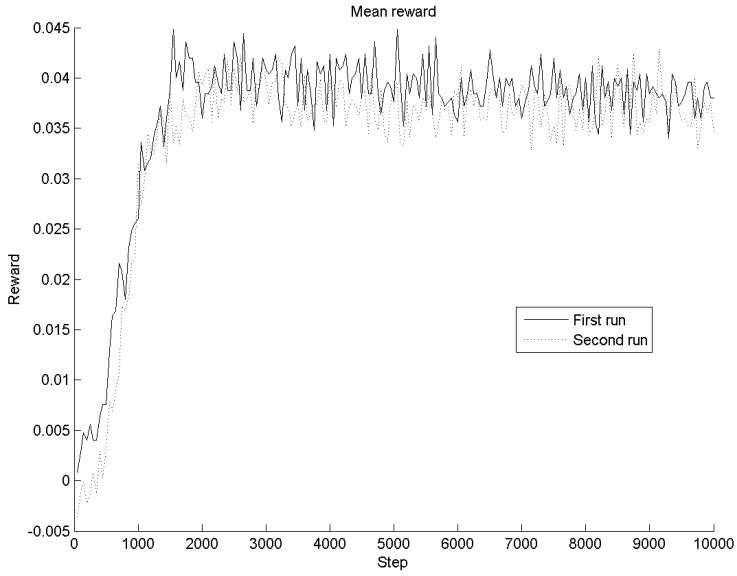

Figure B-28: Reward value without and with collision penalty

B-4 Adding a collision penalty

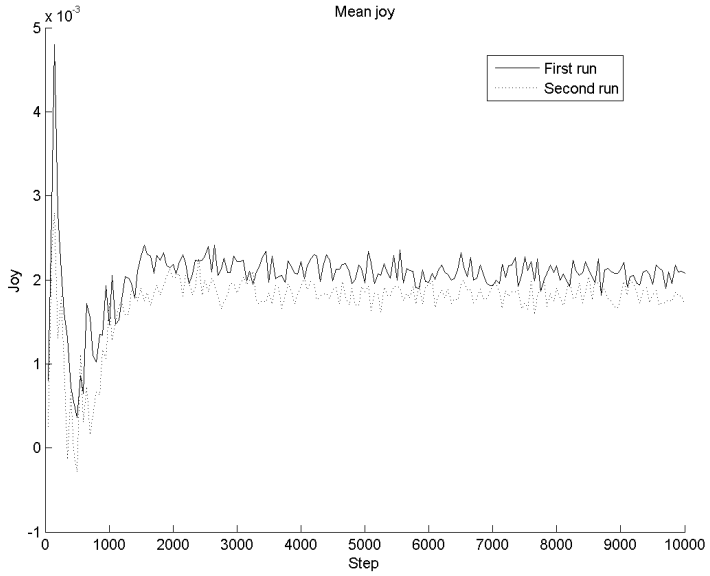

Figure B-29: Happiness value without and with collision penalty

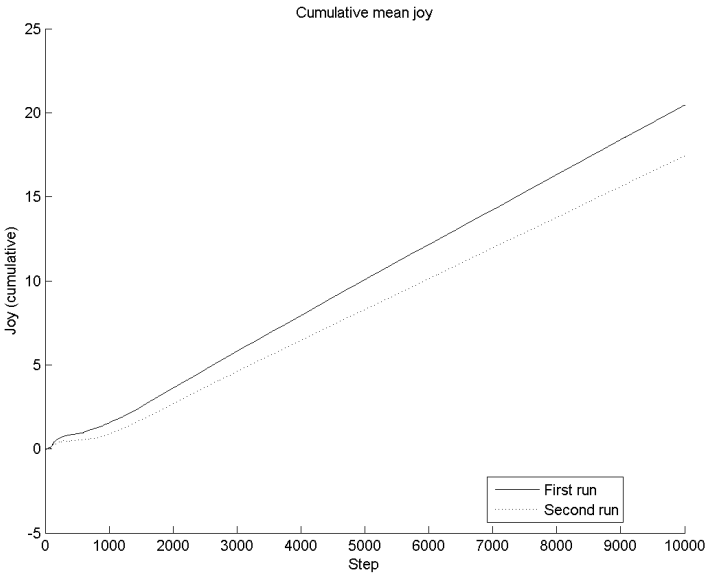

Figure B-30: Cumulative happiness value without and with collision penalty

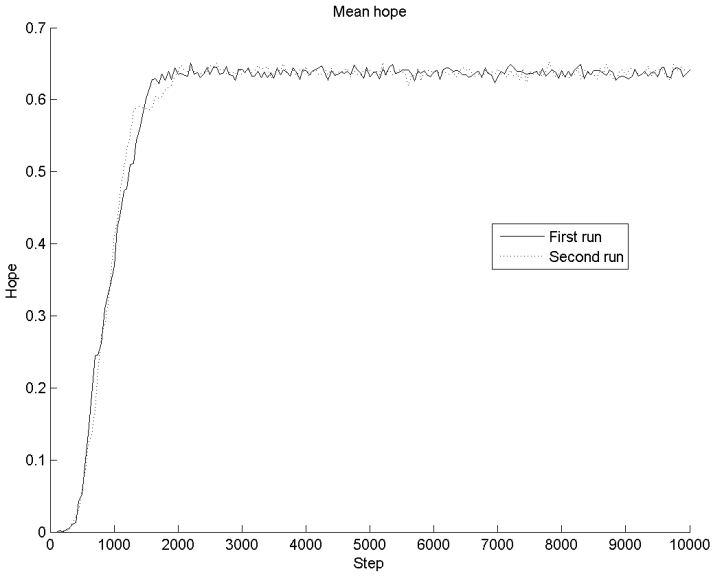

Figure B-31: Cumulative hope value without and with collision penalty

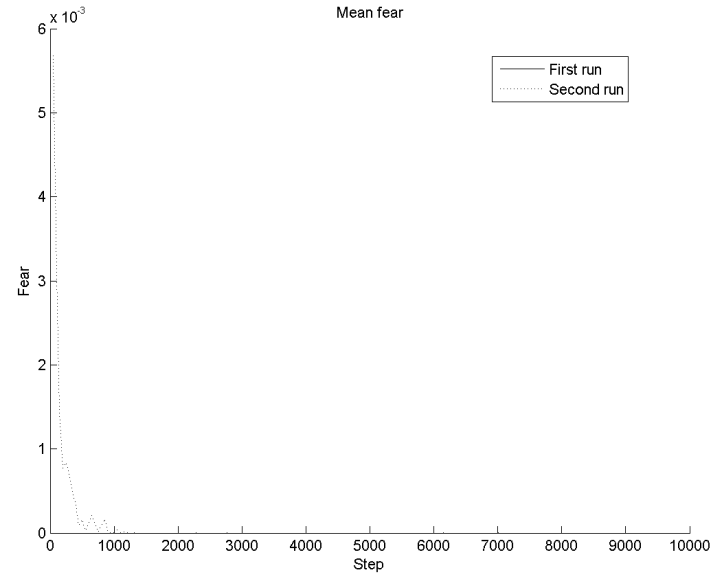

Figure B-32: Cumulative fear value without and with collision penalty

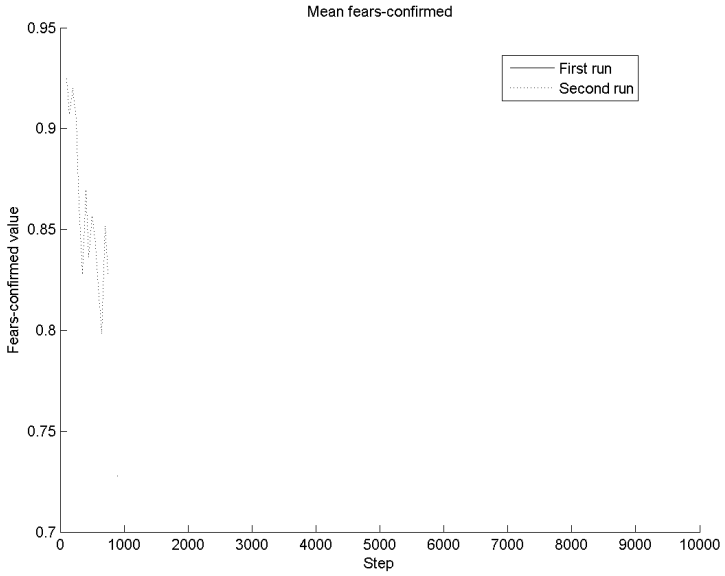

Figure B-33: Fears-confirmed value without and with collision penalty

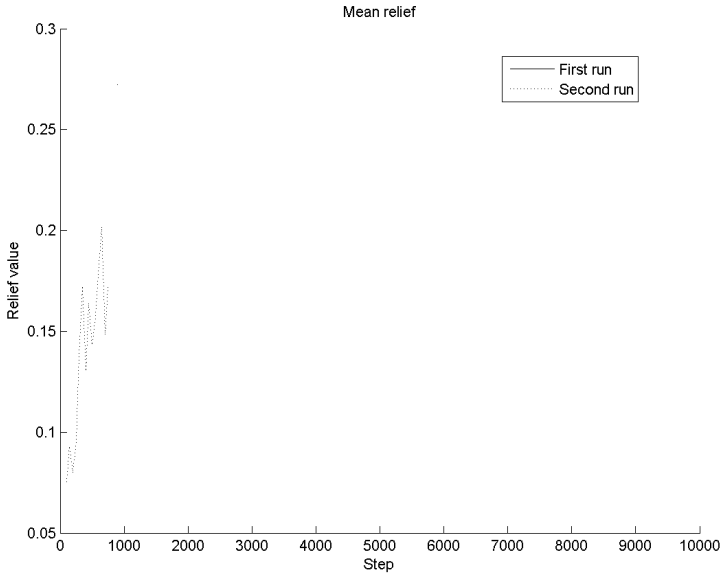

Figure B-34: Relief value without and with collision penalty

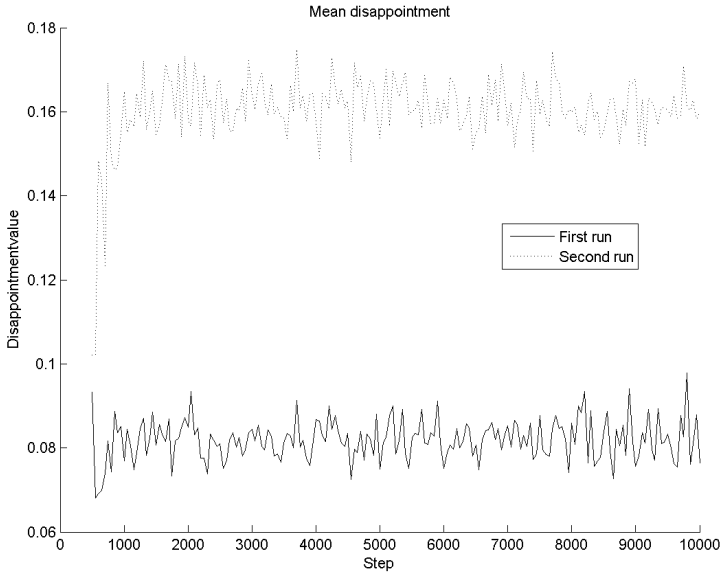

Figure B-35: Disappointment value without and with collision penalty

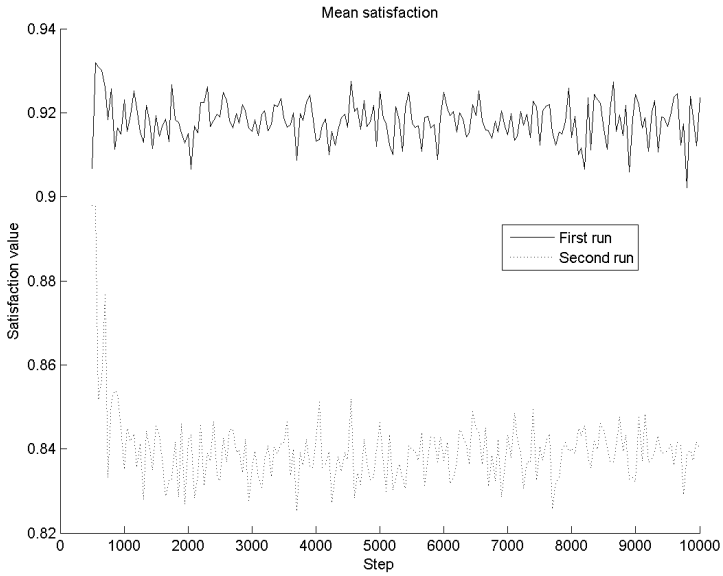

Figure B-36: Satisfaction value without and with collision penalty

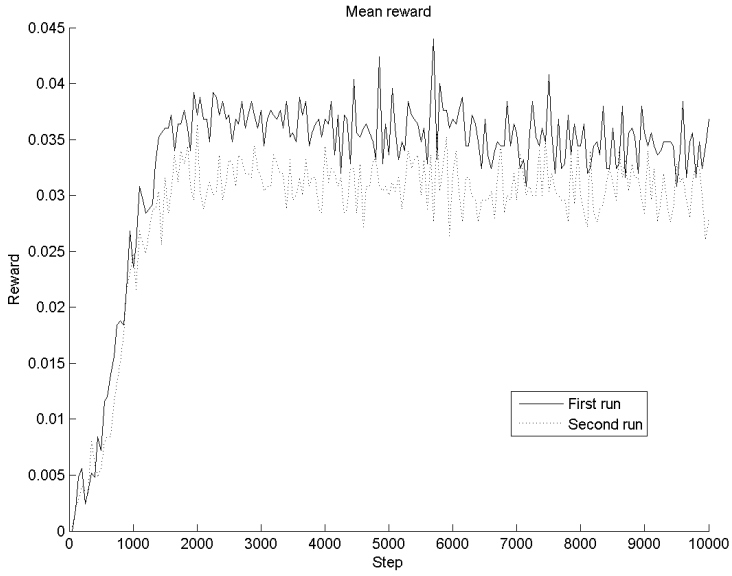

Figure B-37: Reward value in a stochastic and deterministic world

B-5 Deterministic or stochastic

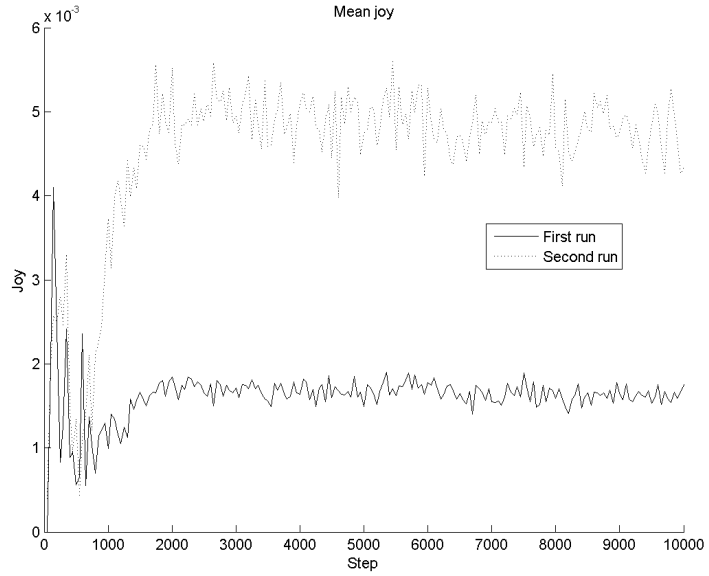

Figure B-38: Happiness value in a stochastic and deterministic world

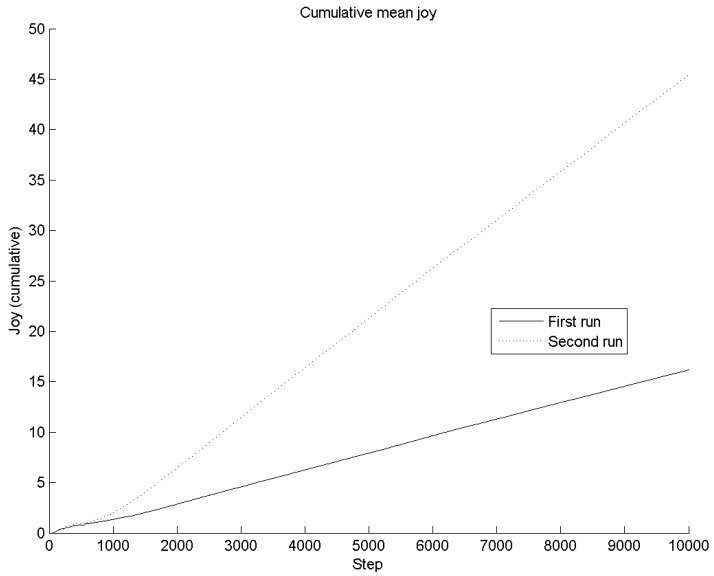

Figure B-39: Cumulative happiness value in a stochastic and deterministic world

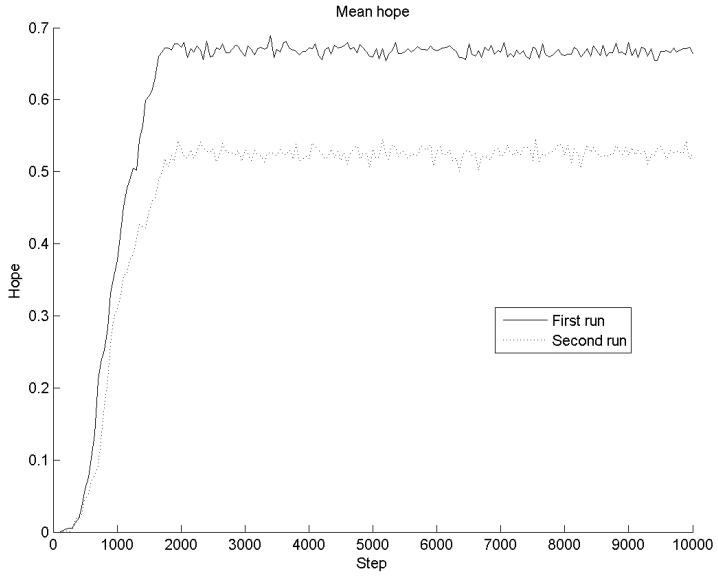

Figure B-40: Cumulative hope value in a stochastic and deterministic world

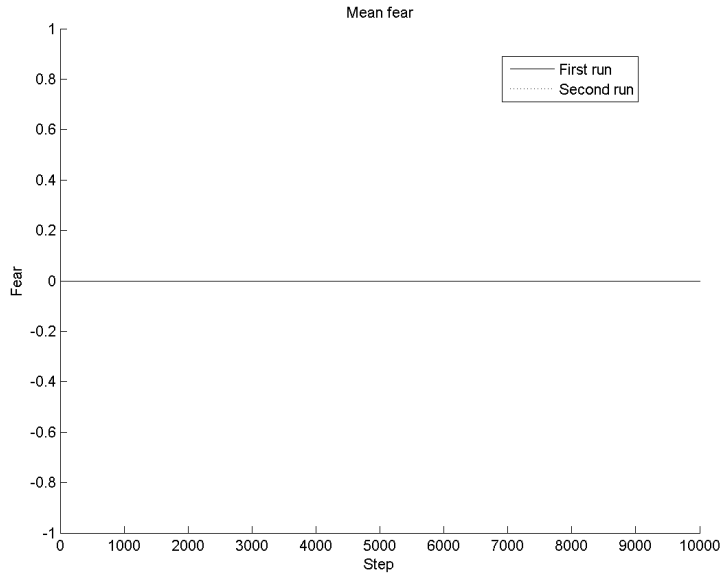

Figure B-41: Cumulative fear value in a stochastic and deterministic world

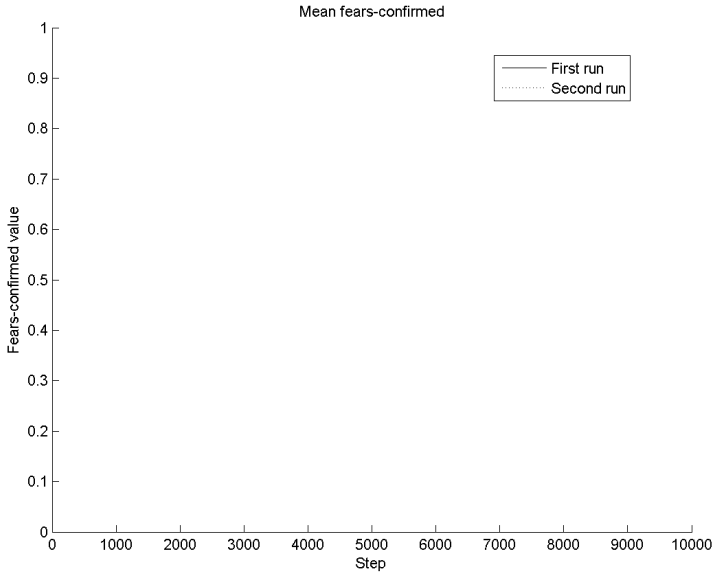

Figure B-42: Fears-confirmed value in a stochastic and deterministic world

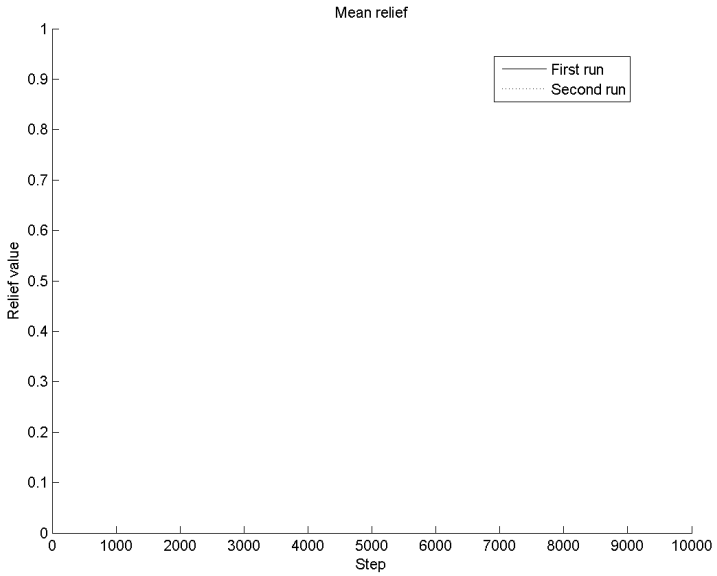

Figure B-43: Relief value in a stochastic and deterministic world

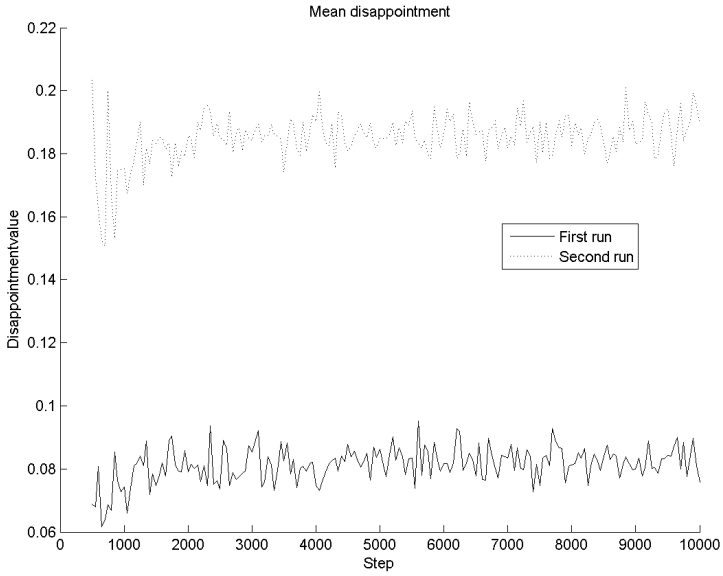

Figure B-44: Disappointment value in a stochastic and deterministic world

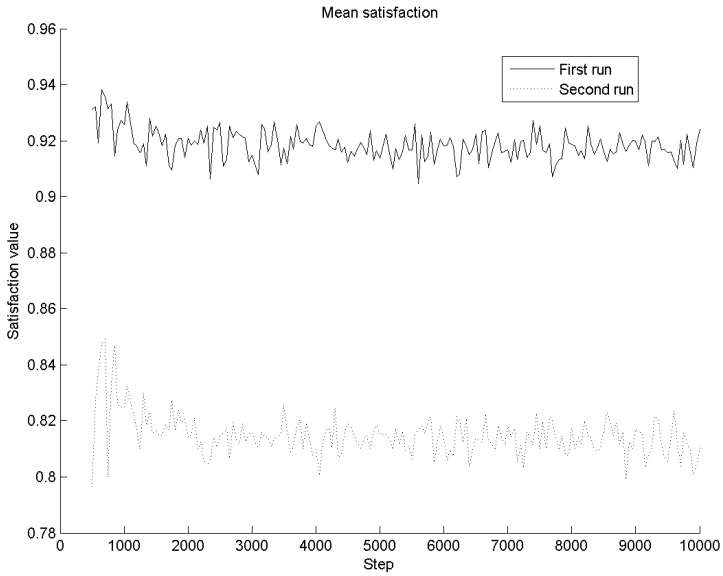

Figure B-45: Satisfaction value in a stochastic and deterministic world

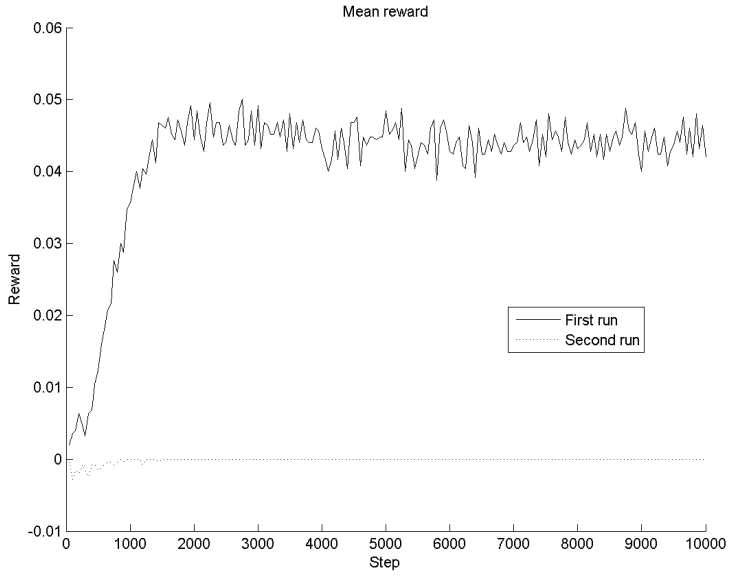

Figure B-46: Reward value for a positive and negative reward

B-6 Negative vs positive reward

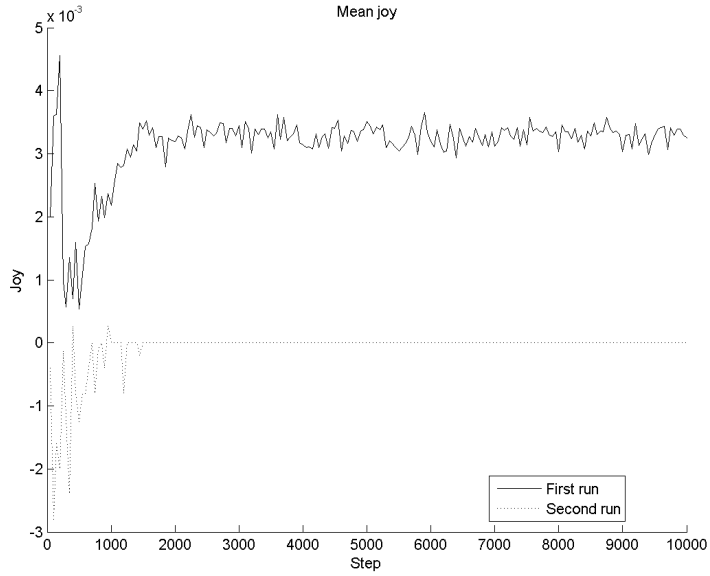

Figure B-47: Happiness value for a positive and negative reward

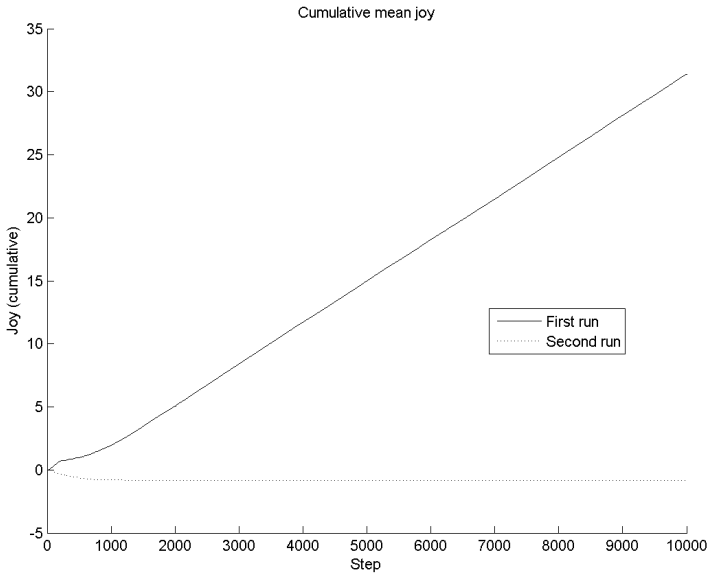

Figure B-48: Cumulative happiness value for a positive and negative reward

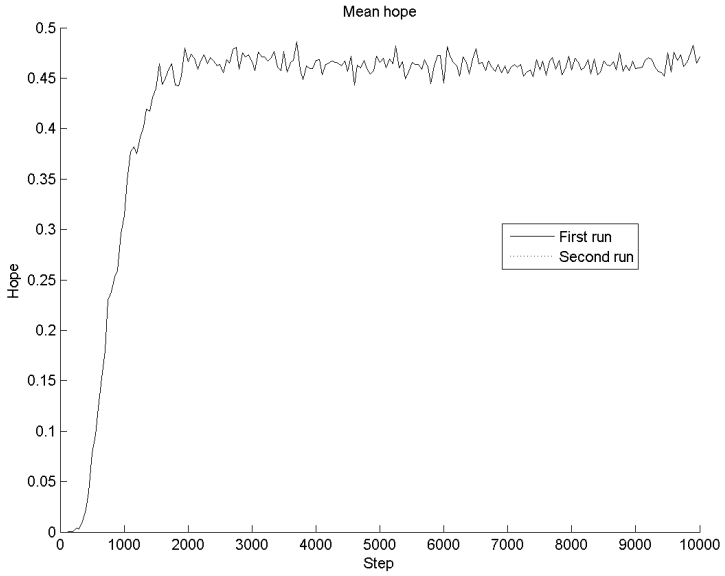

Figure B-49: Cumulative hope value for a positive and negative reward

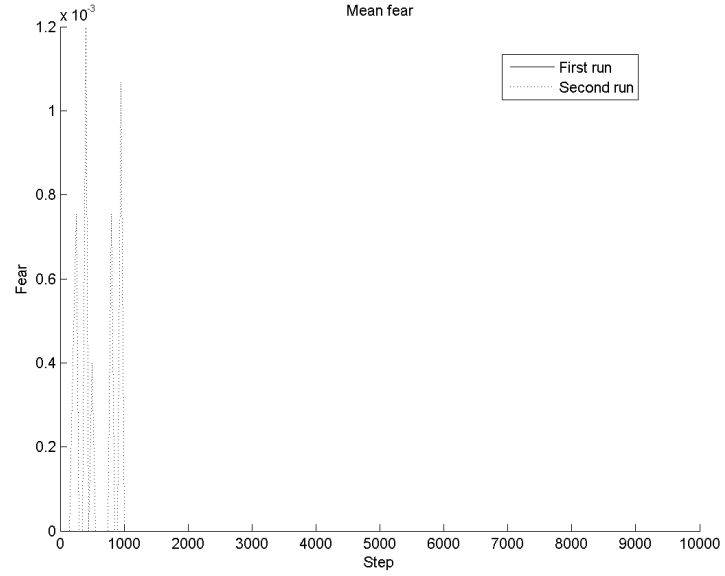

Figure B-50: Cumulative fear value for a positive and negative reward

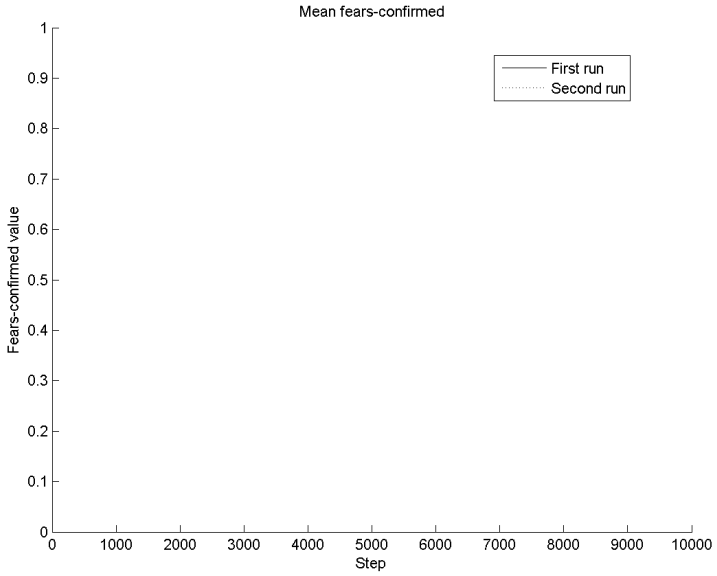

Figure B-51: Fears-confirmed value for a positive and negative reward

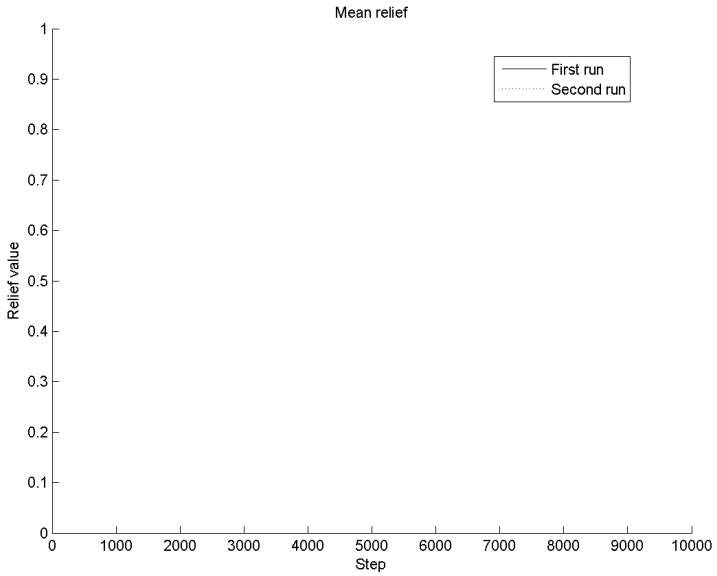

Figure B-52: Relief value for a positive and negative reward

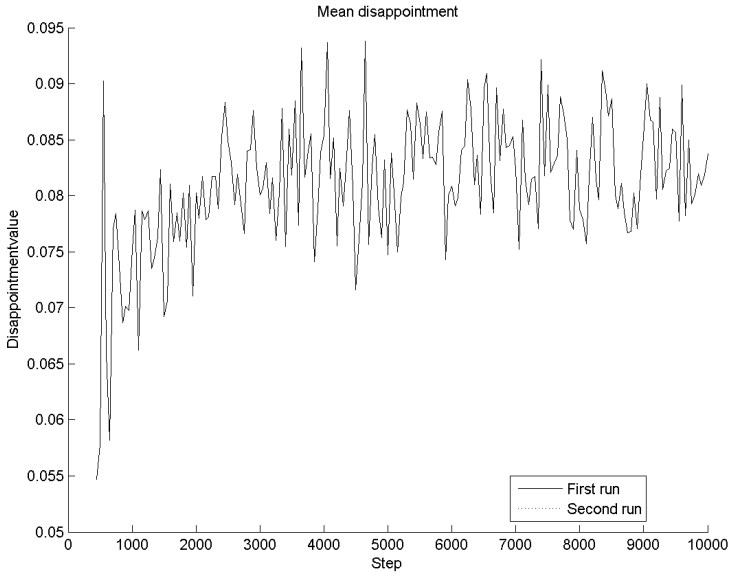

Figure B-53: Disappointment value for a positive and negative reward

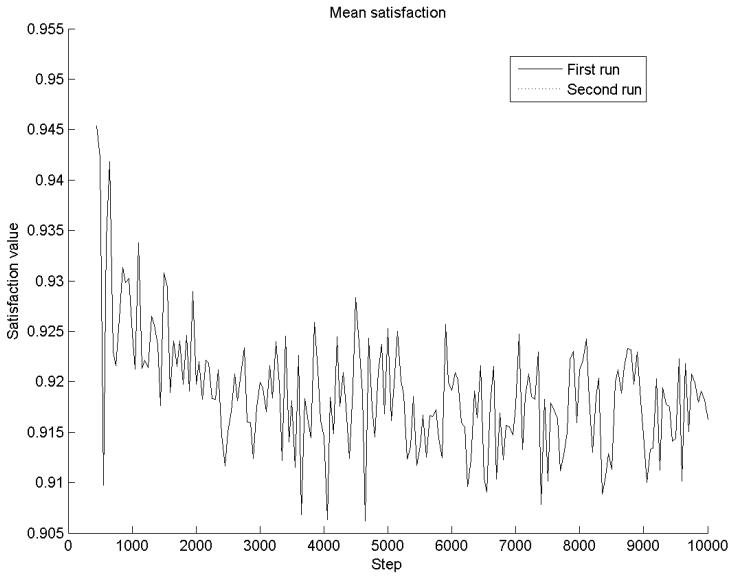

Figure B-54: Satisfaction value for a positive and negative reward

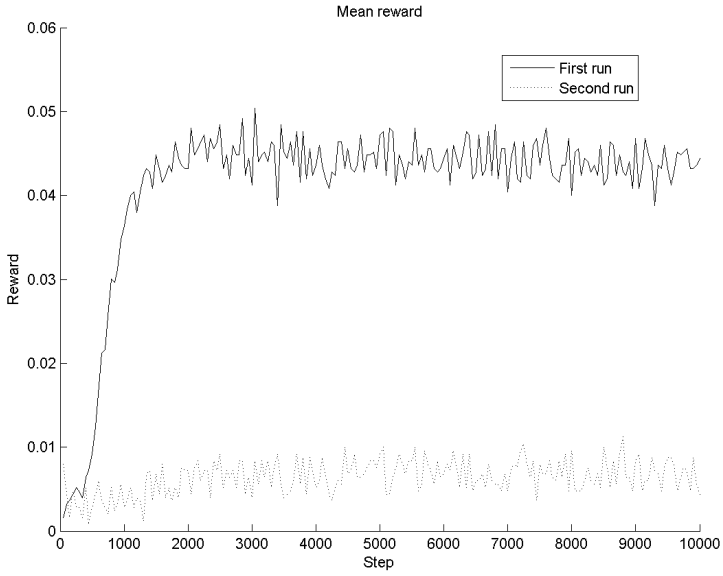

Figure B-55: Reward value for relocating the agent and relocating the reward

B-7 Relocating the reward

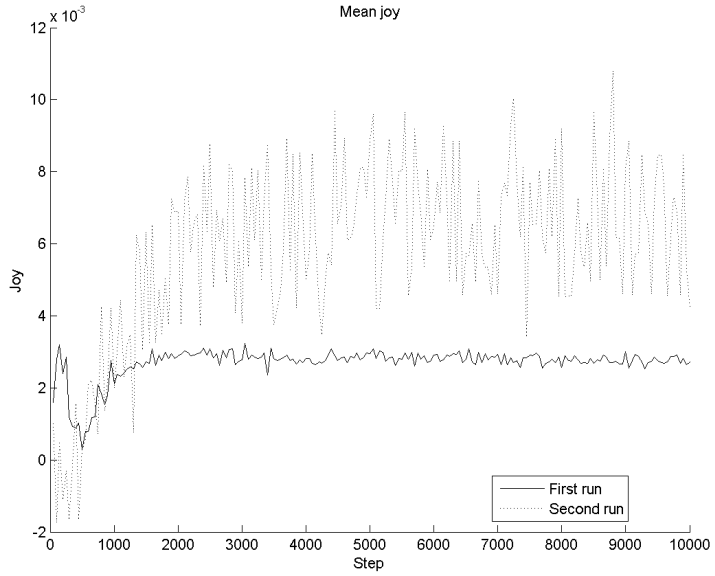

Figure B-56: Happiness value for relocating the agent and relocating the reward

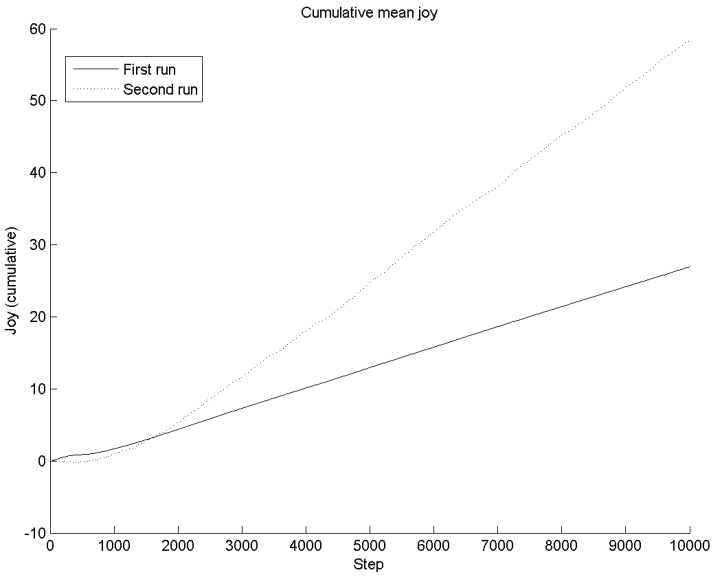

Figure B-57: Cumulative happiness value for relocating the agent and relocating the reward

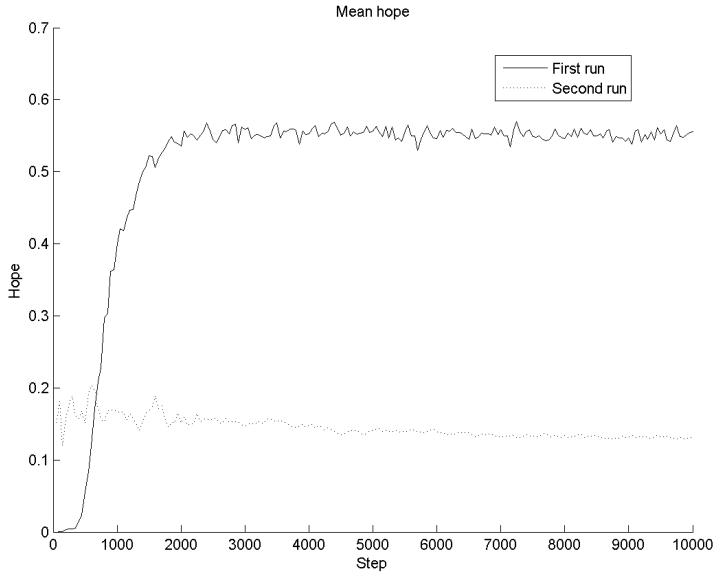

Figure B-58: Cumulative hope value for relocating the agent and relocating the reward

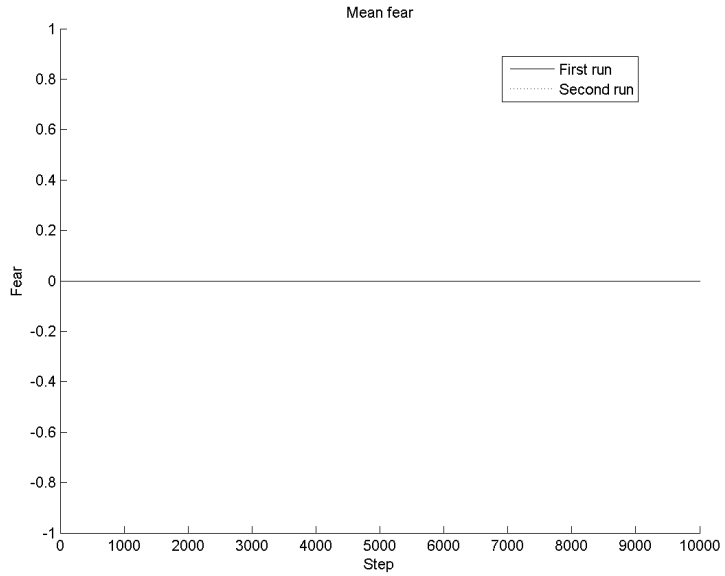

Figure B-59: Cumulative fear value for relocating the agent and relocating the reward

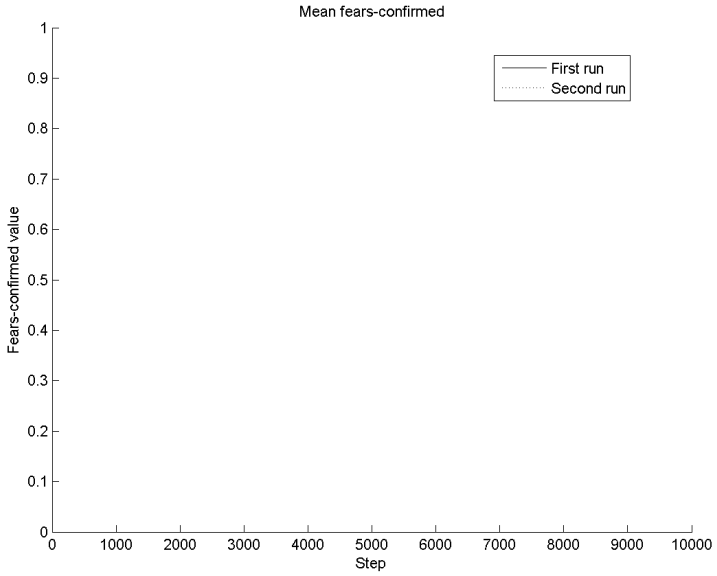

Figure B-60: Fears-confirmed value for relocating the agent and relocating the reward

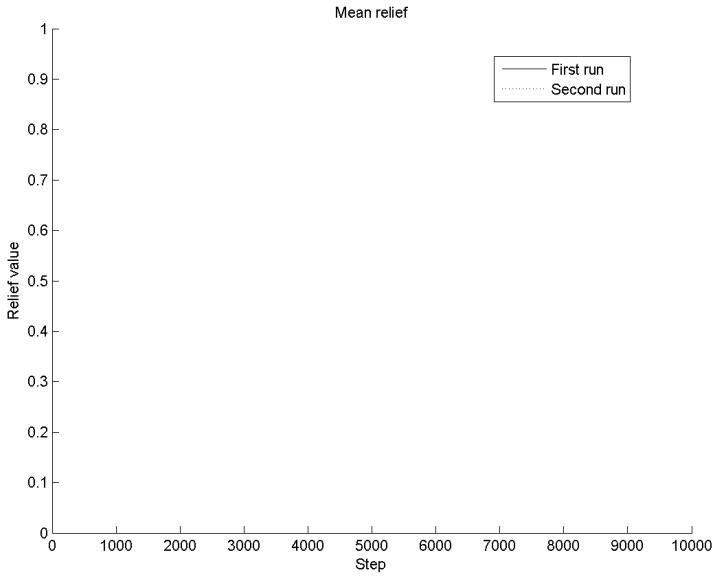

Figure B-61: Relief value for relocating the agent and relocating the reward

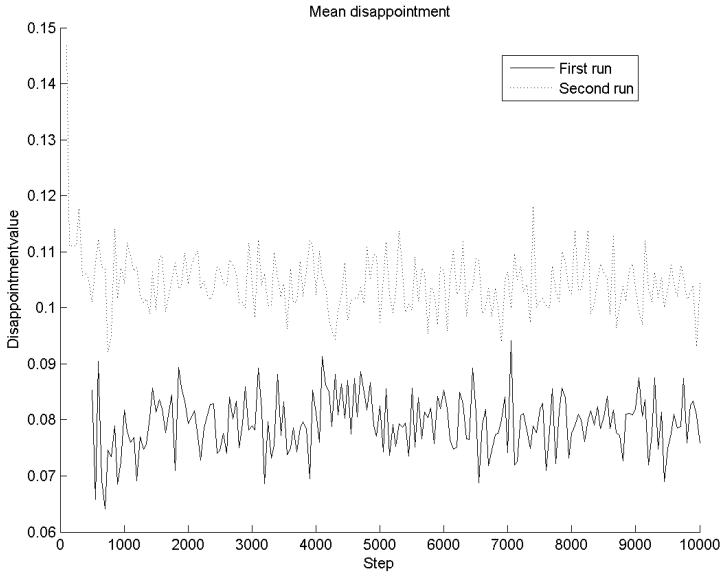

Figure B-62: Disappointment value for relocating the agent and relocating the reward

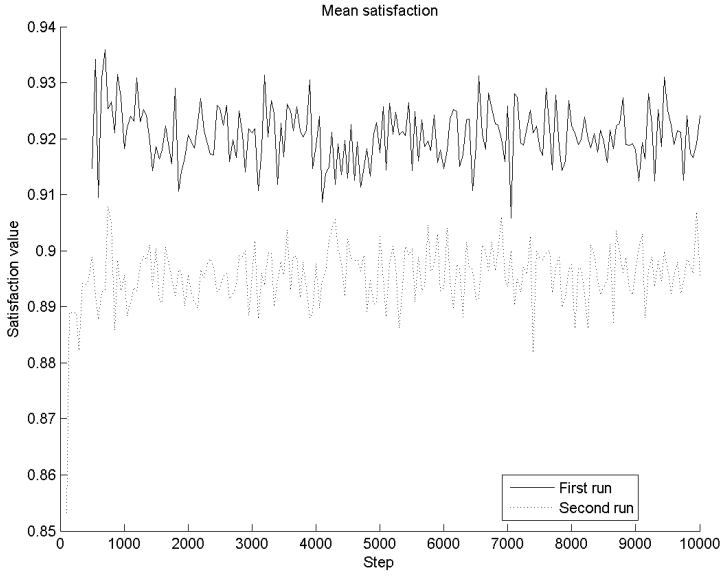

Figure B-63: Satisfaction value for relocating the agent and relocating the reward

Bibliography

[1] R. F. Baumeister, K. D. Vohs, C. N. DeWall, and L. Zhang, "How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation," *Personality and Social Psychology Review*, vol. 11, no. 2, pp. 167–203, 2007.

[2] L. A. Sroufe, *Emotional development: The organization of emotional life in the early years*. Cambridge University Press, 1997.

[3] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, vol. 1. Cambridge Univ Press, 1998.

[4] E. Hogewoning, J. Broekens, J. Eggermont, and E. Bovenkamp, "Strategies for affect-controlled action-selection in soar-rl," *Nature Inspired Problem-Solving Methods in Knowledge Engineering*, pp. 501–510, 2007.

[5] J. Broekens, W. A. Kosters, and F. J. Verbeek, "Affect, anticipation, and adaptation: Affect-controlled selection of anticipatory simulation in artificial adaptive agents," *Adaptive behavior*, vol. 15, no. 4, pp. 397–422, 2007.

[6] P. Sequeira, *Socio-Emotional Reward Design for Intrinsically Motivated Learning Agents*. PhD thesis, Universidada Técnica de Lisboa, 2013.

[7] N. L. Bottan and R. Perez Truglia, "Deconstructing the hedonic treadmill: Is happiness autoregressive?," *Journal of Socio-Economics*, vol. 40, no. 3, pp. 224–236, 2011.

[8] P. Brickman, D. Coates, R. Janoff-Bulman, *et al.*, "Lottery winners and accident victims: is happiness relative?," *Journal of personality and social psychology*, vol. 36, no. 8, p. 917, 1978.

[9] R. Veenhoven, "Is happiness relative?," *Social Indicators Research*, vol. 24, no. 1, pp. 1–34, 1991.

[10] N. Frijda, P. Kuipers, and E. Ter Schure, "Relations among emotion, appraisal, and emotional action readiness.," *Journal of Personality and Social Psychology*, vol. 57, no. 2, p. 212, 1989.

[11] H. R. Schaffer, "Cognitive components of the infant's response to strangeness," *The origins of fear*, vol. 2, p. 11, 1974.

[12] P. Ekman, "Basic emotions," *Handbook of cognition and emotion*, vol. 4, pp. 5–60, 1999.

[13] M. Lewis, "The emergence of human emotions," *Handbook of emotions*, vol. 2, pp. 265–280, 2000.

[14] E. B. Foa and M. J. Kozak, "Emotional processing of fear: exposure to corrective information.," *Psychological bulletin*, vol. 99, no. 1, p. 20, 1986.

[15] W. H. Thorpe, *Learning and instinct in animals.* Harvard University Press, 1956.

[16] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári, "Convergence results for single-step on-policy reinforcement-learning algorithms," *Machine Learning*, vol. 38, no. 3, pp. 287–308, 2000.

[17] M. S. El-Nasr, J. Yen, and T. R. Ioerger, "FlameŮfuzzy logic adaptive model of emotions," *Autonomous Agents and Multi-agent systems*, vol. 3, no. 3, pp. 219–257, 2000.

[18] R. Marinier and J. E. Laird, "Emotion-driven reinforcement learning," *Cognitive Science*, pp. 115–120, 2008.

[19] T. Myhrer, "Neurotransmitter systems involved in learning and memory in the rat: a meta-analysis based on studies of four behavioral tasks," *Brain Research Reviews*, vol. 41, no. 2, pp. 268–287, 2003.

[20] W. Schultz, "Predictive reward signal of dopamine neurons," *Journal of neurophysiology*, vol. 80, no. 1, pp. 1–27, 1998.

[21] C. Holroyd and M. Coles, "The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity.," *Psychological review*, vol. 109, no. 4, p. 679, 2002.

[22] K. Doya, "Metalearning and neuromodulation," *Neural Networks*, vol. 15, no. 4, pp. 495–506, 2002.

[23] A. Blokland, "Acetylcholine: a neurotransmitter for learning and memory?," *Brain Research Reviews*, vol. 21, no. 3, pp. 285–300, 1995.

[24] T. S. Critchfield, E. M. Paletz, K. R. MacAleese, and M. C. Newland, "Punishment in human choice: Direct or competitive suppression?," *Journal of the Experimental analysis of Behavior*, vol. 80, no. 1, pp. 1–27, 2003.

[25] P. R. Montague, B. King-Casas, and J. D. Cohen, "Imaging valuation models in human choice," *Annu. Rev. Neurosci.*, vol. 29, pp. 417–448, 2006.

[26] H. A. Simon, "A behavioral model of rational choice," *The quarterly journal of economics*, vol. 69, no. 1, pp. 99–118, 1955.

[27] H. A. Simon, "Rational choice and the structure of the environment," *Psychological review*, vol. 63, no. 2, pp. 129–138, 1956.

[28] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge university press, 1990.

[29] E. André, M. Klesen, P. Gebhard, S. Allen, and T. Rist, "Integrating models of personality and emotions into lifelike characters," *Affective interactions*, pp. 150–165, 2000.

[30] C. Bartneck, "Integrating the occ model of emotions in embodied characters," in *Workshop on Virtual Conversational Characters*, Citeseer, 2002.

[31] S. Kshirsagar, "A multilayer personality model," in *Proceedings of the 2nd international symposium on Smart graphics*, pp. 107–115, ACM, 2002.

[32] P. Ekman, "Universals and cultural differences in facial expressions of emotion.," in *Nebraska symposium on motivation*, University of Nebraska Press, 1971.

[33] B. Steunebrink, M. Dastani, and J. Meyer, "The occ model revisited," in *Proceedings of the 4th Workshop on Emotion and Computing*, vol. 65, pp. 2047–2056, 2009.

[34] H. Leventhal and K. Scherer, "The relationship of emotion to cognition: A functional approach to a semantic controversy," *Cognition and Emotion*, vol. 1, no. 1, pp. 3–28, 1987.

[35] K. Scherer, "Appraisal considered as a process of multilevel sequential checking," *Appraisal processes in emotion: Theory, methods, research*, vol. 92, p. 120, 2001.

[36] T. Wehrle and K. R. Scherer, "Towards computational modeling of appraisal theories," 2001.

[37] D. Sander, D. Grandjean, and K. R. Scherer, "2005 special issue: A systems approach to appraisal mechanisms in emotion," *Neural networks*, vol. 18, no. 4, pp. 317–352, 2005.

[38] J. Broekens, D. DeGroot, and W. A. Kosters, "Formal models of appraisal: Theory, specification, and computational model," *Cognitive Systems Research*, vol. 9, no. 3, pp. 173–197, 2008.

[39] J. Sprott, "Dynamical models of happiness," *Nonlinear Dynamics, Psychology, and Life Sciences*, vol. 9, no. 1, pp. 23–36, 2005.

[40] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica: Journal of the Econometric Society*, pp. 263–291, 1979.

[41] L. Lopes and G. Oden, "The role of aspiration level in risky choice: A comparison of cumulative prospect theory and sp/a theory," *Journal of mathematical psychology*, vol. 43, no. 2, pp. 286–313, 1999.

[42] J. Day, "The anatomy of hope and fear," *Mind*, vol. 79, no. 315, pp. 369–384, 1970.

[43] F. B. Bryant and J. A. Cvengros, "Distinguishing hope and optimism: Two sides of a coin, or two separate coins?," *Journal of Social and Clinical Psychology*, vol. 23, no. 2, pp. 273–302, 2004.

Glossary

List of Acronyms

RL	Reinforcement Learning
MDP	Markov Decision Process
MC	Monte Carlo
DP	Dynamic Programming
TD	Temporal Difference Learning
SEC	Stimulus Evaluation Check
VI	Value-iteration

List of Symbols

α	Learning rate
β	Inverse temperature in Boltzmann Action selection
δ	Temporal Difference Error
ϵ	Probability of random action selection in ϵ-Greedy
γ	Discount factor
λ	Eligibility trace factor
π	Policy
$\mathcal{A}(s_t)$	Set of actions in current state
$\mathcal{R}$	Set of rewards
$\mathcal{S}$	Set of states
a	Action
$AD(s')$	Evaluation of adjustment in the given state

$C(s, a, s')$	Confirmation emotion resulting from the given state-transition
$CO(s, a, s')$	Conduciveness of the given state-transition
$CT(s')$	Evaluation of control in the given state
$D(s, a, s')$	Desirability of state transition
$D_p(s, a, s')$	Prospect desirability of given state transition
$DI(s, a, s')$	Discrepancy from expectation of given state-transition
$F(s)$	Fear in state s
$H(s)$	Hope in state s
$HF(s)$	Hope or fear value in the (action-)value-based representation of prospect emotions
$IP(s, a, s')$	Intrinsic Pleasantness of given state-transition
$L(s, a, s')$	Likelihood of prospect state-transition
$n(...)$	Number of entity given between parentheses
$OP(s', s'')$	Outcome probability of state-transition from state just arrived in to given future state
$P(s)$	Attributed prospect emotion in state s
$P^a_{ss'}$	Probability of getting in state s' from state s after executing action a
$PR(s, a, s')$	Predictability of given state-transition
$PW(s')$	Evaluation of power in the given state
$Q(s, a)^\pi$	Value of state-action combination (s, a) under policy π
R	Expected return
r	Reward
$R^a_{ss'}$	Expected reward upon transferring from state s to state s' after executing action a
$RE(s, s')$	Relevance of given state-change
s	State
s'	Next state
s''	State after next state
$UR(s, a, s')$	Urgency created by the given state-transition
$V^\pi(s)$	Value of state s under policy π
$WB(s, a, s')$	Well-being emotion of state transition
U(s,a,s')	Unexpectedness of state transition
$_p$	Prospect-type internal emotion variable
$_Q$	Action-value based emotion variable
$_T$	Final time step
$_t$	Current time
$_V$	State-value based emotion variable
*	Given optimal policy