A TDRL Model for the Emotion of Regret

1st Joost Broekens
Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Leiden, The Netherlands
joost.broekens@gmail.com

2nd Laduona Dai
Human Media Interaction
University of Twente
Enschede, The Netherlands
laduona.dai@gmail.com

***Abstract*—To better understand the nature, function and elicitation conditions of emotion it is important to approach studying emotion from a multidisciplinary perspective involving psychology, neuroscience and affective computing. Recently, the TDRL Theory of Emotion has been proposed. It defines emotions as variations of temporal difference assessments in reinforcement learning. In this paper we present new evidence for this theory. We show that regret - a negative emotion that signifies that an alternative action should have been taken given new outcome evidence - is modelled by a particular form of TD error assessment. In our model regret is attributed to each action in the state-action trace of an agent for which - after new reward evidence - an alternative action becomes the best action in that state (the new argmax) after adjusting the action value of the chosen action in that state. Regret intensity is modeled as the difference between this new best action and the adjusted old best action, reflecting the additional amount of return that could have been received should that alternative have been chose. We show in simulation experiments how regret varies depending on the amount of adjustment as well as the adjustment mechanism, i.e. Q-trace, Sarsa-trace, and Monte Carlo (MC) re-evaluation of action values. Our work shows plausible regret attribution to actions, when this model of regret is coupled with MC action value update. This is important evidence that regret can be seen as a particular variation of TD error assessment involving counterfactual thinking.**

***Index Terms*—Emotion, TDRL Emotion Theory, Computational Modelling, Regret**

I. Introduction

Emotion is a multifaceted phenomenon involving a relation between action, motivation, expression, information processing, feelings and social interaction [1]–[4]. Emotion is (the result of) the process of assessing the personal relevance of a situation (appraisal) thereby providing feedback on the suitability of past, current and future behavior [5]–[7]. This sets emotion apart from mood, which is a longer-term transient affective state often not attributed to a particular situation or cause, and from affective attitude, which is an existing affective association with a particular stimulus or situation not necessarily involving appraisal.

To better understand the nature, function and elicitation conditions of emotion it is important to study emotion from a multidisciplinary perspective involving psychology, neuroscience and affective computing. To highlight this interdisciplinary importance, consider the recent proposal, inspired by computational modelling and agent interaction experiments, that facial expressions communicate appraisal [8]. Consider also recent work in facial expression analysis and synthesis proposing a more detailed view on cultural universality of basic emotions [9]. Finally, consider the work in appraisal theory emphasizing the role of computational modelling for better understanding of the structure and processes involved in emotion elicitation [10]–[13].

In line with this interdisciplinary approach to understand emotion, and building on initial ideas by Brown and Wagner [14] and Redish [15], and extending ideas of Baumeister [5] and Rolls [16], and work on intrinsic motivation [17], the TDRL (temporal difference reinforcement learning) Theory of Emotion [18], [19] proposes that all emotions are manifestations of neural temporal difference assessment. The theory argues that emotion shares its essential elements - event triggered, feedback providing, action tendency related, valenced experience, and grounded in primary reinforcers - with the assessment of temporal difference errors [18]. Simulation results show that the TDRL theory of Emotion can replicate plausible elicitation conditions and dynamics of joy, distress [20], hope and fear [21].

Reinforcement Learning (RL) [22], [23] is a well-established computational technique enabling agents to learn skills by trial and error. Although there is still considerable work to be done on the sampling efficiency of in particular the exploration process of RL [24], a recent survey [25] reviewed a large variety of robot tasks that can be learned using RL, including walking, navigation, table tennis, and industrial arm control. The learning of a skill in RL is mainly determined by a feedback signal, called the reward, r. In contrast to the definition of reward in the psychological conditioning literature where a negative "reward" is referred to as punishment, reward in RL can be positive or negative. In RL cumulative reward is also referred to as $return$. Through trial and error, a robot or virtual agent adjusts its estimates of action values so that they reflect the expected cumulative future reward. The adjustments are referred to as temporal difference errors (below we provide a slightly more formal definition). RL has been argued to be a neurologically plausible mechanisms for task learning [26], [27]

Explaining emotions as a result of TD error assessment is important as it provides a simple basis for the elicitation mechanisms for different emotions during learning [18], it provides a way to bridge emotion elicitation in task learning with cognitive appraisal theories [28], and it provides a com-

978-1-7281-3888-6/19/$31.00

putational approach towards understanding the role of action values and goal-directed processes in the causation of emotion [29]. On top of that it has been proposed as a method to enhance transparency of the learning process in interactive learning agents and robots [19].

In this paper we present new evidence for the TDRL Theory of Emotion. We show it can be used to model the elicitation mechanisms for regret. We show that regret, classically defined as "a comparison between the outcome of a choice (reality) and the better outcome of foregone rejected alternatives (what might have been)" [30] is modelled by a particular form of TD error assessment in reinforcement learning agents. We propose that regret is attributed to each action in the state-action trace for which - after new reward evidence - an alternative action becomes the best action in that state (the new argmax) after adjusting the action value of the chosen action in that state. Regret intensity is modeled as the difference between this new best action and the adjusted old best action, reflecting the additional amount of return that could have been received should that alternative have been chosen. We show in simulation experiments how regret varies depending on the amount of adjustment as well as adjustment mechanism (Q-trace, Sarsa-trace, MC evaluation of actions).

II. Related Work

There is a lot of related work on computational modelling of emotion. For work on computational modelling of appraisal processes we refer to the recent review by Gratch and Marsella [31]. For work on computational modelling of emotion in reinforcement learning agents we refer to a recent review by Moerland and others [32]. In this related work section we focus on regret and computational modelling thereof.

Regret is an emotion that results from the realization that a decision turned out worse than expected [30]. Regret is generally considered to be a more complex emotion than joy and sadness. According to a study by Guttentag and Ferrell [33] children might feel regret at about the age of 7 since it requires the ability to imagine and compare different outcomes (counterfactuals). Regret has been shown to exist in apes as well [34], indicating the link between regret and re-evaluation of decision outcome. Regret needs the cognitive ability to remember actions and action values, and attribute a change in outcome to an action taken in the past. This requires sophisticated neural machinery, involving processing in the (medial) orbito-frontal cortex and the amygdala [35]. This is evidence for the link between neural reward (re)processing structures, emotion, and behavior modification, and reflective thought. In fact, there is accumulating evidence that neural centers (most notably the amygdala and the orbito- and pre-frontal cortex) are involved in both RL-like reward processing and emotion processing (see [18] for a short overview).

To better understand the elicitation conditions and processes involved in regret, we propose to computationally model it based on TD error processing. Computational modelling enables detailed variation of variables, that are otherwise difficult to vary in vivo. For example, different value update mechanism can be used to incorporate important TD errors, enabling the investigation of what kind of value update mechanisms would be minimally needed for the replication of a plausible human regret signal.

Others have also computationally modelled regret. For example in [36] regret (the form we focus on, i.e., regretting an action you did given a disappointing outcome of that action) is proposed as an assessment of the difference between the utility of the actual outcome and the anticipated outcome. So, from a TD perspective that would mean that regret is a negative TD error attributed to past actions. In our model we propose instead that there is a threshold one needs to get over before regret is elicited due to a negative TD error: regret only arises when a new action becomes the best action after updating action values due to the negative TD. In our model regret is felt when a new action should have been taken given new outcome evidence. This makes more sense from an emotion perspective. "Regret embodies the painful lesson that things would have been better under a different choice" [30].

Regret in the field of RL is usually defined as some form of the cumulative reward gained from the best policy minus the cumulative reward gained form the chosen policy [37], [38]. This is directly compatible with our proposed model, except that we are interested in the actions to which regret is attributed. Humans (and apparently also other apes, see above) feel regret *about* an action. To model the emotion of regret it is therefore important to investigate how correct attribution takes place as well.

III. TDRL model of Regret

In this article we focus on modelling regret using value-function based RL methods. Further, as we want to study minimal elicitation conditions for regret, we adopt a model-free learning paradigm for the action value updates (please note we do use a simulated model in one of the value update mechanisms, but only for sampling purposes, i.e., we do not learn and use environmental dynamics $T(s, a, s')$, with T referring to the transition function specifying how a state s' follows from an action a in a state s). In model-free RL we iteratively approximate the value-function through temporal difference (TD) reinforcement learning (TDRL), thereby avoiding having to learn the transition function. Well-known algorithms are Q-learning [39], SARSA [40] and TD(λ) [22]. TDRL approaches share the following: at each value update, the value is changed using the difference between the current estimate of the value and a new estimate of the value. This new estimate is calculated as the current reward and the return of the next state. This difference signal is called the temporal difference error. It reflects the amount of change needed to the current estimate of the value of the state the agent is in. The update equation for Q-learning is given by:

$$Q(s,a)_{new} \leftarrow Q(s,a)_{old} + \alpha\Big[TD\Big] \tag{1}$$

$$TD = r + \gamma \max_{a'} Q(s',a') - Q(s,a)_{old} \tag{2}$$

where α specifies a learning rate, γ the discount factor and r the reward received when executing action a, and $Q(s,a)$ the action value of action a in state s. Here the TD error is equal to the update taking the best action into account, while in the case of SARSA, the update is based on the actual action taken:

$$TD = r + \gamma Q(s', a') - Q(s,a)_{old} \tag{3}$$

Note that although model-based RL methods typically do not explicitly define the TD error, it still exist and can be calculated [19].

From the definition of regret [30], one can conclude two preconditions for regret. First, an agent must be able to evaluate its chosen actions at corresponding states in history, i.e., it should evaluate an action trace. Second, it must be able to realize there is an alternative action that would have resulted in a better outcome. In RL $Q(s,a)$ represents the value of an action a in state s. A better outcome at a certain situation s after evaluation means that there is a better action b with $Q_{new}(s,b) > Q_{new}(s,a)$.

Therefore, regret at state s associated to action a can be defined as:

$$Regret(a_t) = \max_b(Q_{new}(s_t,b)) - Q_{new}(s_t,a_t) \tag{4}$$

In the above equation, a is the action taken in state s. If after evaluation of the TD error, there is an action b with a higher Q value than the new $Q(s,a)$, then regret is the difference between these two. Otherwise, if the new $Q(s,a)$ is still the highest Q value for state s, then regret for this state is 0. Note that subscript t refers to the position in the state-action trace (the memory, or eligibility trace). So, in words, regret attributed to an action defined in TDRL terms equals the portion of the TD error that pushes the chosen action value below the best alternative action.

IV. Experiment

We tested the model described in the previous section in a maze task with 3 target reward changes and 3 value update methods (3x3 conditions). An agent acts in this maze (see Figure 1), available actions for the agent are $A = (up, down, left, right)$, collision with the wall will result in the agent staying in the same state with a reward of -0.5. This maze has two terminal states, a candy (top left corner) and a target (top right corner). Initially, the candy has a reward of +20 and the target has a reward of +30. To learn the task, we allow the agent to converge using standard Q-learning.

Then, to test the effect on regret intensity of different target reward changes, the target reward is set to 3 different new values after the agent converges to a policy that favors the target. The large change of reward is from +30 to -30, medium change is from +30 to +10 and small change is from +30 to +29.

Then, we allow the agent to follow its converged policy to walk to the target. Here, we keep track of the state-action trace in this one episode (start to target). This trace is not used in the standard Q-learning mentioned above, but we need it to evaluate regret for the actions chosen along the policy of the agent, after the reward change mentioned above.

After the reward change and after the agent arrived at the target state, the TD error is calculated and 3 different value update methods are tried (Sarsa-trace, Q-learning-trace and Monte Carlo (MC) reward averages). This is done to compare the effect of different value update methods on the elicitation of regret. This is important for studying regret attribution, as we will see later. The algorithms used are shown in Algorithm 2,3 and 4.

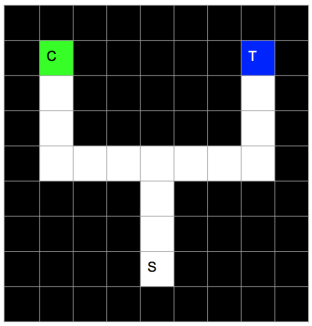

Fig. 1. Maze grid world. The agent stats at S and terminal states are candy(C) and target(T).

```
initiate Q, Episodes;
for e in range(Episodes-1) do
    Q(s,a) ← Q-Learning;
end
Target reward ← new reward;
Trace =[];
for e in range(1) do
    Q(s,a) ← Q-Learning;
    Trace.append(s,a);
end
trace back = Trace[::-1];
new Q(s,:) ← Evaluate s for s in trace back;
regret =[];
for (s,a) in trace back do
    regret[s] = max(new Q(s,:)) - new Q(s,a);
end
```

Algorithm 1: Regret elicitation and intensity calculation

```
for (s,a) in trace back do
    s', a', r ← according to (s,a);
    if s' is terminal state then
        δ ← r - Q(s,a);
    else
        δ ← r + γ * Q(s',a') - Q(s,a);
    end
    Q(s,a) ← Q(s,a) + α * δ
end
```

Algorithm 2: Sarsa-trace based Q value updates

```
for (s,a) in trace back do
    s', r ← according to (s,a);
    if s' is terminal state then
        δ ← r - Q(s,a);
    else
        δ ← r + γ * max(Q(s',:)) - Q(s,a);
    end
    Q(s,a) ← Q(s,a) + α * δ
end
```

Algorithm 3: Q-trace based Q value updates

```
for (s,a) in trace back do
    for action in A do
        for t in range(sampling times) do
            reward ← sampling a episode for (s,action);
            Total reward +=reward;
        end
        Q(s,action) ← (Total reward) / (sampling times);
    end
end
```

Algorithm 4: MC average based Q value updates

For each of the 9 experimental conditions (3 reward changes X 3 update mechanisms), the general steps are shown in Algorithm 1. In the beginning, the Q table is initialized with 0 values and the max number of episodes equals 40. An episode is defined as the agent moving from the start state to the candy or target state, or, a maximum number of steps. In all of the episodes except the last one, the agent learns the maze using standard Q-Learning with learning rate α=0.9, γ=0.9 and ϵ-greedy(ϵ=0.1). In these 39 episodes, the agent will finally converge to the target. Then the change in target reward is introduced, and the agent performs one more episode. During this last episode we store the actions it chose at corresponding states in $Trace$. For all states in this trace, we calculate the new Q values (using one of the three methods mentioned) after which we calculate regret attributed to the actions in the trace, using the definition in Section III. To examine regret, we plot regret intensities for the trace in these 9 different settings.

A. Results

We found that a small change in reward does not elicit any regret for all three value update methods (Figure 8). This is exactly as expected as we defined regret, in accordance with regret literature, to be a signal that occurs when a better alternative exists. With a small reward modification, there is no such alternative, and of course none of the update methods will result in new Q-values that change the policy.

We further found that when using the Q-trace to update Q-values, regret is attributed to the last action taken (Figure 3 and 6). This attribution is not dependent on the size of the reward change. This can easily be explained by the fact that a Q-trace does not really do much, if the alternative actions in states do not get updated as well. These alternative actions will function as new $\max_{a'} Q(s', a')$, so this will immediately cut off the propagation. The larger reward change did result, not surprisingly, in more regret about the last action taken.

We also found that in Sarsa-trace updates, regret is distributed over the complete action trace, with more regret attributed to actions close to the target (close to the reward change) (Figure 2 and 5). This follows immediately from the working of repeated sarsa updates executed in reverse order as well. TD errors will propagate back according to the chosen policy and will be scaled down according to the discount factor γ. The larger reward change resulted in more regret.

Finally, we observed that MC value updates results in regret attributed to the expected split point in the maze (Figure 4 and 7). This is because the MC roll-outs will sample new returns, and will, given sufficient roll-outs, relearn the value function for all alternative actions in the state-action trace. The medium reward change results in regret at the maze's split point about going right at that point. The large reward change results in regret attributed to the complete ending of that learning episode, with the most regret still being attributed to the split point. Some regret is attributed to the other actions in the right arm of the maze, due to the fact that also in those states the alternative action to go back to the left arm would have been better.

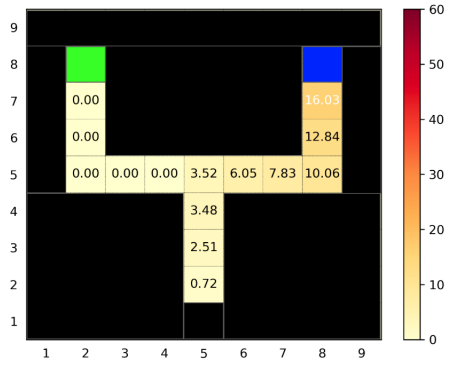

Fig. 2. Regret with Sarsa trace and medium reward change.

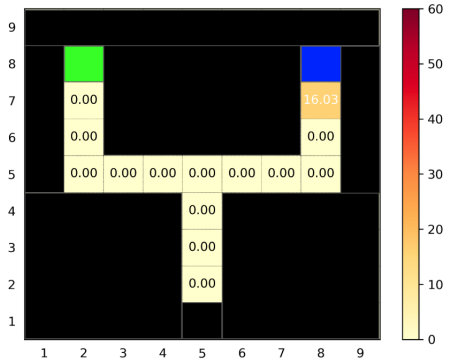

Fig. 3. Regret with Q trace and medium reward change.

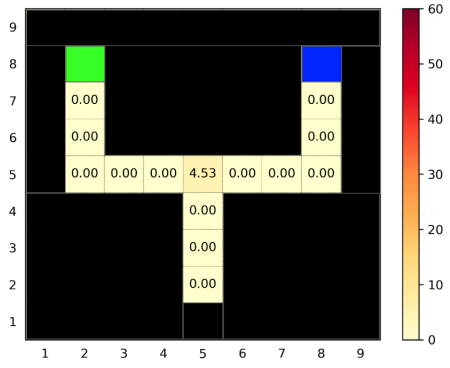

Fig. 4. Regret with MC average sampling and medium reward change.

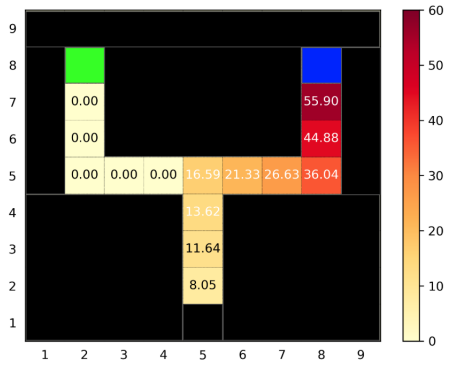

Fig. 5. Regret with Sarsa trace and big reward change.

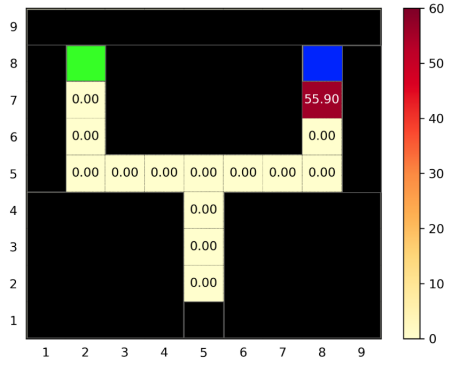

Fig. 6. Regret with Q trace and big reward change.

V. Discussion

Our results clearly show that a simple computational model of regret, coupled with MC value updates after a disappointing outcome result in regret attributed to the most responsible action, and for humans the most logical action. In this task this is the maze's split point. The other two versions of value updates using the same trace - Sarsa-trace and Q-trace updates - do result in regret intensity that is modified by the size of the reward change, but the action to which regret is attributed is not the plausible.

Regret based on Sarsa-trace updates results - for larger

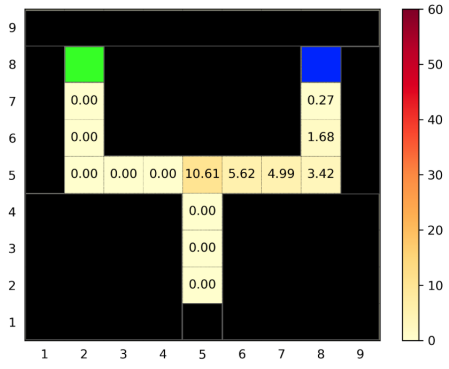

Fig. 7. Regret with MC average sampling and big reward change.

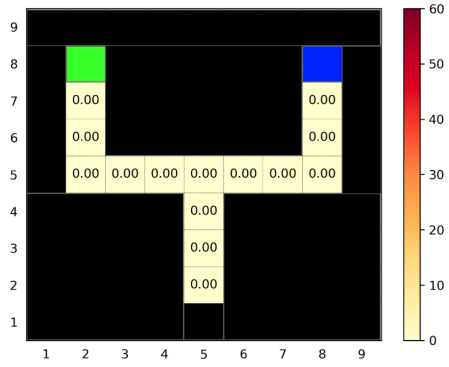

Fig. 8. Regret with small reward change for all 3 evaluation methods.

reward changes - in regret attributed to all actions in the trace, even to the beginning of the trace. This is not plausible, as there is no reason to regret an action in hindsight if you know you can still fully repair the situation later on. Regret should be attributed to key actions that could still make a change, as is the case with regret after Monte Carlo updates.

Regret based on Q-trace updates results in regret felt only about the last action taken. This is not plausible either as this is clearly not the action that bares the most responsibility in the trace, which is the choice to go right rather than left at the junction.

Our results align well with evidence that regret processing is in need of higher level cognitive processing and value processing of counterfactuals [30], [35], [41]. MC updates along the trace represent such counterfactural thinking, as they are random explorations of "what if" actions along the trace. This fits with behavioral data on regret showing that people who imagine a situation in which they felt regret also imagine changing actions to get to a better outcome [41]. This is in contrast to felt disappointment, which is associated with changes in the situation [41]. In TDRL terms this means that regret is a manifestation of the assessment of the degree to which a negative TD pushes the taken action's value below an imagined alternative action's value.

Our results show that this model for regret, formalized as the processing of a negative TD error resulting in alternative

actions becoming the better option, is a better model for regret than just the observed loss (TD error) (such as proposed in [36]). The latter would produce regret for chosen actions, even in the case where the reward was less than expected but the actions chosen were still the best policy. This is not plausible. One does not regret an action that in hindsight still is the best action. With our model this is indeed not the case.

Finally, our work shows that in addition to joy, distress, hope and fear, regret can be defined as a variation of TD error processing combined with simple counterfactual thinking, and, that regret can be simulated in computo using RL also for delayed action consequences. This, in our view, is additional evidence that the TDRL Theory of Emotion [18] is able to bridge the gap between cognitive views on emotion and emotion as emergent phenomenon from a learning and adapting agent.

VI. Conclusion

We proposed to model regret as a variation of TD error assessment. In particular, regret is attributed to those actions in an action trace for which - after new reward evidence - an alternative action becomes the best action in that state (the new argmax), after adjusting the action value of the chosen action in that state. Regret intensity is modeled as the difference between this new best action and the adjusted old best action, reflecting the additional amount of return that could have been received should that alternative have been chosen. We showed in simulation experiments how regret varies depending on the amount of adjustment as well as adjustment mechanism, i.e. Q-trace, Sarsa-trace, and Monte Carlo (MC) re-evaluation of action values. Our work shows plausible elicitation patterns of regret, when this model of regret is coupled with MC action value updates. This is important evidence that regret can be seen as a particular variation of TD error assessment coupled to counterfactural thinking and gives additional evidence for a TDRL Theory of Emotion that can help to better understand the relation between emotion, cognition and adaptive behavior.

References

[1] Nico H. Frijda, Antony S. R. Manstead, and Sacha Bem. *Emotions and Beliefs: How Feelings Influence Thoughts*. Cambridge University Press, 2000.

[2] Lisa Feldman Barrett, Batja Mesquita, Kevin N. Ochsner, and James J. Gross. The experience of emotion. *Annual Review of Psychology*, 58(1):373–403, 2007.

[3] A. H. Fischer and A.S.R. Manstead. *Social Functions of Emotion*, pages 456–468. Guilford Press, 2008.

[4] A. R. Damasio. *Descartes' Error: emotion reason and the human brain*. Putnam, New York, 1994.

[5] R. F. Baumeister, K. D. Vohs, and C. Nathan DeWall. How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review*, 11(2):167, 2007.

[6] Joost Broekens, Tibor Bosse, and Stacy C Marsella. Challenges in computational modeling of affective processes. *Affective Computing, IEEE Transactions on*, 4(3):242–245, 2013.

[7] Carien M Van Reekum and Klaus R Scherer. Levels of processing in emotion-antecedent appraisal. In *Advances in Psychology*, volume 124, pages 259–300. Elsevier, 1997.

[8] Celso M de Melo, Peter J Carnevale, Stephen J Read, and Jonathan Gratch. Reading peoples minds from emotion expressions in interdependent decision making. *Journal of Personality and Social Psychology*, 106(1):73, 2014.

[9] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

[10] Klaus R. Scherer. Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3459–3474, 2009.

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

[12] R. Reisenzein, E. Hudlicka, M. Dastani, J. Gratch, K. Hindriks, E. Lorini, and J. J. C. Meyer. Computational modeling of emotion: Toward improving the inter- and intradisciplinary exchange. *Affective Computing, IEEE Transactions on*, 4(3):246–266, 2013.

[13] Jonathan Gratch and Stacy Marsella. A domain-independent framework for modeling emotion. *Cognitive Systems Research*, 5(4):269–306, 2004.

[14] Robert T. Brown and Allan R. Wagner. Resistance to punishment and extinction following training with shock or nonreinforcement. *Journal of Experimental Psychology*, 68(5):503–507, 1964.

[15] A. David Redish, Steve Jensen, Adam Johnson, and Zeb Kurth-Nelson. Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. *Psychological Review*, 114(3):784–805, 2007.

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

[17] S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. *Autonomous Mental Development, IEEE Transactions on*, 2(2):70–82, 2010.

[18] Joost Broekens. A temporal difference reinforcement learning theory of emotion: unifying emotion, cognition and adaptive behavior. *arXiv preprint arXiv:1807.08941*, 2018.

[19] Joost Broekens and Mohamed Chetouani. Towards transparent robot learning through tdrl-based emotional expressions. *IEEE Transactions on Affective Computing*, in press, 2019.

[20] Joost Broekens, Elmer Jacobs, and Catholijn M. Jonker. A reinforcement learning model of joy, distress, hope and fear. *Connection Science*, pages 1–19, 2015.

[21] Thomas Moerland, Joost Broekens, and Catholijn M. Jonker. *Fear and Hope Emerge from Anticipation in Model-Based Reinforcement Learning*, pages 848–854. AAAI Press, 2016.

[22] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 1998.

[23] G. Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.

[24] Jacqueline Gottlieb, Pierre-Yves Oudeyer, Manuel Lopes, and Adrien Baranes. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in Cognitive Sciences*, 17(11):585–593, 2013.

[25] Jens Kober, J. Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[26] Daeyeol Lee, Hyojung Seo, and Min Whan Jung. Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience*, 35(1):287–308, 2012.

[27] Peter Dayan and Bernard W. Balleine. Reward, motivation, and reinforcement learning. *Neuron*, 36(2):285–298, 2002.

[28] Elmer Jacobs. *Representing human emotions using elements of Reinforcement Learning*. Master thesis, 2013.

[29] Agnes Moors, Yannick Boddez, and Jan De Houwer. The power of goal-directed processes in the causation of emotional and other actions. *Emotion Review*, 9(4):310–318, 2017.

[30] Giorgio Coricelli, Raymond J. Dolan, and Angela Sirigu. Brain, emotion and decision making: the paradigmatic example of regret. *Trends in Cognitive Sciences*, 11(6):258–265, 2007.

[31] Jonathan Gratch and Stacy Marsella. *Appraisal Models*, page 54. 2014.

[32] Thomas M. Moerland, Joost Broekens, and Catholijn M. Jonker. Emotion in reinforcement learning agents and robots: a survey. *Machine Learning*, 107(2):443–480, 2018.

[33] R. Guttentag and J. Ferrell. Reality compared with its alternatives: Age differences in judgments of regret and relief. *Developmental Psychology*, (5):764–775, 2004.

[34] Alexandra G Rosati and Brian Hare. Chimpanzees and bonobos exhibit emotional responses to decision outcomes. *PloS one*, 8(5):e63058, 2013.

[35] Giorgio Coricelli, Hugo D. Critchley, Mateus Joffily, John P. O'Doherty, Angela Sirigu, and Raymond J. Dolan. Regret and its avoidance: a neuroimaging study of choice behavior. *Nature Neuroscience*, 8(9):1255–1262, 2005.

[36] Stephen Marsh and Pamela Briggs. *Examining trust, forgiveness and regret as computational concepts*, pages 9–43. Springer, 2009.

[37] Sbastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[38] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of Articial Intelligence Research*, 4:237–285, 1996.

[39] Christopher John Cornish Hellaby Watkins. *Learning from delayed rewards*. Thesis, 1989.

[40] G. A. Rummery and M. Niranjan. On-line q-learning using connectionist systems. Report, Cambridge University Engineering Department., 1994.

[41] Marcel Zeelenberg, Wilco W. van Dijk, Joop van der Pligt, Antony S. R. Manstead, Pepijn van Empelen, and Dimitri Reinderman. Emotional reactions to the outcomes of decisions: The role of counterfactual thought in the experience of regret and disappointment. *Organizational Behavior and Human Decision Processes*, 75(2):117–141, 1998.