Emergent Dynamics of Joy, Distress, Hope and Fear in Reinforcement Learning Agents

Elmer Jacobs
Interactive Intelligence, TU Delft
Delft, The Netherlands
elmer.j.jacobs@gmail.com

Joost Broekens
Interactive Intelligence, TU Delft
Delft, The Netherlands
joost.broekens@gmail.com

Catholijn Jonker
Interactive Intelligence, TU Delft
Delft, The Netherlands
C.M.Jonker@tudelft.nl

ABSTRACT

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. Joy/distress is a signal that is derived from the RL update signal, while hope/fear is derived from the utility of the current state. Agent-based simulation experiments replicate psychological and behavioral dynamics of emotion including: joy and distress reactions that develop prior to hope and fear; fear extinction; habituation of joy; and, task randomness that increases the intensity of joy and distress. This work distinguishes itself by assessing the dynamics of emotion in an adaptive agent framework - coupling it to the literature on habituation, development, and extinction. Our results support the idea that the function of emotion is to provide a complex feedback signal for an organism to adapt its behavior. We show this feedback signal can be operationalized for RL agents. This is important because (a) RL-based models can help understand the relation between emotion and adaptation in animals, (b) the emotional state might be used to increase adaptive potential, and (c) expression of an emotion to a human observer that it is grounded in the learning mechanism of the agent should help interpret the meaning of the emotion.

Categories and Subject Descriptors

I.2.6 [**Computing Methodologies/Artificial Intelligence**]: Learning

General Terms

Human Factors

Keywords

Reinforcement Learning, Emotion Dynamics, Affective computing

1. INTRODUCTION

Emotion and reinforcement learning play an important role in shaping behaviour. Emotions drive adaptation in behaviour and are therefore often coupled to learning [1]. Further, emotions inform us about the value of alternative actions [9] and directly influence action selection, for example through action readiness [15]. Reinforcement Learning (RL) [45] 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 [10, 26, 44]. 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 [31]).

While most research on computational modeling of emotion is based on cognitive appraisal theory [22], above mentioned similarities have inspired computational studies into how emotion-like signals influence RL to create an adaptive benefit for the agent [18, 36, 16, 35, 38, 43]. For example, it has been shown that emotion-like signals can emerge as part of the intrinsic reward function[36], that emotion-like signals can function as metalearning parameters [18, 35], and that emotional signals coming from others [5] or coming from a cognitive assesment of the agent itself [16] can provide additional reward information for an RL learner. Other work can be considered 'in between' RL and cognitive emotion models because it either looks at the emotion-cognition-RL relation [21], or because it explicitly looks at utility (or payoff) models for emotion intensity [17, 23]. Our work is different in that we aim to show a direct mapping between RL primitives and emotions, and assess the validity by replicating psychological findings on emotion dynamics, the latter being an essential difference with [12]. 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.

There are many reasons for wanting a valid affective labeling of RL-related signals. From a theoretical point of view, the function of emotions is to provide complex feedback signals aimed at informing the agent about the current state of affairs during learning and adaptation [14, 20, 27, 28, 29, 30, 6]. What do such signals look like in an adaptive agent? If we can operationalize such signals for RL agents, a popular computational model for reward-based learning in animals [10, 26], we can computationally tie emotion to adaptation. From a learning optimization point of view, correct labeling will further our understanding of how emotions can increase the agent's adaptive potential [18, 36, 16, 35, 38, 43]. From a human-robot interaction point of view the emotional signal can be expressed to a human observer. If this signal is grounded in the learning mechanism of the agent it should

help interpret the meaning of the expressed emotion[7].

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 [27], emotion development[42], and habituation fear extinction [3, 4, 46, 25]. Using agent-based simulation where an RL-based agent collects rewards in a maze, we show that the emerging emotion dynamics are consistent with this 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 [45]. 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 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 a final time step, but can become infinite in applications where such an end state 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, such that:

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

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. In these types of problems, two additional elements are available: the transition probability $P^a_{ss'}$ and the expected reward $R^a_{ss'}$.

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

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

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:

$$V_{k+1}(s) = \max_a \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V_k(s') \right]. \quad (5)$$

After convergence of the values, the policy simplifies to:

$$\pi(s) = \arg\max_a \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V(s') \right] \quad (6)$$

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 prposed as a plausible model for human learning based on feedback [34, 19, 11, 2]. The simplest method, one-step Temporal Difference Learning, updates values according to:

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

with α now representing the learning rate. After convergence, the values can be used to determine actions. Several types of action selection exist, from completely random to simply choosing the action resulting in the maximum predicted value. The Boltzmann distribution, argued to be a model for human action selection [8, 24, 39, 40], gives a probability of choosing each action and is given by:

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

where β is a positive parameter called inverse temperature and $Q(s,a)$ is the value of taking a specific action according to:

$$Q(s,a) = \sum_{s'} P^a_{ss'} \left[V(s') + R^a_{ss'} \right]. \quad (9)$$

3. MAPPING EMOTIONS

In essence, the computational model of emotion we propose is a mapping between RL primitives (reward, value, update 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[27], a well-known computation-ready psychological model of cognitive emotion elicitation. We now detail the rationale for our mapping.

3.1 Emotional development and habituation

Learning not only drives adaptation in human behaviour, 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 responsivity to the environment [42]. 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 fearfulness, usually reported first at 7 or 8 months. Fearfulness requires a comparison of multiple

events [32] 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 [25, 3, 4, 46, 13]. A mapping of RL primitives to emotion should be consistent with habituation and extincition, and in particular fear extinction as this is a well studied phenomenon [25].

3.2 Mapping joy and distress

Joy and distress are the first emotions to be observable in infants. A first possible choice to map joy would be to use the reward r_t. Any state transition that yields some reward therefore causes joy in the agent. However, anticipation and unexpected improvement can result in joy [41] 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 expectation [27]. We take care of this by adding an unexpectedness term, derived from the expected probability of the state-transition that just took place, which is $(1 - P^{a_{t-1}}_{s_{t-1}s_t})$. We let:

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

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 [27].

3.3 Mapping hope and fear

According to theory about emotional development, joy and distress are followed by anger and fearfullness. Hope is the anticipation of a positive outcome and fear the anticipation of a negative outcome [27]. 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 $P^{a_t}_{s_t s_{t+1}}$. 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 decision may be to use $V(s_t)$ as the hope and fear representation. Under this mapping, fear extinction can happen by a mechanism similar to *new learning* [25]. 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.

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 ($HF > 0$) and intenstity of fear ($HF < 0$) equals to:

$$HF(s_t) = V(s_t) \quad (11)$$

The OCC model states that hope and fear are dependent on the expected joy/distress and likelihood of a future event [27], which is again consistent with our mapping.

4. VALIDATION

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 first state four requirements based on habituation, development and extinction literature.

REQUIREMENT 1. *In all simulations, joy/distress is the first emotion to be observed followed by hope/fear. As mentioned earlier, human emotions have an order in their developent in individuals from simple to complex [42].*

REQUIREMENT 2. *Simulations should show joy habituation and fear extinction over time. Habituation of joy when the agent is presented repeatedly to the same reinforcement [3, 4, 46, 13] should be observed. as well as fear extinction through the mechanism explained above [25]. In our instrumentation, if joy and fear show this behavior, then distress and hope will also show this as these are symetrical, so we test only for joy and fear.*

REQUIREMENT 3. *Lowered expectation decreases hope and results in a higher intensity for joy/distress [46, 27]. So, when we lower the expectation of an agent of its environment (by adding a wall collision penalty), we should observe an increase in its joy reaction to positively reinforced situations as well as a decrease in hope due to lower expectation.*

REQUIREMENT 4. *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 [27, 33].*

4.1 Experimental setup

We ran our validation tests in an agent-based simulation implemented in Java. A 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.

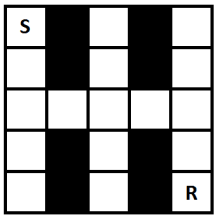

Figure 1: The maze used in the experiments. Each square is a state. The agent starts at S and the reward can (initially) be found at R.

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

	Control setting	Variations
Collision penalty	0.0	0.01
P(remain in place)	0.1	0.25
Reward action	Return to start	Relocate reward

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 γ equals 0.9. To test the effect of expectation and predictability of the world, we varied several parameters of the task (see Table 1). To manipulate an agent's high versus low expectation of the task, we varied collision penalty (the reward for bumping into a wall was 0 or -0.01, representing a less favorable world in the latter case). To manipulate predictability we varied 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). A simulation consists of a population of 50 different agents, each agent runs the task once for a total of 10000 steps, which appeared in pre-testing to be long enough to show policy convergence. To reduce the probability that our results are prodcued by a 'lucky parameter setting', each run has gaussian noise over the parameter values for β, γ, the collision penalty 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, with the exception of the case where the collision penalty is 0.

The mappings from RL primitives to emotions as defined in the emotional model require knowledge of 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:

$$V(s_t) \longleftarrow \max_a \sum_s P^a_{ss'} \left[R^a_{ss'} + \gamma V(s')\right]. \qquad (12)$$

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 Habituation, development, fear extinction

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 a reaction to the update of the Value function in combination with 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 state value updates, 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 value update. Remember that joy/distress is derived from the change in value *and* the unexpectedness of that update. 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 succesful transitions toward higher valued states (and these continue to occur, as the discount factor is non-zero), while the negative spikes are transitions toward lower valued states, both with low intensities caused by high expectedness. 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 [27]. Habituation of joy/distress is also confirmed: the intensity of joy goes down each time the reward is gathered.

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). We explained earlier that state values can only be updated once a reward is gained, and therefore hope and fear can only emerge *after* the agent has some expected gain/loss as represented by the values of states. This order of emergence becomes clear from the simulation.

To test for the occurance of fear extinstion, we introduce a wall collision penalty. We ran a novel simulation (50 agents, 10000 steps) with the same settings as the previous simulation except for the collision penalty (see Table 1). We plot the average intenisty of fear for 50 agents in Figure 5. Fear is caused by running into the wall at the beginning of the simulation. Fear goes to 0 after a small number of steps. However, the agent always has options available that never cause it to collide. Because of the max function in Equation 12, these types of actions take precedence in the value updates. Using this update function, if any action is available

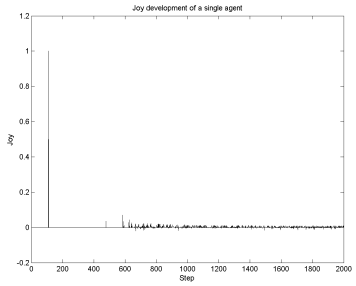

Figure 2: Intensity of joy/distress for a single agent, observed in the first 2000 steps

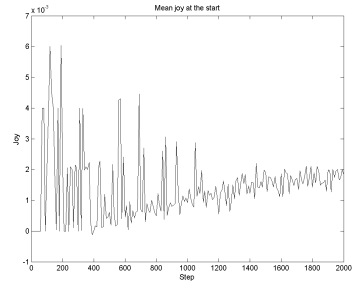

Figure 3: Intensity of joy/distress, mean over 50 agents, observed in the first 2000 steps

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 once it knows how to avoid penalties. The agent updates its utility of the state and the initial fear associated with that state extincts to 0. This demonstrates a know mechanism for the habituation of fear, called *new learning* [25]). New learning explains fear extinctions 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. Our mode thus replicates findings on fear extinction.

5.2 Lowered expectation

To test our requirement (in this case more so a hypothe-

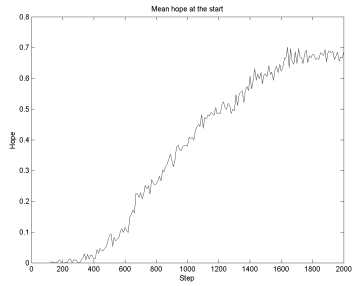

Figure 4: Intensity of hope, mean over 50 agents, observed in the first 2000 steps

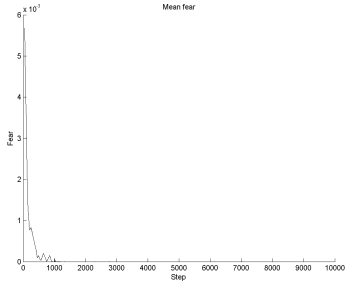

Figure 5: Intensity of fear, mean over 50 agents, in the presence of a collision penalty

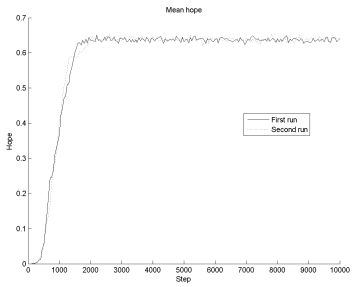

Figure 6: Intensity of hope, mean over 50 agents, without (first run) and with (second) collision penalty

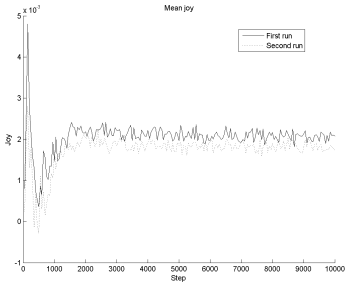

Figure 7: Intensity of joy/distress, mean over 50 agents, without (first run) and with (second) collision penalty

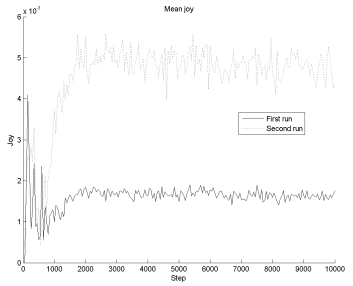

Figure 8: Intensity of joy/distress, mean over 50 agents, with a probability of 0.1 (first run) and 0.25 (second run) of failing an action

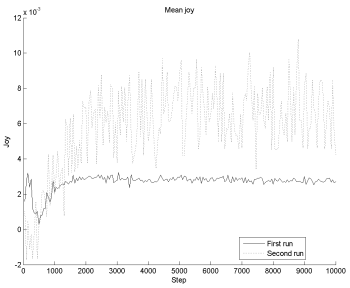

Figure 9: Intensity of joy/distress, mean over 50 agents, returning the agent (first run) or relocating the reward (second run)

sis) that lowered expectation decreases hope and results in a higher intensity for joy, we contrast the results of a task in which a wall collision penalty is present against a task in which this penalty is not present (see Table 1). Adding a collision penalty lowers the expectations in any state, since a chance exists for making such a punished move. The mean hope and joy over 50 agents with and without collison penalty are shown in Figures 6 and 7. The intensity of hope is almost unaffected by the collision penalty, while joy decreases a bit. At first sight, this seems to falsify our hypothesis, which stated that the collision penalty would result in a decrease in expectation resulting in an increase in joy intensity and a decrease in hope. Upon further 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 12 results in the agent not considering the possibility for a penalty in the updates of values. Very quickly, the penalty is not represented anymore in the state values. Secondly, the Boltzmann action selection method selects actions with potentially negative reward less often than neutral actions, resulting in faster convergence toward the optimal policy. As a result, state values are approximately the same in both conditions. Our manipulation of expectation is not adequate to manipulate hope and joy. As such, the hypothesis is neither falsified nor supported.

5.3 The effect of unexpectedness

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 three ways to vary predictability, two of which match with unexpectedness. First, we can make action selection a random process. This results in unpredictable behaviour and an inefficient policy, but does not change the predictability of the *effects of an action* once it is chosen. Second, we can make the results of an action stochastic, for example by letting the action fail completely every once in a while. Third, we give rewards at random points rather than at the same transition all the time. This randomizes the reinforcing part of an experiment. The last two options increase the unexpectedness of the result of an action and we test both of them.

First of all, we increased the probability for an action to fail (failure results in no state change). The resulting mean intensity of joy for 50 agents is shown in Figure 8. The second simulation consisted of randomly relocating the reward after each time it was collected, instead of returning the agent to the starting position. The mean intensity of joy for 50 agents is shown in Figure 9. We can see that in both simulations the intensity of joy/distress reactions is larger (bigger 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 $6\%(1/17 states)$. 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 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 [27, 33].

6. DISCUSSION

The speed at which the habituation takes place is high as a result of how unexpectedness is calculated in the model. In the beginning unexpectedness is 0 for many state transitions because the probablity 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. This differs from humans, who in general have certain expectations about alternative outcomes. We can simulate such experiences by starting with a default model of the environment, or by assuming a small default probablity for all states as potential outcome of an action in an arbitrary state (i.e., all $P_{sas'}$ are non-zero) The latter is probably closer to reality.

Technically speaking, joy/distress in our model is not derived from the update 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 and an update function that uses the max_a value to update states then this signal is equivalent to the update signal because the expected value of the any previous state $V(s_{t-1})$ converges to $r_t + V(s_t)$ (the policy ensures this). The main difference then is that a model that is derived from the RL update signal includes a discount factor gamma and a learning rate alpha. Inclusion or exclusion of these variables result in different joy/distress signals. 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 update signal, then joy/distress would become 0 at convergence because the update signal 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 simulations as habituation of joy/distress intensity does not end up at 0. The discount factor ensures that expectations are always smaller than the actual outcome. This translates to humans that still experience a little bit of joy,

even when getting a reward that they have often received. In our model habituation predicts a diminishing intensity but not a complete loss of joy/distress response. Further study should investigate the plausibility of the two alternative models for joy/distress.

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. This is because our agents use the maximum of the possible outcomes to update the value of states, as is common in RL. This means that they are very optimistic, and assume complete control over their actions. Our model predicts that fear extinction rate does not depend on the strength of the negative reinforcer, as even a collision penalty of -1000 would still show extinction as soon as a better alternative outcome is available (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. The influence of update functions and learning rates should be investigated further.

Hope (expectation) did not decrease when adding a collision penalty. This can also be attributed to the use of the best possible outcome in the update function, since bad actions are simply not taken into account in $V(s)$. We were not able to validate hypothesis 3 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 a strong argument for taking utmost care when constructing a task to test hypotheses. All settings for the parameters of the simulation should be considered with respect to their counterparts in the animal world to ensure a correct test, pointng towards the need for standardized benchmark tests. To elaborate a little further on this point, many of the task/learning parameters including the update function, discount factor, action-selection and learning rate, of which the effects were not investigated in this paper, can influence the signals we label as emotions. It is important that future research focuses on drawing a correct parallel between parameters of Reinforcement Learning and human behaviour. An understanding of the effect of each of these parameters allows us to construct more thorough hypotheses as well as benchmark tests. Comparing the results of a simulation to emotions shown by human subjects in a similar setting [17] is essential to further our understanding of the validity of different RL-based models of emotion.

We limited ourselfs to joy/distress and hope/fear. A direct next step would be the mapping of confirmation emotions as these are the consequence of hope/fear. If we assume that confirmation is the percentage or ratio reflecting how much of an expectation has actually come true the next state, then this can be expressed as follows:

$$C(s_{t-1}, a_{t-1}, s_t) = \frac{V(s_t) + r_t}{V(s_{t-1})} P^{a_{t-1}}_{s_{t-1}s_t} \tag{13}$$

This mapping is in accordance with the OCC model [27], which states that confirmation depends on the intensity of the prospect emotion, the degree of realization and unexpectedness of the event. It is mathematically only defined if $V(s_{t-1}) \neq 0$, which is in accordance with reality as having no expectations whatsoever can never result in confirmation.

None of the emotions we modelled involve the appraisal of agency, which is by definition not available in Reinforcement Learning. However, adding agency (e.g., in multi-agent RL), it is not unthinkable that social emotions could emerge from RL primitives. Attempts at using social values in reward construction have been tried recently in a multi-agent RL setting [37]. This would also be a good moment to try to map anger to RL primitives. Validation of such signals should be further investigated.

7. CONCLUSION

We have proposed a computational model of emotion based on reinforcement learning primitives. We derive joy/distress from the RL update signal weighted by the unexpectedness of the outcome, and hope/fear from the learned value of the current state. Our contribution is that we replicate important (cited) properties of emotion dynamics in humans with our model, including habituation of joy, extinction of fear, and the occurance of hope and fear after joy and distress. We conclude that our model is a plausible RL-based instrumentation for joy/distress and hope/fear. However, we are aware of the difficulties of labeling RL-based signals as particular emotions, as discussed. Also, we feel that in general a more structured approach is needed to develop scenarios (tasks/learning approach/RL parameters) to test for the plausibility of affective labeling of RL-based signals.

8. REFERENCES

[1] Roy F Baumeister, Kathleen D Vohs, C Nathan DeWall, and Liqing Zhang. How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review*, 11(2):167–203, 2007.

[2] A. Blokland. Acetylcholine: a neurotransmitter for learning and memory? *Brain Research Reviews*, 21(3):285–300, 1995.

[3] Nicolas Luis Bottan and Ricardo Perez Truglia. Deconstructing the hedonic treadmill: Is happiness autoregressive? *Journal of Socio-Economics*, 40(3):224–236, 2011.

[4] P Brickman, D Coates, R Janoff-Bulman, et al. Lottery winners and accident victims: is happiness relative? *Journal of personality and social psychology*, 36(8):917, 1978.

[5] Joost Broekens. *Affect and Learning: a computational analysis*. Phd thesis, 2007.

[6] Joost Broekens, Stacy Marsella, and Tibor Bosse. Challenges in computational modeling of affective processes. *IEEE Transactions on Affective Computing*, 4(3), 2013.

[7] L. Canamero. Emotion understanding from the perspective of autonomous robots research. *Neural networks*, 18(4):445–455, 2005.

[8] Thomas S Critchfield, Elliott M Paletz, Kenneth R MacAleese, and M Christopher Newland. Punishment in human choice: Direct or competitive suppression? *Journal of the Experimental analysis of Behavior*, 80(1):1–27, 2003.

[9] A. R. Damasio. *Descartes' Error: emotion reason and the human brain*. Penguin Putnam, 1996.

[10] Peter Dayan and Bernard W. Balleine. Reward,

motivation, and reinforcement learning. *Neuron*, 36(2):285–298, 2002.

[11] K. Doya. Metalearning and neuromodulation. *Neural Networks*, 15(4):495–506, 2002.

[12] Magy Seif El-Nasr, John Yen, and Thomas R Ioerger. Flame: fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-agent systems*, 3(3):219–257, 2000.

[13] Edna B Foa and Michael J Kozak. Emotional processing of fear: exposure to corrective information. *Psychological bulletin*, 99(1):20, 1986.

[14] N. H. Frijda. *Emotions and action*, pages 158–173. Cambridge University Press, 2004. Feelings and emotions: The Amsterdam symposium.

[15] N.H. Frijda, P. Kuipers, and E. Ter Schure. Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, 57(2):212, 1989.

[16] Sandra Clara Gadanho. *Reinforcement Learning in Autonomous Robots: An Empirical Investigation of the Role of Emotions.* PhD thesis, 1999.

[17] Jonathan Gratch, Stacy Marsella, Ning Wang, and Brooke Stankovic. Assessing the validity of appraisal-based models of emotion. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–8. IEEE, 2009.

[18] 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*, pages 501–510, 2007.

[19] C.B. Holroyd and M.G.H. Coles. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological review*, 109(4):679, 2002.

[20] Marc D. Lewis. Bridging emotion theory and neurobiology through dynamic systems modeling. *Behavioral and Brain Sciences*, 28(02):169–194, 2005.

[21] RP Marinier and John E Laird. Emotion-driven reinforcement learning. *Cognitive Science*, pages 115–120, 2008.

[22] Stacy Marsella, Jonathan Gratch, and Paolo Petta. Computational models of emotion. *K. r. Scherer, t. BÃd'nziger and e. roesch (eds.), A blueprint for affective computing*, pages 21–45, 2010.

[23] Stacy C. Marsella and Jonathan Gratch. Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90, 2009. doi: DOI: 10.1016/j.cogsys.2008.03.005.

[24] P Read Montague, Brooks King-Casas, and Jonathan D Cohen. Imaging valuation models in human choice. *Annu. Rev. Neurosci.*, 29:417–448, 2006.

[25] K. M. Myers and M. Davis. Mechanisms of fear extinction. *Mol Psychiatry*, 12(2):120–150, 2006.

[26] John P. O'Doherty. Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Current opinion in neurobiology*, 14(6):769–776, 2004.

[27] Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.

[28] Rainer Reisenzein. Emotional experience in the computational belief-desire theory of emotion. *Emotion Review*, 1(3):214–222, 2009.

[29] Peter Robinson and Rana el Kaliouby. Computation of emotions in man and machines. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3441–3447, 2009.

[30] Edmund T. Rolls. Precis of the brain and emotion. *Behavioral and Brain Sciences*, 20:177–234, 2000.

[31] Edmund T. Rolls and Fabian Grabenhorst. The orbitofrontal cortex and beyond: From affect to decision-making. *Progress in Neurobiology*, 86(3):216–244, 2008.

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

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

[34] W. Schultz. Predictive reward signal of dopamine neurons. *Journal of neurophysiology*, 80(1):1–27, 1998.

[35] N. Schweighofer and K. Doya. Meta-learning in reinforcement learning. *Neural Networks*, 16(1):5–9, 2003.

[36] PDB Sequeira. *Socio-Emotional Reward Design for Intrinsically Motivated Learning Agents*. PhD thesis, Universidada TÏ£¡cnica de Lisboa, 2013.

[37] Pedro Sequeira. *Socio-Emotional Reward Design for Intrinsically Motivated Learning Agents*. PhD thesis, 2013.

[38] Pedro Sequeira, FranciscoS Melo, and Ana Paiva. *Emotion-Based Intrinsic Motivation for Reinforcement Learning Agents*, volume 6974 of *Lecture Notes in Computer Science*, chapter 36, pages 326–336. Springer Berlin Heidelberg, 2011.

[39] Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118, 1955.

[40] Herbert A Simon. Rational choice and the structure of the environment. *Psychological review*, 63(2):129–138, 1956.

[41] JC Sprott. Dynamical models of happiness. *Nonlinear Dynamics, Psychology, and Life Sciences*, 9(1):23–36, 2005.

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

[43] John E. Steephen. Hed: A computational model of affective adaptation and emotion dynamics. *IEEE Transactions on Affective Computing*, 4(2):197–210, 2013.

[44] Roland E. Suri. Td models of reward predictive responses in dopamine neurons. *Neural networks*, 15(4-6):523–533, 2002.

[45] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. Cambridge Univ Press, 1998.

[46] Ruut Veenhoven. Is happiness relative? *Social Indicators Research*, 24(1):1–34, 1991.