

## RESEARCH ARTICLE

## A Reinforcement Learning Model of Joy, Distress, Hope and Fear

Joost Broekens\*, Elmer Jacobs, Catholijn M. Jonker <sup>a</sup><sup>a</sup> *4 Mekelweg, Delft University of Technology, Delft, The Netherlands*

( 8-september-2014)

In this paper we computationally study the relation between adaptive behavior and emotion. Using the Reinforcement Learning framework, we propose that learned state utility,  $V(s)$ , models fear (negative) and hope (positive) based on the fact that both signals are about anticipation of loss or gain. Further, we propose that joy/distress is a signal similar to the error signal. We present agent-based simulation experiments that show that this model replicates psychological and behavioral dynamics of emotion. This work distinguishes itself by assessing the dynamics of emotion in an adaptive agent framework - coupling it to the literature on habituation, development, extinction, and hope theory. Our results support the idea that the function of emotion is to provide a complex feedback signal for an organism to adapt its behavior. Our work is relevant for understanding the relation between emotion and adaptation in animals, as well as for human-robot interaction, in particular how emotional signals can be used to communicate between adaptive agents and humans.

**Keywords:** Reinforcement Learning, Emotion Dynamics, Affective computing, Emotion

## 1. Introduction

Emotion and reinforcement learning (RL) play an important role in shaping behavior. Emotions drive adaptation in behavior and are therefore often coupled to learning (Baumeister, Vohs, DeWall, & Zhang, 2007). Further, emotions inform us about the value of alternative actions (Damasio, 1996) and directly influence action selection, for example through action readiness (N. Frijda, Kuipers, & Ter Schure, 1989). Reinforcement Learning (RL) (Sutton & Barto, 1998) is based on exploration and learning by feedback and relies on a mechanism similar to operant conditioning. The goal for RL is to inform action selection such that it selects actions that optimize expected return. There is neurological support for the idea that animals use RL mechanisms to adapt their behavior (Dayan & Balleine, 2002; O’Doherty, 2004; Suri, 2002). This results in two important similarities between emotion and RL: both influence action selection, and both involve feedback. The link between emotion and RL is supported neurologically by the relation between the orbitofrontal cortex, reward representation, and (subjective) affective value (see (Rolls & Grabenhorst, 2008)).

Emotion and feedback-based adaption are intimately connected in natural agents via the process called action selection. A broadly agreed-upon function of emotion in humans and other animals is to provide a complex feedback signal for a(n) (synthetic) organism to adapt its behavior (N. H. Frijda, 2004; Lewis, 2005; Ortony, Clore, & Collins, 1988; Reisenzein, 2009; Robinson & el Kaliouby, 2009; Rolls,

---

\*Corresponding author. Email: joost.broekens@gmail.com

2000; Broekens, Marsella, & Bosse, 2013). Important for the current discussion is that emotion provides feedback and that this feedback ultimately influences behavior, otherwise we can not talk about the adaptation of behavior. Behavior can be conceptualized as a sequence of actions. So, the generation of behavior eventually boils down to selecting appropriate next actions, a process called action selection (Bryson, 2007; Prescott, Bryson, & Seth, 2007). Specific brain mechanisms have been identified to be responsible for, or at the very least involved in, this process (Bogacz & Gurney, 2007; Houk et al., 2007). An important signal that influences action selection in humans is how alternative actions feel. In neuroscience and psychology this signal is often referred to as somatic marker (Damasio, 1996), affective value (Rolls & Grabenhorst, 2008) or preference (Zajonc & Markus, 1982). Another way in which emotion influences action selection is through emotion-specific action tendencies (N. H. Frijda, 2004), such as the tendency to flee or startle when affraid.

RL is a plausible computational model of feedback-based adaptation of behavior in animals. In RL an (artificial) organism learns, through experience, estimated utility of situated actions. It does so by solving the credit assignment problem, i.e., how to assign a value to an action in a particular state so that this value is predictive of the total expected reward (and punishment) that follows this action. After learning, the action selection process of the organism uses these learned situated action values to select actions that optimize reward (and minimize punishment) over time. Here we refer to situated action value as *utility*. In RL, reward, utility, and utility updates are *the* basic elements based on which action selection is influenced. These basic elements have been identified in the animal brain including the encoding of utility (Tanaka et al., 2004), changes in utility (Haruno & Kawato, 2006), and reward and motivational action value (Berridge, 2003; Berridge & Robinson, 2003; Berridge, Robinson, & Aldridge, 2009; Tanaka et al., 2004). In these studies it is argued that these processes relate to instrumental conditioning, in particular to the more elaborate computational model of instrumental conditioning called Reinforcement Learning (Dayan & Balleine, 2002; O'Doherty, 2004).

Computational modeling of the relation between emotion and reinforcement learning is useful from an emotion-theoretic point of view. If emotion and feedback-based adaptation of behavior is intimately connected in natural agents, and, RL is a computational model of feedback-based adaptation of behavior in animals, then computationally studying the relation between emotion and reinforcement learning seems promising. The hypothesis that emotion and RL are intimately connected in animals is supported by the converging evidence that both RL and emotion seem to influence action selection using a utility-like signal. Our idea that there is a strong connection between emotion, or affective signals in general, and reinforcement learning is confirmed by the large amount of neuroscientific work showing a relation between the orbitofrontal cortex, reward representation, and (subjective) affective value (for review see (Rolls & Grabenhorst, 2008)). Computational modeling generates information that helps to gain insight into the dynamics of psychological and behavioral processes of affective processing (K. R. Scherer, 2009; Broekens et al., 2013; Marsella, Gratch, & Petta, 2010). Studying the relation between RL and affect is no exception (see e.g. (Schweighofer & Doya, 2003; Cos, Canamero, Hayes, & Gillies, 2013; Krichmar, 2008; Redish, Jensen, Johnson, & Kurth-Nelson, 2007)).

Computational modeling of the relation between emotion and reinforcement learning is useful from a computer science point of view for two reasons: affective signals can enhance the adaptive potential of RL agents, and, affective signals help human-agent interaction. We briefly discuss related work in this area. While most research on computational modeling of emotion is based on cognitive

appraisal theory (Marsella et al., 2010), computational studies show that emotion-like signals can benefit Reinforcement Learning agents (Broekens, 2007; Hogewoning, Broekens, Eggermont, & Bovenkamp, 2007; Sequeira, 2013a; Gadanho, 1999; Schweighofer & Doya, 2003; Sequeira, Melo, & Paiva, 2011; Sequeira, 2013b; Sequeira, Melo, & Paiva, 2014; Cos et al., 2013; Marinier & Laird, 2008; Broekens, Kusters, & Verbeek, 2007). For example, in the area of optimization of exploration, different groups have shown that mood-like signals emerging from the interaction between the agent and its environment can be used to optimize search behavior of an adaptive agent (Broekens, 2007; Broekens et al., 2007; Hogewoning et al., 2007; Schweighofer & Doya, 2003) by manipulating the amount of randomness in the action selection process. Other studies show that affective signals coming from human observers (Broekens, 2007) (in essence an interactive form of reward shaping) or affective signals generated by a cognitive assessment of the agent itself (Gadanho, 1999) (in essence a form of intrinsic reward) can speed up convergence. It has also been shown that adaptive agents can use intrinsic reward signals that resemble emotional appraisal processes, and that these signals help the agent to be more adaptive than standard RL agents in tasks that lack complete information (Sequeira, 2013b; Sequeira et al., 2011, 2014), similar to the work presented in (Marinier & Laird, 2008). In these two approaches emotional appraisal generates an intrinsic reward signal that is used for learning. A key difference between these latter two works is that the emotional appraisal in (Marinier & Laird, 2008) also appraises based on the semantics of the state of the RL agent, such as if there is a wall or not, while in (Sequeira et al., 2014) appraisal is solely based on RL primitives. Furthermore, already in 1999 an exhaustive attempt has been made to investigate different ways in which both emotion and RL can jointly influence action selection (Gadanho, 1999).

To be able to claim that one can model emotions with reinforcement learning it is essential to replicate psychological and behavioral findings on emotion and affect. An example in the context of cognitive appraisal theory is the work by (Gratch, Marsella, Wang, & Stankovic, 2009) investigating how different computational models predict emotions as rated by human subjects. Examples related to conditioning and emotion include (Moren, 2000; Steephen, 2013; Lahnstein, 2005). However, to ground the affective labeling of RL-related signals we believe that replication of emotion dynamics is needed *in the context of learning a task*.

Solid grounding, for example of fear being negative predicted value, opens up the way towards a domain independent model of emotion within the RL framework which reduces the need to design a specific emotion model for each particular agent. Although such domain independent computational models exist for cognitive appraisal theory (Marsella & Gratch, 2009; Dias, Mascarenhas, & Paiva, 2011), none exist for RL-based agents. Solid grounding of emotion in RL-based artificial agents also means we know what an emotion means in terms of the *functioning* of the agent (Canamero, 2005; Kiryazov, Lowe, Becker-Asano, & Ziemke, 2011). This is important because grounding emotion in adaptive behavior helps human-agent interaction; the expression of that emotion by a virtual agent or robot becomes intrinsically meaningful to humans, i.e., humans can relate to why the emotional signal arises. An adaptive robot that shows, e.g., fear grounded in its learning mechanism will be much easier to understand for humans, simply because we humans know what it means to have fear when learning to adapt to an environment. So, solving the grounding problem directly helps human-robot and human-agent interaction. This means that for an emotional instrumentation to be useful, adaptive benefit per se is not a requirement. Even if the emotion has no direct influence on the agent, it is still useful for human-agent interaction and for understanding

the fundamentals of how emotion emerges from adaptive behavior.

The contribution of our work in this field is that we aim to show a direct mapping between RL primitives and emotions, and assess the validity by replicating psychological findings on emotion dynamics. This focus on replication of psychological findings is an essential difference with (El-Nasr, Yen, & Ioerger, 2000). Our focus on emotions is also a key difference with the approach by (Sequeira et al., 2014), which focuses on appraisal processes. Another defining element of our approach is that we do not focus on learning benefit for artificial agents per se, as we believe that before affectively labelling a particular RL-based signal, it is essential to investigate if that signal behaves according to what is known in psychology and behavioral science. The extent to which a signal replicates emotion-related dynamics found in humans and animals is a measure for the validity of giving it a particular affective label.

In this paper we report on a study that shows plausible emotion dynamics for joy, distress, hope and fear, emerging in an adaptive agent that uses Reinforcement Learning (RL) to adapt to a task. We propose a computational model of joy, distress, hope, and fear instrumented as a mapping between RL primitives and emotion labels. Requirements for this mapping were taken from emotion elicitation literature (Ortony et al., 1988), emotion development (Sroufe, 1997), hope theory (Snyder, 2002) the theory on optimism (Carver, Scheier, & Segerstrom, 2010), and habituation and fear extinction (Bottan & Perez Truglia, 2011; Brickman, Coates, & Janoff-Bulman, 1978; Veenhoven, 1991; Myers & Davis, 2006). Using agent-based simulation where an RL-based agent collects rewards in a maze, we show that emerging dynamics of RL primitives replicate emotion-related dynamics from psychological and behavioral literature.

## 2. 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 (Sutton & Barto, 1998). 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  $\pi$ , with  $\pi_t(s, a)$  the probability that ( $a_t = a$ ) if ( $s_t = s$ ). In Reinforcement Learning, 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.

The total expected reward  $R$  at time  $t$  is finite in applications with a natural notion of an end state. To deal with situations without an end state, a discount factor  $\gamma$ , where  $0 \leq \gamma \leq 1$ , discounts rewards that are further in the future to ascertain a finite sum, such that:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}. \quad (1)$$

Standard RL focuses on problems satisfying the Markov Property, which states that the probability distribution of the future state depends only on the previous state and action. We can define a transition probability  $P_{ss'}^a$  and the expected reward  $R_{ss'}^a$ .

$$P_{ss'}^a = Pr\{s_{t+1} = s' | s_t = s, a_t = a\} \quad (2)$$

$$R_{ss'}^a = E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\}. \quad (3)$$

With these elements it is possible to determine a value  $V^\pi(s)$  for each state. The values are specific per policy and are defined such that:

$$V^\pi(s) = E_\pi\{R_t | s_t = s\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right\}. \quad (4)$$

The value of a state is typically arbitrarily initialized and updated as the state is visited more often. Since the values are policy dependent, they can be used to evaluate and improve the policy to form a new one. Both are combined in an algorithm called value iteration, where the values are updated after each complete sweep of the state space  $k$  such that:

$$V_{k+1}(s) = \max_a \sum_{s'} P_{ss'}^a [R_{ss'}^a + \gamma V_k(s')]. \quad (5)$$

After convergence of the values, the policy simplifies to:

$$\pi(s) = \arg \max_a \sum_{s'} P_{ss'}^a [R_{ss'}^a + \gamma V(s')] \quad (6)$$

The use of this algorithm requires a complete knowledge of the state-space, which is not always available. Temporal Difference learning estimates values and updates them after each visit. Temporal Difference learning has been proposed as a plausible model of human learning based on feedback (Schultz, 1998; Holroyd & Coles, 2002; Doya, 2002). The simplest method, one-step Temporal Difference learning, updates values according to:

$$V(s_t) \leftarrow V(s_t) + \alpha [r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \quad (7)$$

with  $\alpha$  now representing the learning rate. After convergence, the values can be used to determine actions. We use softmax action selection instantiated by a Boltzmann (Gibbs) distribution, argued to be a model of human action selection (Critchfield, Paletz, MacAleese, & Newland, 2003; Montague, King-Casas, & Cohen, 2006; Simon, 1955, 1956). Actions are picked with a probability equal to:

$$p(a) = \frac{e^{\beta Q(s,a)}}{\sum_{b=1}^n e^{\beta Q(s,b)}} \quad (8)$$

where  $a$  is an action,  $\beta$  is a positive parameter controlling randomness and  $Q(s, a)$  is the value of taking a specific action according to:

$$Q(s, a) = \sum_{s'} P_{ss'}^a [V(s') + R_{ss'}^a]. \quad (9)$$

### 3. Mapping emotions

In essence, the computational model of emotion we propose is a mapping between RL primitives (reward, value, error signal, etc..) and emotion labels. Our mapping focuses on well-being emotions and prospect emotions, in particular joy/distress and hope/fear respectively, two emotion groups from the OCC model (Ortony et al., 1988), a well-known computation-ready psychological model of cognitive emotion elicitation. In this section we detail the rationale for our mapping, after which, in section 4, we evaluate the mapping using carefully set-up experiments to replicate emotion dynamics.

#### 3.1. Emotional development, habituation, and high versus low hope

Learning not only drives adaptation in human behavior, but also affects the complexity of emotions. Humans start with a small number of distinguishable emotions that increases during development. 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 responsiveness to the environment (Sroufe, 1997). 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 the actions required to reach a certain goal. In other words, the capability of feeling anger reflects the child's early knowledge of its abilities. Anger is followed by fear, usually reported first at 7 or 8 months. Fear requires a comparison of multiple events (Schaffer, 1974) and is therefore more complex than earlier emotions. Surprise can also be noted within the first 6 months of life.

Apart from the development of emotions, habituation and extinction are important affective phenomena. Habituation is the decrease in intensity of the response to a reinforced stimulus resulting from that stimulus+reinforcer being repeatedly received, while extinction is the decrease in intensity of a response when a previously conditioned stimulus is no longer reinforced (Myers & Davis, 2006; Bottan & Perez Truglia, 2011; Brickman et al., 1978; Veenhoven, 1991; Foa & Kozak, 1986). A mapping of RL primitives to emotion should be consistent with habituation and extinction, and in particular fear extinction as this is a well studied phenomenon (Myers & Davis, 2006).

Further, hope theory defines hope as a positive motivational state resulting from successful goal directed behavior (behavior that is executed and is along pathways that lead to goal achievement) (Snyder, 2002; Carver et al., 2010). This is very much in line with the theory on optimism (Carver et al., 2010). A mapping of hope (and fear) should also be consistent with major psychological findings in this area of work. For example, high-hope individuals perform better (and have more hope) while solving difficult problems as compared to low-hope individuals (see (Snyder, 2002) for refs). An underlying reason for this seems to be that high-hope individuals explore more (and more effective) pathways towards the goal they are

trying to achieve and feel confident ('I can do it'), while low-hope individuals get stuck in rumination ('this can go wrong and that...') (Snyder, 2002).

### 3.2. Mapping joy and distress

Joy and distress are the first emotions to be observable in infants. As such, these emotions should be closely related to reward signals. A first possible choice to map joy/distress would be to use the reward  $r_t$ . Any state transition that yields some reward therefore causes joy in the agent (and punishment would cause distress). However, anticipation and unexpected improvement can also result in joy (Spratt, 2005) and this contradicts the previous mapping. We need to add an anticipatory element of RL. So, we could represent joy by  $r_t + V(s_t)$ . However, this contradicts our knowledge about habituation, which states that the intensity of joy attributed to a state should decrease upon visiting that state more often. So, we should incorporate the convergence of the learning algorithm by using the term  $r_t + V(s_t) - V(s_{t-1})$ , which continuously decreases as values come closer to convergence. This mapping still lacks the concept of expectedness (Ortony et al., 1988). We add an unexpectedness term, derived from the expected probability of the state-transition that just took place, which is  $(1 - P_{s_{t-1}s_t}^{a_{t-1}})$ . We let:

$$J(s_{t-1}, a_{t-1}, s_t) = (r_t + V(s_t) - V(s_{t-1}))(1 - P_{s_{t-1}s_t}^{a_{t-1}}) \quad (10)$$

where  $J$  is the joy (or distress, when negative) experienced after the transition from state  $s_{t-1}$  to state  $s_t$  through action  $a_{t-1}$ . Joy should be calculated before updating the previous value, since it reflects the immediate emotion after arriving in the given state. This mapping coincides with the mapping in the OCC model, which states that joy is dependent on the desirability and unexpectedness of an event (Ortony et al., 1988).

### 3.3. Mapping hope and fear

According to theory about emotional development, joy and distress are followed by anger and fear. Hope is the anticipation of a positive outcome and fear the anticipation of a negative outcome (Ortony et al., 1988). Anticipation implies that some representation of the probability of the event actually happening must be present in the mapping of both of these emotions. The probability of some future state-transition in Reinforcement Learning is  $P_{s_t s_{t+1}}^{a_t}$ . This is implicitly represented in the value  $V(s_t)$  which after conversion is a sampling of all chosen actions and resulting state transitions, so a first assumption may be to map  $V(s_t)$  to hope and fear. Under this mapping, fear extinction can happen by a mechanism similar to *new learning* (Myers & Davis, 2006). If action-selection gives priority to the highest valued transition, then a particular  $V(s)$  that was previously influenced by a negatively reinforced next outcome will, with repeated visits, increase until convergence, effectively diminishing the effect of the negative association by developing new associations to better outcomes, i.e., new learning.

Alternatively, we can use probability and expected joy/distress explicitly in order to determine the hope/fear value for each action. However, as any transition in a direction that decreases reward translates to a loss in value this would also be a source of fear. As a result, the agent would experience fear even in a situation with only positive rewards. In some situations, loss of reward should trigger fear (losing all your money), but it is difficult to ascertain if fear is then in fact a precursor to

actual negativity, or a reaction to the loss of reward. As such we stick to the simpler model where the intensity of hope and fear equals  $V_+(s)$  and  $V_-(s)$  respectively:

$$\text{Hope}(s_t) = \max(V(s_t), 0), \text{Fear}(s_t) = \max(-V(s_t), 0) \quad (11)$$

The OCC model states that hope and fear are dependent on the expected joy/distress and likelihood of a future event (Ortony et al., 1988), which is again consistent with our mapping. This also fits well with hope theory, explicitly defining hope in terms of goal value, possible pathways towards goals, and commitment to actions that go along these pathways (Snyder, 2002). The value function in RL nicely captures these aspects in that goal values are encoded in the value of each state, based on the potential pathways (state action sequences) of actions to be committed to (the policy). In addition, this is consistent with the finding that expected utility models are predictive of the intensity of the prospect-based emotions (i.e. hope and fear) (Gratch et al., 2009).

#### 4. Validation requirements

The main research question in this paper concerns the validity of the mapping we propose between the emotion labels joy/distress/fear/hope and the RL primitives as detailed above. To test the validity, we state requirements based on habituation, development, fear extinction and hope literature.

*Requirement 1.* Habituation of joy should be observed when the agent is presented repeatedly to the same reinforcement (Bottan & Perez Truglia, 2011; Brickman et al., 1978; Veenhoven, 1991; Foa & Kozak, 1986).

*Requirement 2.* In all simulations, joy/distress is the first emotion to be observed followed by hope/fear. As mentioned earlier, human emotions develop in individuals from simple to complex (Sroufe, 1997).

*Requirement 3.* Simulations should show fear extinction over time through the mechanism of new learning (Myers & Davis, 2006).

*Requirement 4.* Lowered expectation of return decreases hope (Veenhoven, 1991; Ortony et al., 1988; Snyder, 2002), and low-hope individuals as opposed to high-hope individuals are particularly sensitive to obstacles. This dynamic should be visible due to at least the following two interventions: First, if we lower an agent's expectation of return by lowering the return itself, i.e., the goal reward, then we should observe a decrease in hope intensity. This does not need to be tested experimentally as it is a direct consequence from the model: if the goal reward is lowered, then the value function must reflect this. Any RL learning mechanism will show this behavior as the value function is an estimate of cumulative future rewards. Second, if we model low- versus high-hope agents by manipulating the expectation of return as used in the value update function  $V(s)$ , we should observe that low-hope agents suffer more from punished non-goal states (simulated obstacles) than high-hope agents. Suffering in this case means a lower intensity hope signal and slower learning speed.

*Requirement 5.* Increasing the unexpectedness of results of actions increases the intensity of the joy/distress emotion. Predictability relates to the expectedness of an event to happen, and this can be manipulated by adding randomness to action selection and action outcomes. Increasing unexpectedness should increase the intensity of joy/distress (Ortony et al., 1988; K. Scherer, 2001).



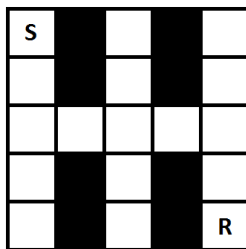


Figure 1. The maze used in the experiments. Each square is a state. The agent starts at S and the reward can (initially) be found at R. For the simulations that test fear extinction and the effect of expectation of return, the end of all non-goal arms in the maze can contain a randomly placed punishment (see text)

Table 1. Control and varied values of different parameters used in the simulations

	Control setting	Variation
Obstacles (punished arms)	no	yes
Dispositional hope	$MAX_A$	<i>Bellman</i>
Predictability (1)	$P(action\ failure) = 0.1$	$P(action\ failure) = 0.25$
Predictability (2)	Return to start	Relocate reward

#### 4.1. Experimental setup

We ran our validation tests in an agent-based simulation implemented in Java. An RL agent acts in a small maze. The maze has one final goal, represented by a single positively rewarded state. The task for the agent is to learn the optimal policy to achieve this goal. The agent always starts in the top-left corner and can move in the four cardinal directions. Collision with a wall results in the agent staying in place. Maze locations are states (17 in total). The agent learns online (no separate learning and performing phases). The maze used in all experiments is shown in Figure 1.

In all simulations the inverse action-selection temperature  $\beta$  equals 10, the reward in the goal state  $r(goal)$  equals 1, and the discount factor  $\gamma$  equals 0.9. To test the effect of expectation of return and predictability of the world, we varied several parameters of the task (see Table 1 for an overview). To model obstacles in the task, we placed negative rewards at all other maze arms except the goal and starting location equal to  $-1$ . To model high-hope versus low-hope agents (dispositional hope) calculation of state values was either based on  $MAX_A$  or *Bellman*.  $MAX_A$  (Equation 12) takes into account only the best possible action in a state for the calculation of that state's new value, while *Bellman* (Equation 13) takes into account all actions proportional to their probabilities of occurring. As such, actions with negative outcomes are also weighted into the value function. Variation in predictability was modeled as a variation in the chance that an action does not have an effect (0.25 versus 0.1, representing an unpredictable versus a predictable world respectively) as well as the consequence of gathering the goal reward (agent returns to start location versus reward is relocated randomly in the maze, the latter representing unpredictable goal locations). Unless mentioned otherwise, a simulation consists of a population of 50 different agents with each agent running the task once for a maximum of 10000 steps, which appeared in pre-testing to be long enough to show policy convergence. To reduce the probability that our results are produced by a "lucky parameter setting", each run has gaussian noise over the parameter values for  $\beta$ ,  $\gamma$ , and the probability that an action fails. We pulled these values from a normal distribution such that 95% of the values are within 5% of the given mean.

To be able to calculate the unexpectedness term in our model of joy/distress,

we need transition probabilities. Temporal Difference learning does not require a model, while Value Iteration requires a complete model. Therefore, we use a form of value iteration that uses an estimate of the transition model to update the value of the current state, such that:

$$V(s_t) \leftarrow \max_a \sum_s P_{ss'}^a [R_{ss'}^a + \gamma V(s')] . \quad (12)$$

$$V(s_t) \leftarrow \sum_a \pi(s, a) \sum_s P_{ss'}^a [R_{ss'}^a + \gamma V(s')] . \quad (13)$$

This is a simple method that converges to the correct values under the same circumstances as any Monte Carlo method. After a transition to some state  $s'$ , the estimated transition model of state  $s$  is updated, allowing  $V(s)$  to be updated at the next visit to that state. This approach is similar to Temporal Difference learning with learning rate  $\alpha = 1$  as presented in Equation 7 but uses a model instead of sampling.

## 5. Experimental results

### 5.1. Joy habituates

To test if joy habituates over time, we ran a simulation using the control settings in Table 1. We analyse a representative signal for joy/distress for a *single* agent during the first 2000 steps of the simulation (Figure 2). We see that, for a number of steps, the agent feels nothing at all, reflecting not having found any rewards yet. A sudden spike of joy occurs the first time the reward is collected. This is caused by the error signal and the fact that the rewarded state is completely novel, i.e., high unexpectedness. Then joy/distress intensity equals 0 for some time. This can be explained as follows. Even though there are non-zero error signals, the unexpectedness associated with these changes is 0. As there is only a 10% chance that an action is unsuccessful (i.e. resulting in an unpredicted next state  $s'$ ), it can take some time before an unexpected state change co-occurs with a state value update. Remember that joy/distress is derived from the change in value *and* the unexpectedness of the resulting state. Only once an action fails, the joy/distress signals start appearing again, reflecting the fact that there is a small probability that the expected (high-valued) next state does not happen. The joy/distress signal is much smaller because it is influenced by two factors: at convergence the unexpectedness goes to 0.1, and, the difference between the value of two consecutive states approaches 0.1 (taking the discount factor into account, see discussion). Individual positive spikes are caused by successful transitions toward higher valued states (and these continue to occur as  $\gamma < 1$ ), while the negative spikes are transitions toward lower valued states, both with low intensities caused by high expectedness.

To summarize, these results show that joy and distress emerge as a consequence of moving toward and away from the goal respectively, which is in accordance with (Ortony et al., 1988). Further, habituation of joy/distress is also confirmed: the intensity of joy goes down each time the reward is gathered.

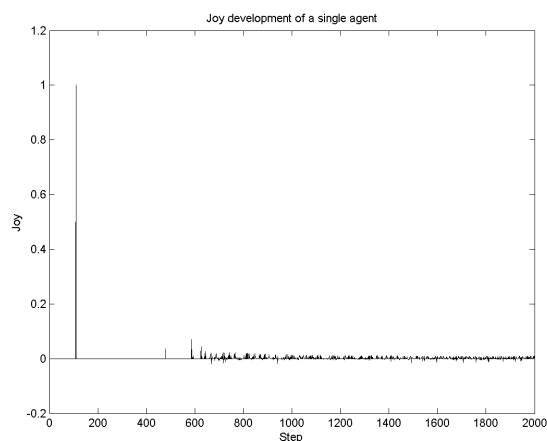


Figure 2. Intensity of joy/distress for a single agent, observed in the first 2000 steps. Later intensity of joy is strongly reduced compared to the intensity resulting from the first goal encounter (spike at  $t=100$ ).

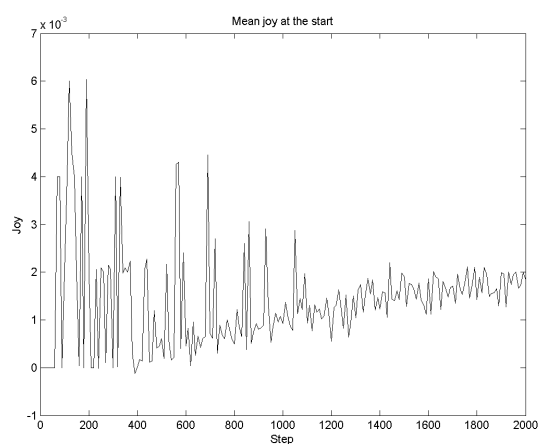


Figure 3. Intensity of joy/distress, mean over 50 agents, observed in the first 2000 steps. The noisy signal in the first 1000 steps reflects the fact that this is a non-smoothed average of only 50 agents, with each agent producing a spike of joy for the first goal encounter.

## 5.2. Joy/distress occur before hope/fear

Based on the same simulation, we test if joy/distress is the first emotion to be observed followed by hope/fear. We plot the mean joy/distress and hope over all 50 agents for the first 2000 steps. We can see that joy (Figure 3) appears before hope (Figure 4). Theoretically, state values can only be updated once an error signal received, and therefore hope and fear can only emerge *after* joy/distress. This order of emergence is of course confirmed by the simulation results.

## 5.3. Fear extincts

To test for the occurrence of fear extinction, we ran the same simulation but now with punished non-goal arms (the agent is relocated at its starting position after the punishment). This introduces a penalty for entering the wrong arm. Further, we varied the value function to be either *Bellman* or  $MAX_a$  modeling low- versus high hope agents. Fear is caused by incorporating the punishment in the value function. We ran the simulation for 500 agents and 5000 steps with all other settings kept the same as in the previous simulation (see Table 1). We plot the average intensity of fear for 500 agents for 5000 steps in and a detailed plot of the first

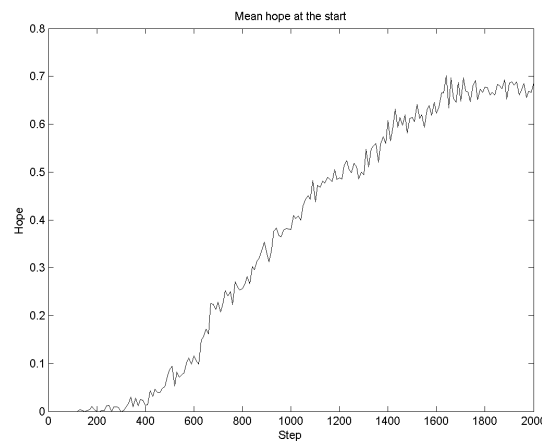


Figure 4. Intensity of hope( $V_+(s)$ ), mean over 50 agents, observed in the first 2000 steps. Over the course of learning, the hope signal grows to reflect the anticipation of the goal reward, while in the first several hundred trials virtually no hope is present.

500 steps (Figure 5). Our model successfully replicates fear extinction. Fear converges to 0 for  $MAX_A$  and *Bellman* agents. Convergence is quicker and intensity is lower for  $MAX_A$  than for *Bellman* agents. There is in fact almost no fear for  $MAX_a$  (high-hope) agents. This can be explained as follows. Actions that avoid the punishment take precedence in the  $MAX_a$  calculation of the state value; if *any* action is available that causes no penalty whatsoever, the value is never negative, i.e., no fear. It represents an agent that assumes complete control over its actions and is therefore not "afraid" once it knows how to avoid penalties, i.e. it is an optimistic (high-hope) agent. The *Bellman* agents take punishment into account much longer and to a greater extent, simply because they weight all possible outcomes in their calculation of the update to a state value. This mechanism demonstrates a known mechanism for fear extinction, called *new learning* (Myers & Davis, 2006). New learning explains fear extinction by proposing that novel associations with the previously fear-conditioned stimulus become more important after repeated presentation of that stimulus. This results in a decrease in fear response, not because the fear association is forgotten but because alternative outcomes become more important.

To summarize, our model thus nicely explains why high-hope agents experience less fear, and for a shorter time than low-hope (or realistic) agents.

#### 5.4. Expectation of return influences hope

To test the requirement (in this case more so a hypothesis) that high-hope agents suffer less from obstacles and produce higher hope than low-hope agents, we varied high- versus low-hope and the presence of obstacles (see Table 1). Results show that low-hope agents perform worse in the presence of punished non-goal maze arms while high-hope agents perform better in the presence of such punishment (Figure 6). This is consistent with hope theory stating that "high-hope people actually are very effective at producing alternative (ed:problem solving) routes - especially in circumstances when they are impeded" (Snyder, 2002). Further, low-hope agents generate more fear in the presence of punishment (Figure 5), resulting in an overall negative "experience" at the start of learning the novel task (Figure 7). This replicates psychological findings that high-hope as opposed to low-hope individuals are more effective problem solvers, especially in the face of obstacles (Snyder, 2002) and that low-hope individuals are more fearful and experience more

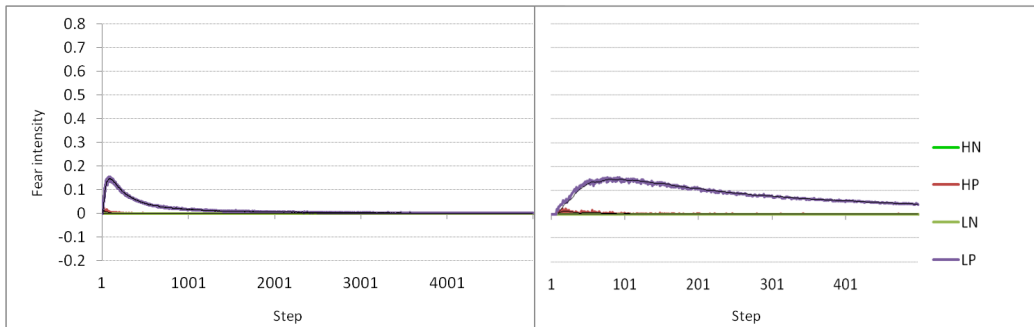
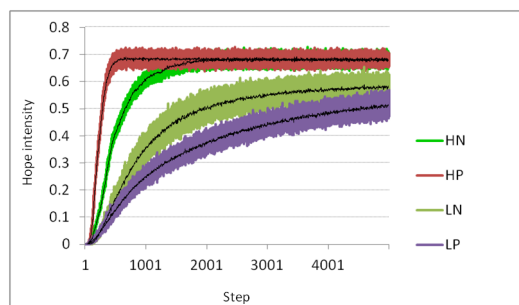


Figure 5. Left: intensity of fear, mean over 500 agents, 5000 steps, for four conditions: H/L (high-hope/low-hope), and N/P (punished arms versus no punishment). Fear extinguishes over time and low-hope agents (Bellman update) generate more fear in the presence of punishment than high-hope agents ( $MAX_a$  update). Right: zoom-in of Left figure for the first 500 steps.



V(s)

Figure 6. Intensity of hope, mean over 500 agents, 5000 steps, for four conditions: H/L (high-hope/low-hope), and N/P (punished arms versus no punishment). Low-hope agents (Bellman update) perform worse in the presence of punished non-goal maze arms while high-hope ( $MAX_a$ ) agents perform better in the presence of such punishment.

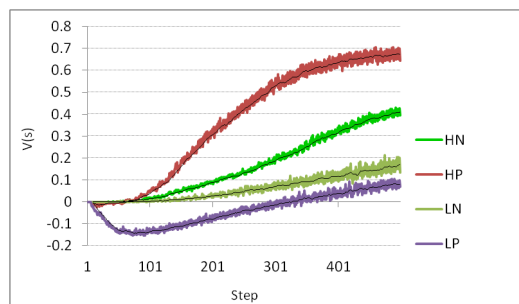


Figure 7. State value interpreted as the agent's "experience", mean over 500 agents, 500 steps, for four conditions: H/L (high-hope/low-hope), and N/P (punished arms versus no punishment).

distress in the face of obstacles. Our manipulation of the update function is successful at varying high- versus low-hope, and this manipulation is consistent with hope theory, stating that low-hope individuals tend to consider many of the possible negative outcomes, thereby blocking progress along fruitful pathways towards the goal. This is exactly in line with what happens when the *Bellman* update is used instead of the  $MAX_A$  update: also consider actions with negative outcomes in your current estimate of value, instead of only the best possible ones. As such, the hypothesis is supported.

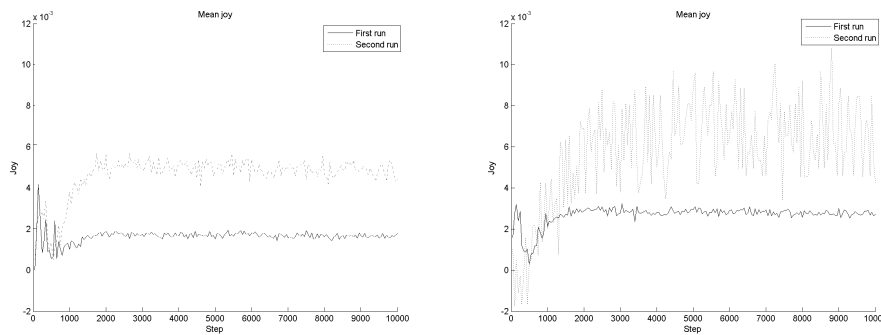


Figure 8. Intensity of joy, mean over 50 agents. Left figure shows the difference between a probability of 0.1 (first run) versus 0.25 (second run) of failing an action. The right figure shows the difference between returning the agent (first run) to its starting position versus relocating the reward (second run)

### 5.5. *Unexpectedness increases joy intensity*

To test our requirement that increasing the unexpectedness of results of actions increases the intensity of joy/distress, we vary predictability of the world. In our current setting, there are two ways to vary predictability. First, we can make the results of an action stochastic, for example by letting the action fail completely every once in a while. Second, we give rewards at random points rather than at the same transition all the time. This randomizes the reinforcing part of an experiment. Note that making action selection a random process does result in unpredictable behavior and an inefficient policy, but does not change the predictability of the *effects of an action* once it is chosen, so this does not manipulate expectedness.

First, we increased the probability for an action to fail (failure results in no state change). The resulting mean intensity of joy for 50  $MAX_a$  agents is shown in Figure 8, left. Second, we randomly relocated the reward after each time it was collected, instead of returning the agent to the starting position. The mean intensity of joy for 50  $MAX_a$  agents is shown in Figure 8, right. In both simulations the intensity of joy reactions is larger as shown by the bigger noise in the signal, indicative of bigger joy spikes. 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% ( $1/17states$ ). 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 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. These results are consistent with the psychological finding that unpredictability of outcomes result in higher intensity of joy (Ortony et al., 1988; K. Scherer, 2001).

## 6. Discussion

Overall our experiments replicate psychological and behavioral dynamics of emotion quite well. Here we discuss several assumptions and findings of the simulations in more detail.

Joy habituation is artificially fast as a result of how unexpectedness is calculated in the model. Once state transitions have been found to lead to the rewarded goal state, the unexpectedness associated with these transitions stays at 0 until variation in outcome has been observed. The cause is that the probability of an action to fail is small and the world model is empty. In other words, the agent does not know

about any alternative action outcomes and thus can only conclude that there is no unexpectedness associated with a particular outcome, i.e., no joy/distress. Due to this mechanism habituation cannot be observed by looking at the average intensity of joy/distress over a population of agents (Figure 3). The reason is twofold. First, the joy/distress signal does not converge to 0 due to the  $\gamma < 1$ , and in fact will become a stable but small signal for each agent. Second, intense spikes of joy/distress happen sporadically and only in the beginning because of the unexpectedness mechanism. Averaging over multiple agents therefore suggests that the joy signal for individual agents becomes bigger, but this is not the case. This is an artifact of low frequency high intensity joy/distress in the beginning of learning, combined with high frequency low intensity joy/distress in the end. This differs from humans, who in general have a model of alternative outcomes. We can model this in several ways: by starting with a default model of the environment; by assuming a small default probability for all states as potential outcome of an action in an arbitrary state (i.e., all  $P_{sas'}$  are non-zero); or, by dropping the unexpectedness term altogether and equating joy/distress to the error signal. These alternatives are closer to reality than our current model.

Technically speaking, joy/distress in our model is not derived from the error signal, but based on the difference in expected value between the current state and the previous state, i.e.,  $r_t + V(s_t) - V(s_{t-1})$ . If we assume a greedy policy, no discount, and an update function that uses the  $MAX_a$  value to update states then this signal is equivalent to the error signal because the expected value of any previous state  $V(s_{t-1})$  converges to  $r_t + V(s_t)$  (the policy ensures this). The main difference with a model that is derived from the error signal is that joy/distress signals react differently to the discount factor  $\gamma$  and the learning rate  $\alpha$ . In our model, the amount of residual distress/joy present at convergence is proportional to the discount factor, because the discount factor determines the difference between  $V(s_{t-1})$  and  $r_t + V(s_t)$  at convergence. If joy/distress were really equal to the error signal, then joy/distress would become 0 at convergence because the error would be 0. Also, the intensity of joy/distress would be proportional to the learning rate  $\alpha$  if joy/distress is derived from the RL update signal. If  $\alpha$  is high, joy/distress reactions are close to  $V(s_{t-1}) - r_t + V(s_t)$ , if  $\alpha$  is small the signal is close to 0. In our model, this is not the case, as the actual  $V(s_{t-1})$  and  $r_t + V(s_t)$  are taken, not the difference weighted by  $\alpha$ . We observed the discount-dependent habituation effect in our model as an habituation of joy/distress intensity that does not end up at 0. In our model habituation is predicted to have diminishing intensity but not a complete loss of the joy/distress response. This translates to humans that still experience a little bit of joy, even after repeatedly receiving the same situational reward. Further study should investigate the plausibility of the two alternative models for joy/distress.

In previous work (Jacobs, Broekens, & Jonker, 2014a, 2014b) we argued that the influence of different update functions on fear should be investigated for the following reason: agents that use a  $MAX_A$  update function are very optimistic, and assume complete control over their actions. Our model correctly predicted that fear extinction rate was very quick (and can not depend on the strength of the negative reinforcer in this specific case, as even an extreme collision penalty would show immediate extinction as soon as a better alternative outcome is available because the value of the state would get updated immediately to the value of the better outcome). This is caused by two factors: first, the assumption of complete control in the update function as explained above; second, the learning rate  $\alpha$  is 1, meaning that  $V(s_{t-1})$  is set to  $r_t + V(s_t)$  at once erasing the old value. In this paper we extend our previous work (Jacobs et al., 2014a) by showing the effect of

update functions on fear behavior. The habituation of fear was demonstrated in an experiment with punished non-goal arms in the maze for two different update functions. The more balanced the value updates, the longer it takes for fear to extinct, and the higher average fear is for the population of agents. This is consistent with the literature on hope and optimism. Optimist and high-hope individuals experience less fear and recover quicker from obstacles when working towards a goal (Carver et al., 2010; Snyder, 2002). Our work thus predicts that optimism/high-hope personalities could in fact be using a value update function more close to  $MAX_A$  during learning (and perhaps also during internal simulation of behavior), while realists (or pessimists) would use a value update function that weighs outcomes based on probabilities that are more evenly distributed (see also (Broekens & Baarslag, 2014) for a more elaborate study on optimism and risk taking.

Also in previous work (Jacobs et al., 2014a) we showed that hope did not decrease when adding a wall collision penalty. This is unexpected, as the introduction of risk should at least influence hope to some extent. We argued that this was also due to the use of a  $MAX_a$  update function, since bad actions are simply not taken into account in  $V(s)$ . In this paper we studied the effect of update function (an operationalization of high- versus low-hope personality) and penalties on hope. We were able to show that only low-hope agents suffered from penalties (in the form of punished non-goal maze arms), while high-hope agents in fact benefited from such penalties because the penalties push the agent away from the non-goal arms while not harming the value of states along the path to the goal. As mentioned, this replicates psychological findings nicely.

Our high-hope / low-hope experiments also show that it is very important to vary RL parameters and methods in order to understand the relation to human (and animal) emotions. It is probably *not* the case that best practice RL methods (i.e., those that perform best from a learning point of view) also best predict human emotion data. In our case, it is clear that one has to take into account the level of optimism of the individual and correctly control for this using, in our case, the way value updates are performed. Again, this shows we need to better understand how task/learning parameters, including reward shaping, update function, discount factor, action-selection and learning rate, influence the resulting policy and emotion dynamics and how this relates to human dispositional characteristics (e.g., high-hope relates to  $MAX_A$  updates). It is important that future research focuses on drawing a correct parallel between parameters of Reinforcement Learning and human behavior. An understanding of the effect of each of these parameters allows us to construct more thorough hypotheses as well as benchmark tests to study emotion in this context. Comparing the results from a computational simulation to emotions expressed by human subjects in a similar setting (Gratch et al., 2009) is essential, also to further our understanding of RL-based models of emotion.

We mentioned that this type of computational modeling of emotion can help human robot interaction by making more explicit what particular feedback signals mean. Our model suggests that if a human observer expresses hope and fear to a learning robot to shape the learning process than this should directly alter the *value* of a state during the learning process (perhaps temporarily, but in any case the signal should not be interpreted as a reward signal). The expression of joy and distress should alter the robot's current TD *error*, and also not be interpreted as a reward signal. In a similar way the model suggests that if a robot expresses its current state during the learning process to a human observer in order to give this observer insight into how the robot is doing, than hope and fear should be used to express the value of the current state and joy and distress should be used to express the TD error.



## 7. Conclusion

We have proposed a computational model of emotion based on reinforcement learning primitives. We have argued that such a model is useful for the development of adaptive agents and human interaction therewith. We model joy/distress based on the error signal weighted by the unexpectedness of the outcome, and model hope/fear as the learned value of the current state. We have shown experimentally that our model replicates important properties of emotion dynamics in humans, including habituation of joy, extinction of fear, the occurrence of hope and fear after joy and distress, and "low-hope" agents having more trouble learning than "high-hope" agents in the presence of punishments around a rewarded goal. We conclude that it is plausible to assume that the emotions of hope, fear, joy and distress can be mapped to RL-based signals.

However, we have also pointed out the current limitations of testing the validity of models of emotion based on RL-based signals. These limitations include: the absence of standard scenarios (learning tasks); the effect of learning parameters and methods on emotion dynamics; the influence of learning approaches on the availability of particular signals; and, last but not least, psychological ramifications of the assumption that certain emotions can be mapped to simple RL-based signals.

## References

- Baumeister, R. F., Vohs, K. D., DeWall, C. N., & Zhang, L. (2007). How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review*, *11*(2), 167–203.
- Berridge, K. C. (2003). Pleasures of the brain. *Brain and Cognition*, *52*(1), 106–128.
- Berridge, K. C., & Robinson, T. E. (2003). Parsing reward. *Trends in Neurosciences*, *26*(9), 507–513.
- Berridge, K. C., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward: liking, wanting, and learning. *Current Opinion in Pharmacology*, *9*(1), 65–73.
- Bogacz, R., & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Computation*, *19*(2), 442–477.
- Bottan, N. L., & Perez Truglia, R. (2011). Deconstructing the hedonic treadmill: Is happiness autoregressive? *Journal of Socio-Economics*, *40*(3), 224–236.
- Brickman, P., Coates, D., & Janoff-Bulman, R. (1978). Lottery winners and accident victims: is happiness relative? *Journal of personality and social psychology*, *36*(8), 917.
- Broekens, J. (2007). *Affect and learning: a computational analysis* (PhD thesis). Leiden University.
- Broekens, J., & Baarslag, T. (2014). Optimistic risk perception in the temporal difference error explains the relation between risk-taking, gambling, sensation-seeking and low fear. *arXiv:1404.2078*.
- Broekens, J., Kusters, W., & Verbeek, F. (2007). On affect and self-adaptation: Potential benefits of valence-controlled action-selection. In *Bio-inspired modeling of cognitive tasks* (p. 357–366).
- Broekens, J., Marsella, S., & Bosse, T. (2013). Challenges in computational modeling of affective processes. *IEEE Transactions on Affective Computing*, *4*(3).

- Bryson, J. J. (2007). Mechanisms of action selection: Introduction to the special issue. *Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems*, 15(1), 5-8.
- Canamero, L. (2005). Emotion understanding from the perspective of autonomous robots research. *Neural networks*, 18(4), 445-455.
- Carver, C., Scheier, M., & Segerstrom, S. (2010). Optimism. *Clinical Psychology Review*, 30(7), 879.
- Cos, I., Canamero, L., Hayes, G. M., & Gillies, A. (2013). Hedonic value: enhancing adaptation for motivated agents. *Adaptive Behavior*, 21(6), 465-483.
- Critchfield, T. S., Paletz, E. M., MacAleese, K. R., & Newland, M. C. (2003). Punishment in human choice: Direct or competitive suppression? *Journal of the Experimental analysis of Behavior*, 80(1), 1-27.
- Damasio, A. R. (1996). *Descartes' error: emotion reason and the human brain*. Penguin Putnam.
- Dayan, P., & Balleine, B. W. (2002). Reward, motivation, and reinforcement learning. *Neuron*, 36(2), 285-298.
- Dias, J., Mascarenhas, S., & Paiva, A. (2011). Fatima modular: Towards an agent architecture with a generic appraisal framework. In *Proceedings of the international workshop on standards for emotion modeling*.
- Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, 15(4), 495-506.
- El-Nasr, M. S., Yen, J., & Ioerger, T. R. (2000). Flame: fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-agent systems*, 3(3), 219-257.
- Foa, E. B., & Kozak, M. J. (1986). Emotional processing of fear: exposure to corrective information. *Psychological bulletin*, 99(1), 20.
- Frijda, N., Kuipers, P., & Ter Schure, E. (1989). Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, 57(2), 212.
- Frijda, N. H. (2004). Emotions and action. In A. S. R. Manstead & N. H. Frijda (Eds.), *Feelings and emotions: the amsterdam symposium* (p. 158-173). Cambridge University Press. (Feelings and emotions: The Amsterdam symposium)
- Gadanhó, S. C. (1999). *Reinforcement learning in autonomous robots: An empirical investigation of the role of emotions*. (Unpublished doctoral dissertation). University of Edinburgh.
- Gratch, J., Marsella, S., Wang, N., & Stankovic, B. (2009). Assessing the validity of appraisal-based models of emotion. In *3rd international conference on affective computing and intelligent interaction and workshops, 2009. acii 2009*.
- Haruno, M., & Kawato, M. (2006). Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *Journal of neurophysiology*, 95(2), 948-959.
- Hogewoning, E., Broekens, J., Eggermont, J., & Bovenkamp, E. (2007). Strategies for affect-controlled action-selection in soar-rl. *Nature Inspired Problem-Solving Methods in Knowledge Engineering*, 501-510.
- Holroyd, C., & Coles, M. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological review*, 109(4), 679.
- Houk, J., Bastianen, C., Fansler, D., Fishbach, A., Fraser, D., Reber, P., . . . Simo, L. (2007). Action selection and refinement in subcortical loops through basal ganglia and cerebellum. *Philosophical Transactions of the Royal Society B*:

- Biological Sciences*, 362(1485), 1573-1583.
- Jacobs, E., Broekens, J., & Jonker, C. (2014a). Emergent dynamics of joy, distress, hope and fear in reinforcement learning agents. In S. Barrett, S. Devlin, & D. Hennes (Eds.), *Adaptive learning agents workshop at aamas2014*.
- Jacobs, E., Broekens, J., & Jonker, C. (2014b). Joy, distress, hope and fear in reinforcement learning. In A. Bazzan, M. Huhns, A. Lomuscio, & P. Scerri (Eds.), *Proceedings of the 2014 international conference on autonomous agents and multi-agent systems* (p. 1615-1616). International Foundation for Autonomous Agents and Multiagent Systems.
- Kiryazov, K., Lowe, R., Becker-Asano, C., & Ziemke, T. (2011). *Modelling embodied appraisal in humanoids: Grounding pad space for augmented autonomy*.
- Krichmar, J. L. (2008). The neuromodulatory system: A framework for survival and adaptive behavior in a challenging world. *Adaptive Behavior*, 16(6), 385-399.
- Lahnstein, M. (2005). The emotive episode is a composition of anticipatory and reactive evaluations. In L. Caamero (Ed.), *Agents that want and like: Motivational and emotional roots of cognition and action. papers from the aisb'05 symposium* (p. 62-69). AISB Press.
- Lewis, M. D. (2005). Bridging emotion theory and neurobiology through dynamic systems modeling. *Behavioral and Brain Sciences*, 28(02), 169-194.
- Marinier, R., & Laird, J. E. (2008). Emotion-driven reinforcement learning. In *Cognitive science* (p. 115-120).
- Marsella, S., & Gratch, J. (2009). Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1), 70-90.
- Marsella, S., Gratch, J., & Petta, P. (2010). Computational models of emotion. *K. r. Scherer, t. Bnziger and e. roesch (eds.), A blueprint for affective computing*, 21-45.
- Montague, P. R., King-Casas, B., & Cohen, J. D. (2006). Imaging valuation models in human choice. *Annu. Rev. Neurosci.*, 29, 417-448.
- Moren, J. (2000). A computational model of emotional learning in the amygdala. In *From animals to animats* (Vol. 6, p. 383-391).
- Myers, K. M., & Davis, M. (2006). Mechanisms of fear extinction. *Mol Psychiatry*, 12(2), 120-150.
- O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Current opinion in neurobiology*, 14(6), 769-776.
- Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge University Press.
- Prescott, T. J., Bryson, J. J., & Seth, A. K. (2007). Introduction. modelling natural action selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1485), 1521-1529.
- Redish, A., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological review*, 114(3), 784.
- Reisenzein, R. (2009). Emotional experience in the computational belief-desire theory of emotion. *Emotion Review*, 1(3), 214-222.
- Robinson, P., & el Kaliouby, R. (2009). Computation of emotions in man and machines. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535), 3441-3447.
- Rolls, E. T. (2000). Precis of the brain and emotion. *Behavioral and Brain Sciences*, 20, 177-234.

- Rolls, E. T., & Grabenhorst, F. (2008). The orbitofrontal cortex and beyond: From affect to decision-making. *Progress in Neurobiology*, *86*(3), 216-244.
- Schaffer, H. (1974). Cognitive components of the infant's response to strangeness. *The origins of fear*, *2*, 11.
- Scherer, K. (2001). Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, *92*, 120.
- Scherer, K. R. (2009). Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3459-3474.
- Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, *80*(1), 1-27.
- Schweighofer, N., & Doya, K. (2003). Meta-learning in reinforcement learning. *Neural networks*, *16*(1), 5-9.
- Sequeira, P. (2013a). *Socio-emotional reward design for intrinsically motivated learning agents* (Unpublished doctoral dissertation). Universidade Tecnica de Lisboa.
- Sequeira, P. (2013b). *Socio-emotional reward design for intrinsically motivated learning agents* (Unpublished doctoral dissertation). Universidade Tecnica de Lisboa.
- Sequeira, P., Melo, F., & Paiva, A. (2011). Emotion-based intrinsic motivation for reinforcement learning agents. In S. DMello, A. Graesser, B. Schuller, & J.-C. Martin (Eds.), *Affective computing and intelligent interaction* (Vol. 6974, p. 326-336). Springer Berlin Heidelberg.
- Sequeira, P., Melo, F., & Paiva, A. (2014). Emergence of emotional appraisal signals in reinforcement learning agents. *Autonomous Agents and Multi-Agent Systems*, in press.
- Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, *69*(1), 99-118.
- Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, *63*(2), 129-138.
- Snyder, C. R. (2002). Hope theory: Rainbows in the mind. *Psychological Inquiry*, *13*(4), 249-275.
- Sprott, J. (2005). Dynamical models of happiness. *Nonlinear Dynamics, Psychology, and Life Sciences*, *9*(1), 23-36.
- Sroufe, L. A. (1997). *Emotional development: The organization of emotional life in the early years*. Cambridge University Press.
- Stephen, J. E. (2013). Hed: A computational model of affective adaptation and emotion dynamics. *IEEE Transactions on Affective Computing*, *4*(2), 197-210.
- Suri, R. E. (2002). Td models of reward predictive responses in dopamine neurons. *Neural networks*, *15*(4-6), 523-533.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). Cambridge Univ Press.
- Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat Neurosci*, *7*(8), 887-893.
- Veenhoven, R. (1991). Is happiness relative? *Social Indicators Research*, *24*(1), 1-34.
- Zajonc, R. B., & Markus, H. (1982). Affective and cognitive factors in preferences. *Journal of Consumer Research*, *9*(2), 123.