A Computational Framework for Emotion-Based Control

Juan D. Velásquez

MIT Artificial Intelligence Laboratory
545 Technology Square, NE43-935
Cambridge, Massachusetts 02139
jvelas@ai.mit.edu

Abstract

This paper describes a computational framework for Emotion-Based Control. Inspired by work in Psychology, Neuroscience, and Ethology, this framework captures important aspects of emotional systems and integrates these with other models of perception, motivation, behavior, and motor control. Its modular nature allows for an incremental approach in which new features are added as more is known about emotional processing in the brain. We have followed this approach to develop and study Emotion-Based Control systems for several autonomous agents in an effort to obtain new insights into the workings of emotional systems.

Introduction

Computational modeling has grown into a very flexible and powerful approach to the study of cognitive, and biological processes in general. With respect to the study of emotions, however, the case is rather different. While some computational models have been developed, most of these are either oriented toward using emotions for entertainment purposes [Bates 1994; Blumberg 1994; Elliot 1992; Maes 1995; Reilly 1996] or are focused on specific aspects, such as modeling very limited psychological problems [Abelson 1963; Colby 1975], recognizing emotions [Picard 1997] or modeling influences in goals and learning [Blumberg, Todd, and Maes 1996; Frijda 1986; Kitano 1995; Pfeifer 1988]. In addition, most of these models have been based on symbolic representations of post-hoc qualitative observations of emotional experiences (i.e., models of cognitive appraisal), and have very little or no relation to the possible underlying brain mechanisms. For a review of some of these models see [Pfeifer 1988].

Recent ideas and discoveries about some of the neural bases of emotional processing have set up a foundation for the development of computational models that are concerned with the neural mechanisms underlying such processes, and which can be used to test and even extend current theories of emotion [Damasio 1994; LeDoux 1993].

This paper proposes an approach to the study of emotion that relies on the use of computational frameworks for what we call *Emotion-Based Control*, control of autonomous agents that relies on, and arises from, emotional processing.

The rest of the paper is divided in three parts: Part I provides some motivation and main ideas related to Emotion-Based Control. Part II presents a framework for the design of Emotion-Based Control Systems. Finally, part III provides a brief overview of an instantiation of the proposed framework in the control of several different agents.

Emotion-Based Control

Like many others, we view emotions as biological functions of the nervous system that have been conserved throughout evolution, and which are necessary for satisfying basic conditions that are crucial to survival (e.g., escape from danger, obtain food and shelter, and mate).

With this in mind, we propose the use of computational frameworks for Emotion-Based Control as an alternative approach to help us understand how emotional systems might work and thus address questions and issues about the nature of emotions as adaptive systems. In particular, we argue that:

- We should develop computational models that are more concerned with the neural mechanisms underlying emotional processing, and less concerned with the cognitive evaluation of emotional experiences, as it is the case with models of emotion appraisal on which most computational models proposed to date are based.
- While developing these models, several issues should be taken into consideration, such as differentiating emotions from other affective phenomena (e.g., moods and temperament), modeling the dynamic nature of emotional systems, considering both innate and learned elicitors of emotion, and considering the influences of emotional systems in other processes and mechanisms (e.g., biasing of perception and behavior).
- In contrast to many computational models proposed to date, we should build complete systems that integrate not only models of emotions, but also drives, perception, behavior, and motor control.
- If we build these systems in a modular fashion, we can incrementally implement new features as more is learned about the different works of emotional systems and their function.
- At every step during this incremental process, we should build complete agents that operate in the real world with real sensing and real action.

We have been following this approach to develop a computational framework that allow us to create Emotion-Based Control systems. These systems are both simple and sufficiently constrained so that we can study them with the hope of providing insights into the inner workings of emotions as they may have evolved in the brain.

The Framework

The architecture, an extension of the Cathexis model described in previous work [Velásquez 1996; Velásquez 1997] and which represents the essential components of *Emotion-Based Control* described earlier, is depicted in Figure 1. Relevant systems are represented as modules. The Sensory Systems obtain information from the world and provide the Emotion Generation and Behavior Systems with stimuli features and objects. These systems also receive error signals from Drive Systems. The Emotional Systems assess the emotional relevance of stimuli and bias behavioral responses and future perception accordingly. Relevant Behavior Systems generate and execute appropriate Motor Actions.

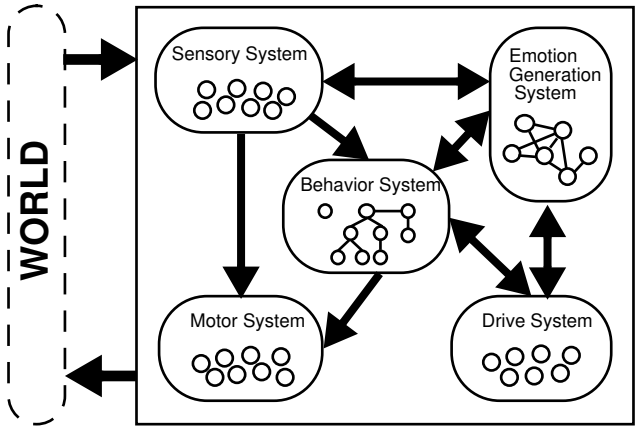

Figure 1 Schematic of the proposed framework for Emotion-Based Control.

The modular representations adopted in the figure exist solely for illustration purposes. In practice, none of these systems exist as single modules that are connected with each other through one single interface, but rather they consist of distributed networks of independent processing units which connect directly to other relevant units.

Computational Units

The basic elements of the framework correspond to nonlinear computational units that are used to represent all relevant systems (i.e., sensory systems, emotional systems, behaviors, drive systems, and motor systems).

Each of these systems can be thought of as consisting of a set of inputs, an appraisal mechanism and a set of outputs (see Figure 2).

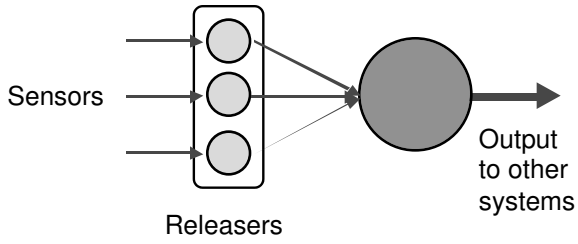

Figure 2 The Basic Computational System

An important component of the appraisal mechanism consists of what we have called *Releasers*, which correspond to computational units that filter sensory data and identify special conditions which will provide excitatory (positive) or inhibitory (negative) input to the system they are associated with.

Releasers can be innate and hard-wired (*Natural Releasers),* or they can be learned (*Learned Releasers*). The latter kind represent stimuli that tend to be associated with and predictive of the occurrence of *Natural Releasers*.

All releasers posses short-term memory and thus can habituate to stimuli using a model of rate-sensitive habituation similar to that proposed by [Staddon 1993].

The general form of the response or activation of each basic system is a nonlinear function of its input. Which in turn is proportional to the output of its *Releasers* and the strengths of their connections, or weights, as described in Equation (1):

$$A_i = f\left(\sum_k (R_{ki} \cdot W_{ki})\right) \quad (1)$$

Where A_i is the activation of System i; R_{ki} is the value of Releaser k, and W_{ki} is its associated weight, where k ranges over the set of releasers for System i, and f is a limiting function such as the standard ramp and logistic functions.

Each specific system (i.e., Sensory System, Emotional System, Behavior, Motor System, or Drive System), however, may compute its activation value in a slightly different manner as it will be described below.

Drive Systems

Drives are motivational systems representing urges that impel the agent into action. For instance, a *Hunger* drive will aid in controlling behaviors that directly affect the level of food intake by the agent. It should be noted, however, that Drive Systems are differentiated from Emotional Systems. In fact, within the proposed framework, Emotional Systems constitute the main motivations for the agent and even Drive Systems exploit their power to bias specific behaviors. In other words, it is both the error signal produced by a *Hunger* drive, and the *Distress* caused by it, that motivates the agent to obtain food.

Drive releasers represent control systems that maintain a

controlled variable within a certain range (see Figure 3). This variable is measured through some of the agent's sensors and compared to a desired value or *set point*. If its value does not match the set point, an error signal is produced. This error signal is fed to the appropriate drive, in which it can be combined with error signals from other relevant control systems. For instance, a *TemperatureRegulation* drive would combine the signals from two different control systems: one controlling peripheral temperature and one controlling brain temperature.

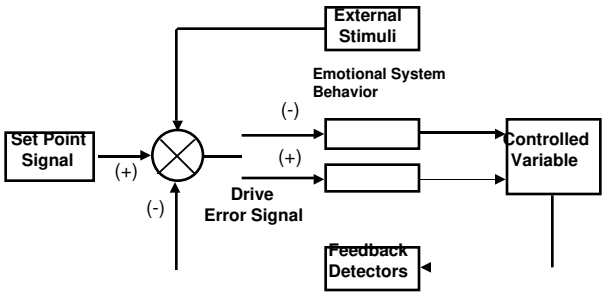

Figure 3 Drive Control Systems

Emotional Systems

Following a neuropsychological perspective, Emotional Systems bear resemblance to some of the aspects in which the interactions between neural systems involving the amygdala, the hippocampus, and the prefrontal cortices have been considered to mediate emotions, such as assigning an emotional valence to different stimuli, activation of emotional behaviors, and emotional learning [Damasio 1994; LeDoux 1993; Panksepp 1995].

Emotional systems represent different families of related affective responses, such as Fright, Fear, Terror, and Panic. Each member of an emotion family shares certain mechanisms and characteristics, including similarities in antecedent events, expression, likely behavioral response, and physiological patterns. These characteristics differ between emotion families, distinguishing one from another.

Drawing on ideas from different theorists [Ekman 1992; Izard 1991; Johnson-Laird and Oatley 1992], we have identified and created explicit models for six different emotion families: *Anger*, *Fear*, *Distress/Sadness*, *Enjoyment/Happiness*, *Disgust*, and *Surprise*. The selection of this core set of emotion types is not arbitrary, but rather it is based on evidence suggesting their universality, including distinctive universal facial expressions, as well as eight other properties [Ekman 1992].

In contrast to other models proposed to date that emphasize on cognitively generated emotions, and influenced by [Izard 1993], we consider both cognitive and noncognitive releasers of emotion. We divide all releasers of emotion into four different groups:

- *Neural*: Includes the effects of neurotransmitters, brain temperature, and other neuroactive agents that can lead to emotion and which can be mediated by hormones, sleep, diet, and environmental conditions.
- *Sensorimotor*: Includes sensorimotor processes, such as facial expressions, body posture, and muscle action potentials, that not only regulate ongoing emotion experiences but can also elicit emotion.
- *Motivational*: Includes all motivations that lead to emotion. In this model, motivations include drives (e.g. *Thirst* and *Hunger*), emotions (e.g. *Anger*, and *Happiness*), and pain regulation. Some examples of elicitors in this system include the innate response to foul odors or tastes producing disgust, as measured in neuropsychological studies by [Fox and Davidson 1986], pain or aversive stimulation causing anger, and emotions like sadness eliciting others such as anger.
- *Cognitive*: Includes all type of cognitions that activate emotion, such as appraisal of events, comparisons, attributions, beliefs and desires, and memories.

Currently, all releasers from the first three categories are *Natural Releasers* and all cognitive ones are *Learned Releasers*. In previous work, these cognitive releasers were based on a cognitive appraisal theory (See [Velásquez 1997] for details). In the proposed framework, however, Emotional Systems do not include any pre-wired cognitive releasers, but rather these are learned throughout the agent's interactions with the world as it is described below.

The main reason behind this change is that although suitable to test theories of cognitive evaluation of emotional experiences, computational models of emotion appraisal are very limited for the understanding on how the brain might process emotions. This is primarily due to the fact that these models are based on post-hoc cognitive recollections of emotional experiences and they are not concerned with the neural mechanisms behind such processes.

The activation of Emotional Systems differs slightly from that of the general form described above. Its main differences include considering the excitatory (positive) and inhibitory (negative) input from other Emotional Systems, as well as considering its temporal decay. This is summarized in Equation (2):

$$A_i(t) = f\left(\psi(A_i(t-1)) + \sum_k R_{ki} \cdot W_{ki} + \sum_l \mu_{li} \cdot A_l(t)\right) \quad (2)$$

Where $A_i(t)$ is the activation of Emotional System i at time t; A_{it-1} is its activation at the previous time step; $\psi()$ is the function that controls the temporal decay of the activation of Emotional System i; R_{ki} is the value of Releaser k, and W_{ki} is its associated weight, where k ranges over the set of releasers for Emotional System i; μ_{li} is the strength of the excitatory (positive) or inhibitory (negative) input from Emotional System l, where $A_l(t)$ is its activation value at time t; and f is a limiting function such as the standard ramp and logistic (sigmoid) functions.

Fast Primary Emotions. This model of Emotional Systems allows for the distinction between different affective phenomena. For instance, primary emotions are modeled as

the activation, via *Natural Releasers*, of one particular Emotional System such as *Disgust* or *Fear.*

These primary emotions play an essential role in the preparation of appropriate emotional responses that are adaptive for the agent. Such is the case, of the Fear Emotional System, which, through these fast primary emotions detects a dangerous situation and generates defensive responses that maximize the probability of surviving it.

Emotional Memories and Secondary Emotions. Following the ideas behind both Damasio's somatic-marker hypothesis [Damasio 1994], and LeDoux's work on fast (low-road) and slow (high-road) pathways for emotion activation [LeDoux 1993], we have extended Emotional Systems with the capacity of acquiring *Learned Releasers,* which, as we previously mentioned, correspond to stimuli that tend to be associated with and predictive of *Natural Releasers*. The activation of Emotional Systems via these *Learned Releasers* correspond to what are referred to by many as secondary emotions. These secondary emotions occur only after we begin experiencing feelings and start making orderly associations between objects and situations, and primary emotions. They require more complex processing, including in most cases, the retrieval of emotional memories of similar previous experiences.

Secondary emotions also play an adaptive role in dealing with situations that have occurred over and over throughout evolution, such as escaping danger, finding food, and mating. For instance, while primary emotions might include pre-wired mechanisms for the detection of a potential mate and the ensuing mating responses, secondary emotions provide the means to associate learned stimuli (e.g., the place where a potential mate was last seen) with these same mechanisms.

These emotional memories have been modeled with an associative network comparable to Minsky's K-lines [Minsky 1986], in which primary emotions are connected to the specific stimuli (e.g., executed behaviors, objects, or agents) that have released them during the agent's interaction with the world.

During emotional learning, connections within this network are changed according to a modified Hebbian rule that prevents saturation of the connection weight between the newly created Cognitive Releaser and the active Emotional System. The sum of the weights of all incoming learned connections to an Emotional System are kept constant, through multiplicative normalization based on the existing excitatory and inhibitory connections between Emotional Systems.

Moods. Following a psychobiological perspective [Panksepp 1995], moods are differentiated from emotions in terms of levels of arousal. While emotions consist of high arousal of specific Emotional Systems, moods may be explained as low tonic levels of arousal within the same systems. This representation is consistent with the enormous subtleties of human moods and feelings, as well as with the common observation that moods seem to lower the threshold for arousing certain emotions. This occurs because Emotional Systems that are aroused, as it happens in the representation of moods, are already providing some potential for the activation of an emotion. Finally, it is consistent with the observation that the duration of moods appears to be longer than that of the emotions, since at low levels of arousal, the intensity of the Emotional Systems will decay more slowly.

Temperament. Finally, temperaments are modeled through the different values that parameters (e.g., thresholds, gains, and decay rates) within each Emotional System can have. Thus, for instance, if we want to model a depressed agent, we might lower the activation threshold and decay rate for the *Sadness* emotion as well as lowering the inhibitory gain between *Happiness* and *Sadness*.

The result is a flexible, distributed framework that can synthesize a variety of affective phenomena simultaneously.

Behavior Systems

Following Damasio's view, reasoning and decision-making define a domain of cognition in which an agent must choose how to respond to a situation. This choice is responsibility of Behavior Systems. The action-selection process is mediated by the reasoning engaged.

Behavior Systems represent interconnected, self-interested behaviors, such as *"approach human"*, "*play*", "*request attention*", and "*avoid obstacle*".

Like Drive Systems and Emotional Systems, Behavior Systems also have Releasers that obtain and filter sensory data in order to identify special conditions which will either increase or decrease their activity. Releasers might represent physical objects and specific conditions, such as "battery recharger is present", as well as motivational states such as "battery level is low" and "distress is high", which would most likely increase the activity of a "*recharge batteries*" behavior.

Behavior Systems may mutually inhibit or excite each other. For instance, "*wag the tail*" might inhibit "*running*" and vice-versa. Whereas behaviors such as "*play with human*" might excite lower-level ones like "*find human*".

In earlier work, the Behavior System followed a winner-take-all strategy in which only one behavior could be active at a time. This made it impossible for non-conflicting Behaviors, such as "*walk*" and "*cry*" to execute at the same time. Given the parallelism of the model, we revised the Behavior System so that active, non-conflicting Behaviors can issue motor actions simultaneously. The value for each Behavior is computed as described in Equation (3):

$$B_j(t) = \sum_n (R_{nj} \cdot W_{nj}) + \sum_l (\mu_{lj} \cdot B_l(t)) \qquad (3)$$

Where $B_j(t)$ is the value of Behavior j at time t; R_{nj} is the value of releaser n and W_{nj} is the weight for releaser n, where n ranges over the releasers for Behavior j; μ_{lj} is the strength of the excitatory (positive) or inhibitory (negative) input from Behavior System l, where $B_l(t)$ is its activation value at time t.

Implementation and Results

The framework for Emotion-Based Control described in the previous sections has been used to develop and control various synthetic agents, including *Simón the Toddler,* a synthetic character, and *Virtual Yuppy*, a simulated emotional pet robot. For descriptions of these systems the interested reader is referred to [Velásquez 1997; Velásquez 1998].

As part of our ongoing research, the same framework is currently being used to control *Yuppy*, an emotional pet robot shown in Figure 4.

The implementation details and results described in this section correspond to observations on both the simulated and physical robots.

The Sensory System

Yuppy has different sensors, including two color CCD cameras (currently using only one as part of its vision system), an active stereo audio system composed of two microphones mounted on Yuppy's ears, IR sensors for obstacle avoidance, an air pressure sensor used to model simple touch perception in the form of painful and pleasurable stimuli, a pyro sensor aligned with the top camera used to detect changes in temperature due to the presence of people, and a simple proprioception system.

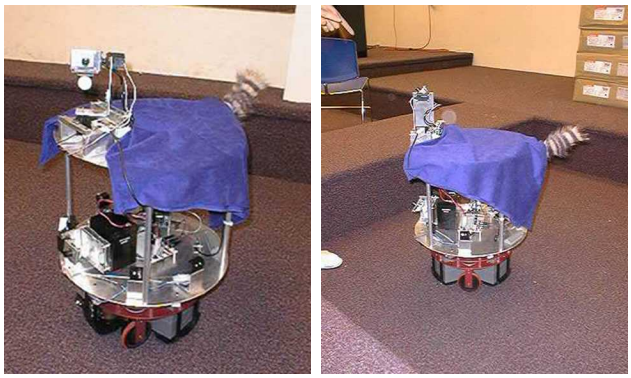

Figure 4 Yuppy, an Emotional Pet Robot

The Drive System

Yuppy's Drive System is composed of four different drives: *RechargingRegulation*, *TemperatureRegulation, Fatigue*, and *Curiosity*, each of which controls internal variables representing the agent's *battery*, *temperature*, *energy*, and *interest* levels, respectively.

The Emotion Generation System

Yuppy's Emotion Generation System includes emotional systems with *Natural Releasers* for the set of basic emotion families described before. Some of the most representative releasers and their associated Emotional Systems are presented below.

Interactions with Drive Systems.

Unsatisfied drives produce *Distress* and *Anger; Distress* is also produced, to a lesser extent, when drives are oversatisfied; and Satiation of drives releases *Happiness*.

Interactions with the environment. In general, all pink-reddish objects release *Happiness* to some extent, although pink Styrofoam "bones" elicit the most*;* Yellow objects found in the environment release *Disgust;* Darkness releases activity in the *Fear* Emotional System; and loud noises release *Surprise*.

Interactions with People. When people interact with the robot, two of the several possible interactions include petting and disciplining the robot. These actions will generate representations of pleasure and pain, respectively. Pleasure releases *Happiness*, and pain releases *Distress*, and *Anger.*

The Behavior System

The robot's Behavior System is composed of a distributed network of approximately nineteen different self-interested behaviors, directed in most part toward satisfying its needs and interacting with humans. Examples of such behaviors include *"search for bone", "approach bone", "recharge battery", "wander", "startle", "avoid obstacle", "approach human", and "express emotion"*.

The user interacts with the robot in two different ways: First, by controlling its affective style, which is done by tweaking the different parameters described above (e.g., thresholds, gains, or inhibitory and excitatory connections) for each Emotional System. And second, by providing stimuli for the robot, whether in the form of internal stimuli, such as modifying the levels of internal variables, or external stimuli, such as hitting the robot, showing the bone, and so on.

Using the model described before, both the simulated and physical Yuppys will exhibit emotional behaviors under different circumstances. For instance, when the robot's *Curiosity* drive is high, Yuppy wanders around, looking for the pink bone which people may carry. When it encounters one, the activity of the *Happiness* Emotional System increases and specific behaviors, such as *"wag the tail"* and *"approach the bone"* become active. On the other hand, as time passes by without finding any bone, the activity of its *Distress* Emotional System rises and appropriate responses, such as "droop the tail", get executed. Similarly, while wandering around, it may encounter dark places which will elicit fearful responses in which it backs up and changes direction.

Besides regulating action-selection and generating emotional behaviors through primary emotions, Yuppy learns secondary emotions which are stored as new or modified cognitive releasers based on the associative network model described before. For instance, consider a scenario in which a person has a bone in her hands and entices Yuppy to approach and "get it". After doing so, the person might pet or discipline the robot, in which case Yuppy will create positive or negative emotional memories with respect to people, and future selection of behaviors such as approaching or avoiding them will be influenced.

Conclusions

We have proposed Emotion-Based Control as a complementary approach to experimental studies for considering issues related to emotional processing in the brain.

A biologically plausible framework that supports Emotion-Based Control has also been presented. This framework captures important aspects of emotional systems and integrates these with separate models for perception, drives, behavior, and motor control. Its design is suitable for an incremental approach in which new features are implemented as more is learned about the different works of emotional systems.

Finally, we have suggested that developing simple Emotion-Based Control systems for robots operating in the real world can prove useful in addressing questions and issues about the nature of emotions as adaptive systems.

Acknowledgments

Support for this research was provided in part by Yamaha Motor Corporation. Work on the physical Yuppy robot is carried out in conjunction with Professor Rodney Brooks, Charles Kemp, Makoto Yoshida, and Milyn Moy.

References

Abelson, R.P. 1963. Computer Simulation of "hot cognition". In: Tomkins, S.S., and Messick, S. eds. *Computer Simulation of Personality.* New York: Wiley.

Bates, J. 1994. The Role of Emotion in Believable Agents. *Communications of the ACM* 37(7):122-125.

Bechara, A., et al. 1997. Deciding Advantageously Before Knowing the Advantageous Strategy. In: *Science*, 275, 1293-1295.

Blumberg, B., Todd, P, and Maes, P. 1996. No Bad Dogs: Ethological Lessons for Learning. In: *Proceedings of SAB96*. Cambridge: MIT Press.

Blumberg, B. 1994. Action-Selection in Hamsterdam: Lessons from Ethology. In *Proceedings of SAB94*, 108-117. Brighton, England: MIT Press.

Colby, K.M. 1975. *Artificial Paranoia*. New York: Pergamon.Adolphs, R et al. 1996. Neuropsychological Approaches to Reasoning and Decision-Making. In: Damasio, A., et al. Eds. *Neurobiology of Decision-Making*. Berlin: Springer-Verlag.

Damasio, A. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Gosset/Putnam.

Ekman, P. 1992. An Argument for Basic Emotions. In: Stein, N. L., and Oatley, K. eds. *Basic Emotions*, 169-200. Hove, UK: Lawrence Erlbaum.

Elliot, C.D. 1992. The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System. Ph.D. Thesis, Institute for the Learning Sciences, Northwestern University.

Fox, N., and Davidson, R. 1986. Taste-Elicited Changes in facial Signs of Emotion and the Asymmetry of Brain Electrical Activity in Human Newborns. In: *Neuropsychologia*, 24, 417-422.

Frijda, N. 1986. *The Emotions*. Cambridge, UK: Cambridge University Press.

Izard, C. 1991. *The Psychology of Emotions*. New York: Plenum Press.

Johnson-Laird, P. and Oatley, K. 1992. Basic Emotions, Rationality, and Folk Theory. In: Stein, N. L., and Oatley, K. eds. *Basic Emotions*, 169-200. Lawrence Erlbaum.

Kitano, H. 1995. A Model for Hormonal Modulation of Learning. In: *Proceedings of IJCAI-95*. Montreal.

LeDoux, J. 1993. Emotional Memory Systems in the brain. *Behavioral and Brain Research*, 58.

LeDoux, J. 1996. *The Emotional Brain*. New York: Simon and Schuster.

Maes P. 1995. Artificial Life meets Entertainment: Lifelike Autonomous Agents. *Communications of the ACM. Special Issue on Novel Applications of AI*.

Minsky. M. 1986. *The Society of Mind*. New York: Simon & Schuster.

Panksepp, J. 1995. The Emotional Brain and Biological Psychiatry. *Advances in Biological Psychiatry*, 1, 263-286.

Pfeifer, R. 1988. Artificial Intelligence Models of Emotion. In: Hamilton, V., Bower, G. H., and Frijda, N. eds. *Cognitive Perspectives on Emotion and Motivation*, 287-320. Netherlands: Kluwer.

Picard, R. 1997. *Affective Computing*. MIT Press.

Reilly S. 1996. Believable Social and Emotional Agents. Technical Report, CMU-CS-96-138, School of Computer Science, Carnegie Mellon University.

Staddon, J. 1993. On Rate-Sensitive Habituation. Adaptive Behavior, 1 (4). MIT Press.

Velásquez, J. 1996. Cathexis: A Computational model for the Generation of Emotions and Their Influence in the Behavior of Autonomous Agents. S.M. Thesis. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Velásquez, J. 1997. Modeling Emotions and Other Motivations in Synthetic Agents. In: *Proceedings of AAAI-97*.

Velásquez, J. 1987. When Robots Weep: Emotional Memories and Decision-Making. In: *Proceedings of AAAI-98*.