Emotion-Driven Reinforcement Learning

Robert P. Marinier III, John E. Laird ({rmarinie,laird}@umich.edu)
Electrical Engineering and Computer Science Department, 2260 Hayward
Ann Arbor, MI 48109 USA

Abstract

Existing computational models of emotion are primarily concerned with creating more realistic agents, with recent efforts looking into matching human data, including qualitative emotional responses and dynamics. In this paper, our work focuses on the functional benefits of emotion in a cognitive system where emotional feedback helps drive reinforcement learning. Our system is an integration of our emotion theory with Soar, an independently-motivated cognitive architecture.

Keywords: Emotion, reinforcement learning, intrinsic reward, cognitive architecture, appraisal theories.

Introduction

Folk psychology often casts emotions in a negative light. For example, in Star Trek, Vulcans are portrayed as superior to humans because they are not distracted by emotions, and thus can make purely logical decisions. As far back as Phineas Gage, however, it has been clear that emotions play a critical role in proper functioning in humans, and over the last several decades psychological research has explored how emotions influence behavior.

We are interested in exploring how some of the functional capabilities of emotion can be utilized in computational agents; that is, we want to bring the functionality of emotions to artificial intelligence. This is in contrast to most existing computational models of emotion, which focused primarily on creating believable agents (Gratch & Marsella, 2004; Hudlicka, 2004), modeling human data (Marsella & Gratch 2006; Gratch, Marsella, and Mao, 2006), or entertainment (Loyall, Neal Reilly, Bates, and Weyhrauch, 2004).

In this paper, we present work in which reinforcement learning is driven by emotion. Intuitively, feelings serve as a reward signal. The agent learns to behave in a way that makes it feel good while avoiding feeling bad. Coupled with a task that the agent wants to complete, the agent learns that completing the task makes it feel good. This work contributes not only to research on emotion in providing a functional computational grounding for feelings, but it also contributes to research in reinforcement learning by providing a detailed theory of the origin and basis of intrinsically-motivated reward.

Background

Our system in implemented in the Soar cognitive architecture (Newell, 1990). Soar is a complete agent framework composed of interacting, task-independent memory and processing modules that include short- and long-term memory, decision making, learning, and perception. We have augmented Soar with a new module, our emotion system, described below.

Integrating Appraisal Theories and Cognition

Our work is grounded in appraisal theories (Roseman & Smith, 2001, for an overview) and Newell's (1990) PEACTIDM (pronounced PEE-ACK-TEH-DIM). Appraisal theories hypothesize that an emotional reaction to a stimulus is the result of an evaluation of that stimulus along a number of dimensions, most of which relate it to current goals. The particular appraisals that our system uses is a subset of the appraisals described by Scherer (2001) (see Table 1). The subset our system uses can be split into two main groups: appraisals that help the agent decide which stimulus attend to (Suddenness, Unpredictability, Intrinsic Pleasantness, Relevance) and those appraisals that help the agent decide what do in response to an attended stimulus (causal agent and motive, outcome probability, discrepancy from expectation, conduciveness, control, power). Appraisal theories generally do not give much detail about how appraisals are generated or why. That is, the details of the process are left unspecified. Thus, computational models must fill in those details, but with little or no direction from appraisal theory, the details are often arbitrary. PEACTIDM, on the other hand, describes necessary and sufficient processes for immediate behavior (see

Table 2). The PEACTIDM hypothesis is that stimuli are Perceived and Encoded so cognition can work with them. Then Attend focuses cognition on one stimulus to process, which is then Comprehended. Tasking is managing tasks and goals (e.g., in response to a change in the situation), whereas Intend is determining what actions to take. Decode and Motor are translating cognitive choices into physical actions. PEACTIDM, however, does not describe the data upon which these processes operate. The basis of our theory is that appraisals are the information upon which the PEACTIDM theory operates (Marinier & Laird, 2006) (see

Table 3). For example, the attend process determines what to process next based on appraisal information generated by the perceive and encode processes (i.e., suddenness, unpredictability, intrinsic pleasantness, and relevance). Thus, appraisals not only determine the information that PEACTIDM processes, PEACTIDM also imposes dependencies between the appraisals (e.g., the appraisals for Comprehend cannot occur until after appraisals for Attend have been generated).

Table 1: Subset of Scherer's (2001) appraisals used by system

Suddenness	Extent to which stimulus is characterized by abrupt onset or high intensity
Unpredictability	Extent to which the stimulus could not have been predicted
Intrinsic pleasantness	Pleasantness of stimulus independent of goal
Relevance	Importance of stimulus with respect to goal
Causal agent	Who caused the stimulus
Causal motive	Motivation of causal agent
Outcome probability	Probability of stimulus occurring
Discrepancy from Expectation	Extent to which stimulus did not match prediction
Conduciveness	How good or bad the stimulus is for the goal
Control	Extent to which anyone can influence the stimulus
Power	Extend to which agent can influence the stimulus

Table 2: Newell's Abstract Functional Operations

Perceive	Obtain raw perception
Encode	Create domain-independent representation
Attend	Choose stimulus to process
Compre-hend	Generate structures that relate stimulus to tasks and can be used to inform behavior
Task	Perform task maintenance
Intend	Choose and action, create a prediction
Decode	Decompose action into motor commands
Motor	Execute motor commands

Table 3: Integration of PEACTIDM and Appraisal

Appraisals	Generated by	Required by
Suddenness	Perceive	Attend
Unpredictability	Encode	
Intrinsic pleasantness		
Relevance		
Causal agent	Comprehend	Comprehend, Task, Intend
Causal motive		
Outcome probability		
Discrepancy from Expectation		
Conduciveness		
Control		
Power		

Emotion, Mood and Feeling

Most existing models of emotion do not explicitly differentiate between emotion, mood and feeling. In our system, emotion is what the appraisals generate, and is thus a set of values – one value for each appraisal. Mood plays the role of a moving history of emotion. Its value is pulled towards the current emotion, and it decays toward a neutral value a little each cycle. Thus, it acts like an average over recent emotions. Mood provides some historical information, so that an agent's emotion, which might change wildly from one moment to the next, does not dominate the agent's interpretation of the situation. Feeling, then, is the combination of emotion and mood, and is what the agent actually perceives, and thus can respond to. Feeling and mood are represented as the same type of structure as emotion: as a set of appraisal values (Marinier & Laird, 2007).

Feeling intensity summarizes the feeling. The feeling is composed of many dimensions (one for each appraisal), whereas the intensity is a single number, valenced by the feeling's conduciveness (Marinier & Laird, 2007). It is the feeling intensity that will act as a reward signal for reinforcement learning.

Reinforcement Learning and Soar-RL

In reinforcement learning (Sutton & Barto, 1998) an agent receives reward as it executes actions in its environment. The agent attempts to maximize its future reward by maintaining a value function that encodes the agent's expected reward for each state-action pair. For a given state, the agent then selects the best action based on the stored values for the available actions. The value function is updated based on external rewards and the agent's own prediction of future reward.

In Soar-RL (Nason & Laird 2004), the value function is encoded as rules that associate expected rewards with state descriptions and operators. For a given state, all of the relevant rules fire, generating expected values, which are then combined for each operator to provide a single expected reward. An epsilon-greedy based decision procedure then selects the next operator. The expected rewards associated with rules are updated using the SARSA equation.

Intrinsically Motivated Reinforcement Learning

In traditional reinforcement learning, an agent perceives states in an environment and takes actions. A critic, located in the environment, provides a rewards and punishments in response to the choices being made. The agent learns to maximize the reward signal (Sutton & Barto, 1998). This model is highly abstract and assumes a source of reward that is specific to every task.

In intrinsically motivated reinforcement learning, the environment is split into internal and external parts. The organism is composed of the internal environment together with the agent (Singh, Barto, and Chentanez, 2004). The critic resides in the internal environment, and thus the organism generates its own rewards.

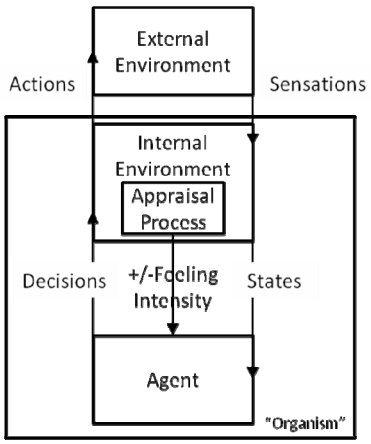

Figure 1: Our system viewed as an intrinsically motivated reinforcement learner. (Adapted from Singh et al., 2004.)

In our system, the appraisal process acts as the critic, and the resulting valenced feeling intensity provides the reward signal over which the agent learns (Figure 1). Appraisal values are generated by rules, which match patterns in perception and internal state. As the situation changes, the appraisal values change with it. The values are then detected by a module that updates the current emotion, mood and feeling states of the agent.

Previous Work

The idea that emotion influences reinforcement learning is not new. Grossberg (1982) describes a connectionist system theory in which drives influence learning. Damasio (1994) describes experiments in which subjects with emotional impairments have difficulty learning.

More recently, there have been other attempts to integrate emotion-like processes with reinforcement learning. Hogewoning, Broekens, Eggermont, and Bovenkamp (2007) describe a system developed in Soar that adjusts its exploration rate based on short- and long-term reward trajectories. They consider the reward histories to be a kind of affect representation. This work is differs from our own in that it is not based on appraisal theories and rewards are not intrinsically generated.

Salichs & Malfaz (2006) describe a system that is capable of happiness, sadness and fear. Happiness and sadness serve as positive and negative rewards, while fear affects the selection of "dangerous" actions. Happiness is generated when an external stimulus is present that is related to current internal drives (e.g., if hungry and food is present, the agent will be happy). Sadness is when the desired external stimulus is not present. Fear is when the state values have a large variance (even if positive overall). This work connects physiology to goals and thus emotion. However, it commits to a categorical view of emotions and thus requires a separate equation for each emotion. In our approach, emotions emerge from the interaction of the appraisals, providing a continuum of emotions (on top of which a categorical view could be imposed if desired).

Experimental Task

To test the system, we created a maze task for the agent (Figure 2). The agent had to learn to navigate the maze, starting in the upper left and going to the far right. While the maze may look simple, the task is actually very difficult because the agent has essentially no knowledge about the environment.

In this environment, the agent's sensing is limited: it can only see the cells immediately adjacent to it in the four cardinal directions. The agent has a sensor that tells it its Manhattan distance to the goal. However, the agent has no knowledge as to the effects of its actions, and thus cannot evaluate possible actions relative to the goal until it has actually performed them. Even then, it cannot always blindly move closer to the goal because given the shape of the maze, it must sometimes increase its Manhattan distance to the goal in order to make progress in the maze.

Thus, the agent must learn such simple things as sometimes moving in directions that reduce the distance to the goal, not walking into walls, and avoiding backtracking. At the lower level, it is actually learning about which PEACTIDM steps to take and when. For example, it learns which direction to attend to, so it can take actions that bring it closer to the goal. When the agent cannot take any actions that bring it closer to the goal, it must learn to do internal actions; specifically, to create subtasks to make progress in the maze while moving away from the goal.

The agent appraises situations in the following manner. The place an agent has just been has low suddenness, while other places have higher suddenness. Directions leading to walls have low intrinsic pleasantness. Directions leading away from the goal have low relevance. Getting closer to the goal has positive conduciveness, while getting further away has negative conduciveness. The agent makes simple predictions that it will make progress with medium probability; if it does not, discrepancy from expectation is high. When the agent is about to accomplish a task or a subtask, discrepancy from expectation is also high (it is pleasantly surprised). If the agent is attending to a wall, control and power are low (since the agent cannot walk through walls); otherwise, they are high. Causal agent and motive do not play much of a role in this domain, since there is only one agent. Unpredictability also does not play much of a role.

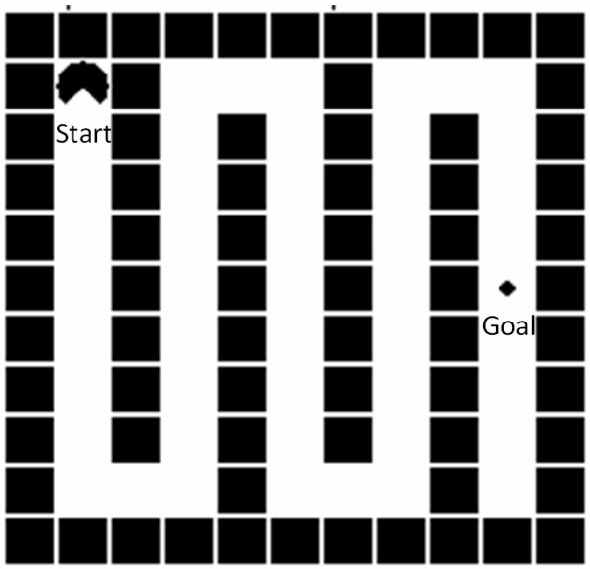

Figure 2: The experimental task. The agent had to learn to navigate this maze.

Methodology

Three agent types were tested: a standard reinforcement learning agent, which only received reward at the end when it accomplished the goal, an agent that had no mood (so its feelings were its emotions) and a full agent that included mood.

We expect the standard reinforcement learning agent to have difficulty since it does not have access to the sensor that tells it how far it is from the goal. This may seem like an unfair comparison; however, creating a standard reinforcement learning agent with this capability but without the other appraisal information is difficult, since the appraisal representations comprise part of the state representation. If we were to remove the appraisal information, then the standard reinforcement learning agent would really be solving a different problem. If we leave the appraisal information, the agent is not really standard. However, the agent without mood can be viewed as a very rough approximation of an agent that would take advantage of this information to generate more frequent rewards. This approximation includes appraisal information, but without mood it is not the complete emotion system. Thus, we have two extremes and an intermediate agent: an agent with no emotion information at all, an agent with emotion but no mood, and an agent with both emotion and mood.

The agents learned across 15 episodes. This was repeated in 50 trials. Each episode and trial took place in the same maze (shown in Figure 2). We recorded the amount of time it took each agent to complete the task (measured in Soar decision cycles). Because of the task difficulty, the agents would sometimes get hopelessly lost; thus, to limit processing time, episodes were cut off at 10000 decision cycles. We report the median to avoid skewing the data.

Results

The results are shown in Figure 3 and Figure 4. The horizontal axis is the episode number, while the vertical axis is the median amount of time it took the agent to complete the task.

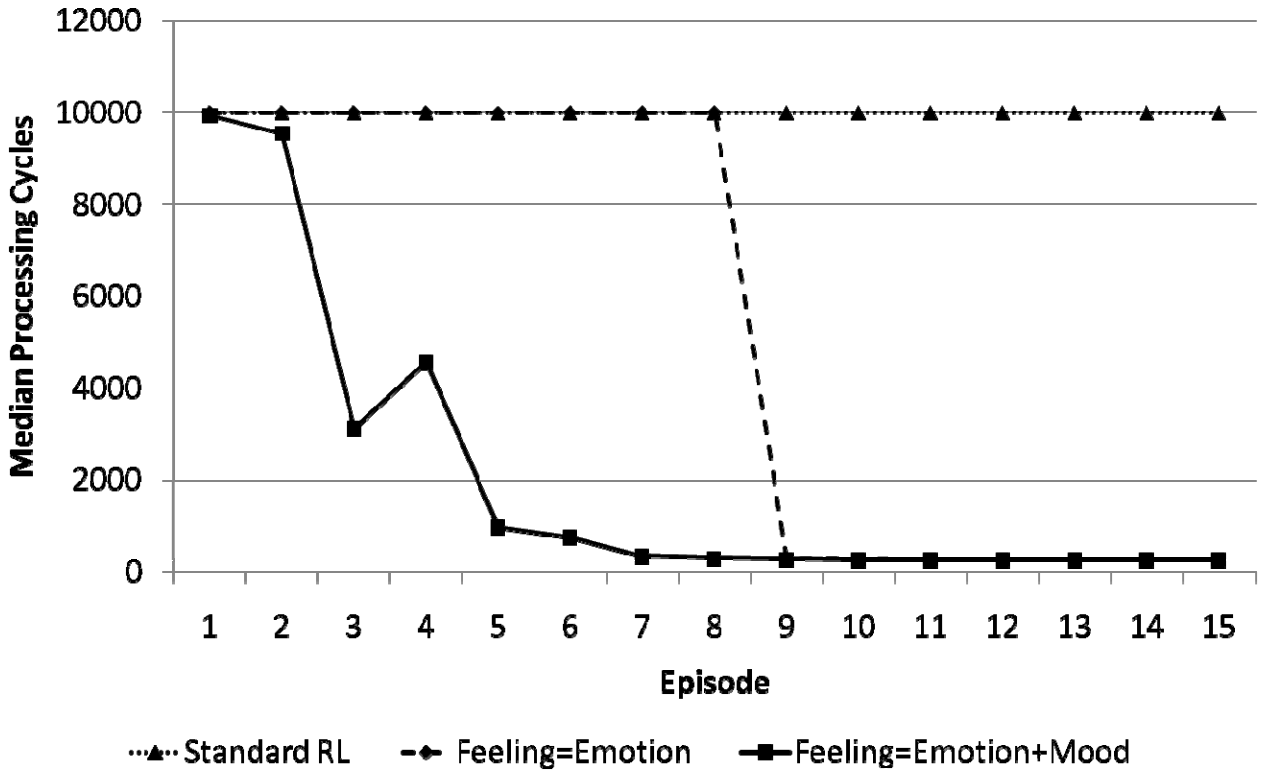

Figure 3: Learning results for three different agents.

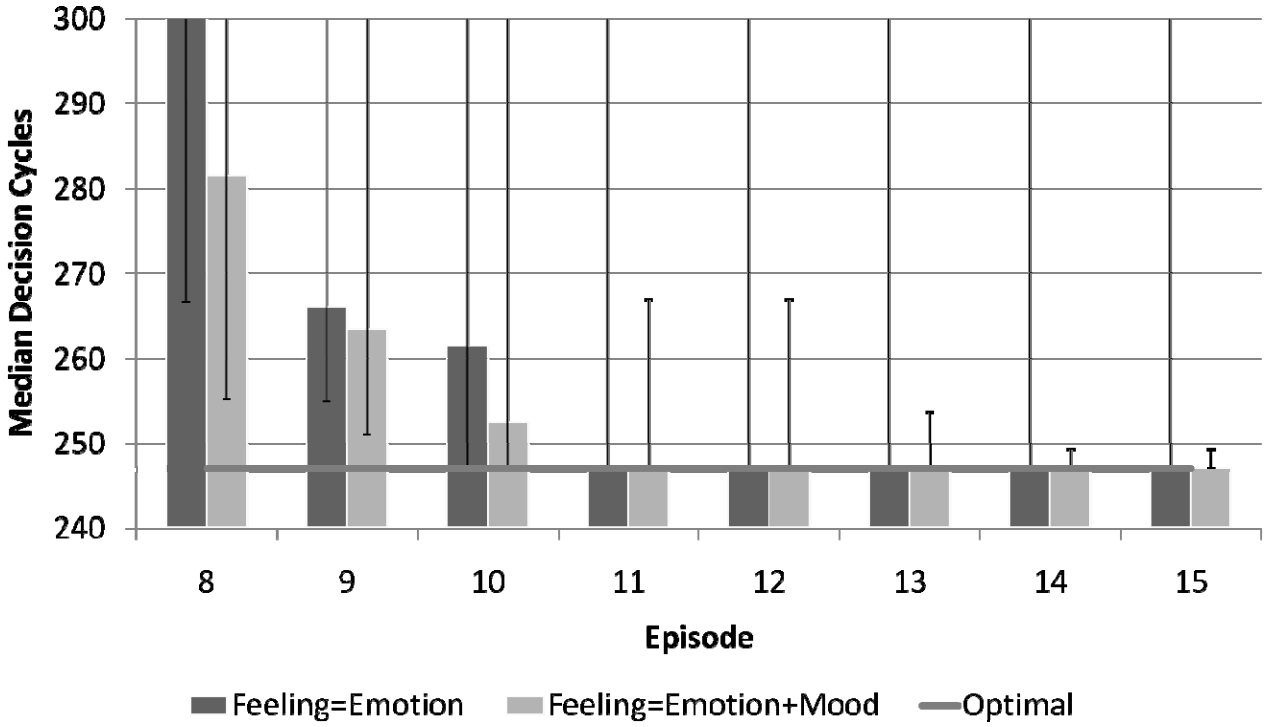

Figure 4: Close-up of last several episodes for agent with just emotion and agent with emotion and mood. "Error" bars show first and third quartiles.

First, consider Figure 3. The standard reinforcement learning agent never made any significant progress. This is expected because 15 training episodes do not provide the agent with enough experience when it only gets a single reward for each. In reinforcement learning, the reward "backs up" only one state per episode, and there are many more than 15 states in this domain.

The agent whose feeling is just its emotion (without mood) does not appear to be learning at first, but the values eventually converge. The agent whose feelings are composed of both emotion and mood does much better earlier on, learning much faster.

Figure 4 shows a close-up of the last several episodes for the two agents with emotions. The "error" bars show the first and third quartiles, which gives an indication of the amount of variability in the agents' behavior at that point. As we can see, both agents reach optimality at the same time, but the variability of the agent with mood is much lower. In fact, the variability of the moodless agent reaches all the way to 10000 even in the final episode, implying that fewer agents did well on the task. In contrast, by the final episode, the agent with mood has minimal variance.

Discussion

The first thing to note is that the agents with emotion learn very fast relative to the standard reinforcement learning agent. This is because they get frequent reward signals (on every decision cycle) and thus get intermediate feedback on how they are doing. The standard reinforcement learning agent only gets reward feedback at the end, and thus it takes a long time for that information to propagate back to earlier actions. This result is not unexpected since part of the agent's feeling is related to whether it is getting closer to the goal or not (albeit with the caveats mentioned earlier, it was not clear that the agent would be able to learn at all).

Next, the agent with mood learns faster. The reason is because, sometimes when the agent is doing some internal bookkeeping kinds of processing, it is not experiencing an emotion. Thus, the agent without mood will get zero reward for those states, and later reward has to propagate back through those states. Propagation takes time (this is why the standard reinforcement learning agent takes so long to learn).

The agent with mood, however, carries a summary of its recent emotions forward into those states (with some decay). Thus, these states get reasonable value estimates, which speeds up the propagation immensely.

In this experiment, a key factor in the success of the agent's using emotion is the availability of the knowledge about the Manhattan distance to the goal, which acts as an intermediate reward. What we have presented is a theory about the origin of those rewards and how they are applied to an integrated system centered on abstract functional operations, and a

demonstration that the rewards generated by that theory do, in fact, speed learning.

Future Work and Conclusion

Much work remains to be done. We are currently working on getting an agent learning in a more complex domain. We also plan to explore which subset of appraisals actually influences the learning. For example, we don't expect that the causal agent has much influence, because there is only one agent in the domain presented. However, we should be able to construct domains in which the causal agent plays a critical role.

In conclusion, we have developed a computational model of emotion that integrates with a theory of cognition (PEACTIDM) and a complete agent framework (Soar). We have also confirmed a functional advantage of that integration that had been proposed by other models; namely, that feelings can drive reinforcement learning. Finally, the system is learning how and when to execute various steps in the PEACTIDM process which, unlike typical reinforcement learning systems, includes learning both external actions and internal actions

Acknowledgments

The authors acknowledge the funding support of the DARPA "Biologically Inspired Cognitive Architecture" program under the Air Force Research Laboratory "Extending the Soar Cognitive Architecture" project award number FA8650-05-C-7253.

References

Damasio, A. (1994). Descartes' Error: Emotion, Reason, and the Human Brain. New York: Avon Books.

Gratch, J. & Marsella, M. (2004). A Domain-independent Framework for Modeling Emotion. *Cognitive Systems Research*, 5:269-306.

Gratch, J., Marsella, S., & Mao, W. (2006). Towards a Validated Model of "Emotional Intelligence". (pp. 1613-1616). In 21st National Conference on Artificial Intelligence (AAAI06), Boston, Massachusetts.

Grossberg, S. (1982). A Psychological Theory of Reinforcement, Drive, Motivation and Attention. Journal of Theoretical Neurobiology, 1:286-369.

Marsella, S. & Gratch, J. (2006). EMA: A computational model of appraisal dynamics. In Robert Trappl, editor, *Cybernetics and Systems 2006* (Volume 2, pp. 601-606). Vienna: Austrian Society for Cybernetic Studies.

Nason, S. & Laird, J. (2005). Soar-RL, Integration Reinforcement Learning with Soar. *Cognitive Systems Research*, 6:51-59.

Hogewoning, E., Broekens, J., Eggermont, J., & Bovenkamp, E. (2007). Strategies for Affect-Controlled Action-Selection in Soar RL. (pp. 501-510). In J. Mira and J.R. Alvarez, editors, *IWINAC 2007 Part II, LNCS 4528*, Berlin Heidelberg: Spinger-Verlag.

Hudlicka, E. (2004). Beyond Cognition: Modeling Emotion in Cognitive Architectures. In *Proc. of the Sixth International Conference on Cognitive Modeling* (pp. 118-123). Mahwah, NJ: Lawrence Earlbaum.

Loyall, A. B., Neal Reilly, W. S., Bates, J. & Weyhrauch, P. (2004). System for Authoring Highly Interactive, Personality-Rich Interactive Characters. (pp. 59-68). In R. Boulic and D. K. Pai, editors, *Eurographics/ACM SIGGRAPH Symposium on Computer Animation*.

Marinier, R. & Laird, J. (2006). A Cognitive Architecture Theory of Comprehension and Appraisal. In Robert Trappl, editor, *Cybernetics and Systems 2006* (Volume 2, pp. 589-594). Vienna: Austrian Society for Cybernetic Studies.

Marinier, R. & Laird, J. (2007). Computational Modeling of Mood and Feeling from Emotion. In Danielle S. McNamara and J. Gregory Trafton, editors, *Proc. of the 29th Meeting of the Cognitive Science Society (CogSci 2007)* (pp. 461-466). Nashville, Tennessee.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Roseman, I. & Smith, C. (2001). Appraisal Theory: Overview, Assumptions, Varieties, Controversies. In Klaus Scherer, Angela Schorr, and Tom Johnstone, editors, *Appraisal Processes in Emotion: Theory, Methods, Research*. New York and Oxford: Oxford University Press, pp. 3-19.

Salichs, M. & Malfaz, M. (2006). Using Emotions on Autonomous Agents. The Role of Happiness, Sadness, and Fear. (pp. 157-164). In *Adaptation in Artificial and Biological Systems (AISB'06)*. Bristol. England.

Scherer, K. (2001). Appraisal considered as a process of multi-level sequential checking. In Klaus Scherer, Angela Schorr, and Tom Johnstone, editors, *Appraisal Processes in Emotion: Theory, Methods, Research*. New York and Oxford: Oxford University Press, pp. 92-120.

Singh, S., Barto, A, & Chentanez, N. (2004). Intrinsically Motivated Reinforcement Learning. In *Proc. of Advances in Neural Information Processing Systems 17 (NIPS)*.

Sutton, R. & Barto, A. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.