When Robots Weep: A Computational Approach to Affective Learning

by

Juan David Velásquez

B.S., Universidad EAFIT (1993)
S.M., Massachusetts Institute of Technology (1996)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2007

Author .....
Department of Electrical Engineering and Computer Science
May 24, 2007

Certified by .....
Rodney A. Brooks
Panasonic Professor of Robotics
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

When Robots Weep: A Computational Approach to Affective Learning

by

Juan David Velásquez

Submitted to the Department of Electrical Engineering and Computer Science
on May 24, 2007, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

This thesis presents a unified computational framework for the study of emotion that integrates several concepts and mechanisms which have been traditionally deemed to be integral components of intelligent behavior. We introduce the notion of *affect programs* as the primary theoretical constructs for investigating the function and the mechanisms of emotion, and instantiate these in a variety of embodied agents, including physical and simulated robots.

Each of these affect programs establishes a functionally distinct mode of operation for the robots, that is activated when specific environmental contingencies are appraised. These modes involve the coordinated adjustment and entrainment of several different systems—including those governing perception, attention, motivation regulation, action selection, learning, and motor control—as part of the implementation of specialized solutions that take advantage of the regularities found in highly recurrent and prototypical environmental contingencies.

We demonstrate this framework through multiple experimental scenarios that explore important features of the affect program abstraction and its function, including the demonstration of affective behavior, evaluative conditioning, incentive salience, and affective learning.

Thesis Supervisor: Rodney A. Brooks
Title: Panasonic Professor of Robotics

Acknowledgments

The road to this point has been filled with magnificent experiences, most of which have resulted in much learning, both academic and personal. In May of 2002, I took a leave of absence from MIT and went back to Colombia, my home country, where I founded a start-up company in the finance sector. The idea was to return after a few weeks and wrap things up. Well, this idea took over four years, time which was certainly rewarding and filled with learning as well, but during which I always yearned to come back. This finally happened in July of 2006, when I returned to the Humanoid Robotics Group at CSAIL, where I completed this work.

Returning to MIT would simply not have been possible without the encouragement and support of my advisor, Rod Brooks. If only emotion terms could be descriptive enough, I would use them to the fullest extent to express the immense gratitude, respect, and affect I have for my relationship with Rod. His groundbreaking and revolutionary thinking, not only with respect to AI, but also in regards to science and entrepreneurship, has been a great source of inspiration to me. I especially thank you, Rod, for your generous support during all these years, for your mentoring, and especially for helping me find a way back to where I really wanted to be. These past months since my return have been, without a doubt, the most interesting and gratifying of my time at MIT. For all of this, I am forever grateful. It feels great to have the number 27 attached to my back. I will wear it with both pride and humility, which are some of the greatest things I have learned these past years.

Throughout my time at MIT, I have also been privileged to interact with several faculty members, who have influenced my thoughts. In particular, I want to thank the members of my thesis committee for their invaluable feedback, and for their encouragement to pursue further the ideas I present in this work. Roz Picard has been an incredible mentor and friend, and as a pioneer in this field, a great inspiration

as well. I thank you, Roz, for all your support and for your interest and patience whenever we discussed the many low-level details of our models of affect. Patrick Winston is a remarkable teacher in the truest sense of the word. He really cares about helping students become what we aspire to be. I thank you Patrick for accepting to be part of my thesis committee when I returned, and for the many sessions in which you helped me find better and more clear ways to communicate my work. I would also like to thank many other professors from which I have benefited while at MIT, including Ann Graybiel, Leslie Kaelbling, Tomás Lozano-Pérez, Daniela Rus, Randy Davis, Lynn Stein, Pattie Maes and Marvin Minsky. To Pattie and Marvin, my most sincere thanks for sending me on the path of emotions when you advised me throughout my S.M. work.

Special thanks to Marilyn Pierce, whose natural kindness and friendly support throughout grad school were paramount. Marilyn, without your help (and your signature!) this thesis could not be complete, literally.

Cognitively, I was always aware of the many challenges that finishing this journey would entail, but I only came to realize what they really meant, as I experienced them emotionally during these past few months. I certainly could not have successfully finished this process had it not been for all of my wonderful and supporting friends. In particular, I would like to give my wholehearted thanks to Eduardo for helping me in more ways than one throughout this final push and for letting me abuse of your Matlab wizardry! I thank you the most, however, for remaining true to yourself during all these years and for allowing me to experience your friendship, which has made the good and the interesting times, great memories to keep. I also want to thank my amazing and spontaneous friend Lijin. Even now that you are far away, your beautiful friendship is cherished and serves as a constant reminder of how wonderful human beings are. I cannot thank you enough for your constant encouragement and your permanent presence throughout this time. Undoubtedly, coming back to MIT

to find you again has been one of my biggest rewards. I most certainly want to give my deepest thanks to Ann, for helping me in uncountable ways. Your kindness has always been present, from the first attempts to reconnect with the lab, to your warm support throughout this process, including the last "completion vibes", which were much needed. You were a marvelous surprise to find at MIT, Ann. One for which I am enormously grateful. Your warmth and strong spirit is a source of light to us all.

Most heartfelt thanks as well to Charlie Kemp, with whom I shared and enjoyed not only an office and several projects throughout much of our time at the lab, but also a wonderful friendship filled with great philosophical discussions about life. Thanks to all of my other lab-mates and friends including the sweet Jessica, Lilla, Aaron, Iris, Una-May, and from earlier times Scaz, Cynthia, Robert Irie, Matt Williamson, the cool Maddog, Bryan Adams, Milyn Moy, Artur, and especially Paul Fitzpatrick.

These acknowledgments would not be complete if I did not give my thanks to the wonderful Varun. I thank you for keeping me aware of my deadlines, for your constant support, for our great philosophical discussions, for teaching me about Hinduism and for staying late and patiently survive my defense practice talks. Also, many thanks to Gabriel Gómez, who came to the lab at a crazy time for me and found so many ways to help me in this last stretch. Thanks Iuliu for helping me with my practice talks in such a disinterested way. Many thanks to all other members of my surrogate group, including Paulina, Marty, Mac, Carrick and the rest of the crew.

Above all, I would like to thank my family, both immediate and extended. Without your encouragement, your companionship, your confidence in me, and your immense love, all of this would be meaningless, in the truest affective sense. Had I not gone back to Colombia to reencounter you, would have meant not to reencounter myself. To my amazing wife, Veronica, my infinite gratitude and respect. Mivi, I can only say that being able to feel my life, through your heart, has been the most wonderful gift I have ever experienced. All my love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *There can be no knowledge without emotion. We may be aware of a truth, yet until we have felt its force, it is not ours. To the cognition of the brain must be added the experience of the soul.*
> — Arnold Bennett (From The Journals of Arnold Bennett (1932), entry for 18 March 1897)

Imagine for a moment what would your life be like if you could not experience the joy of rewarding events or the fear that warns you about dangerous contingencies. What would happen if the events you experience in the world had no particular significance to you? What if you did not, or better put, *could not* prefer some things over others, nor could you choose among them? What if you would never experience the impulse or motivation to act upon this world or decide a course of action based upon events presented to you? Imagine this world where all events and stimuli were presented simultaneously to you, and you had no means for selectively attending to them other than using sensory-perceptual properties such as how close to you they appear to be, how much they move, or what their shape is, but with no information as to whether they are important to you or not. What if you could not derive meaning and learning from the relationships between these events? What if it was impossible for you to determine what your current state is, in relation to this world, nor signal this state

to others? Such a world would be one devoid of emotion. An *affective flatland* that would be so strange and detached from our usual perceived reality that it is certainly difficult to imagine, and perhaps even more difficult, if not impossible, to live in.

This thesis presents a unified computational framework for the study of emotion, and affective phenomena in general, based upon the construction of computational models that integrate several concepts and mechanisms that have been traditionally deemed as integral components of intelligent behavior. Our approach is based on the notion of *Affect Programs*, adaptive biological schemas that have proven useful, throughout our evolutionary past, in helping us deal with life and survival-related fundamental situations.

Given what we have learned thus far with respect to the disparate set of phenomena referred to as emotions and affect, the time is now ripe for computational approaches that supplement and reflect upon many of the more theoretical issues that have been addressed by research in other disciplines, such as Philosophy, Psychology, and more recently Neuroscience, where the emotions have had a more salient interest among researchers.

A computational approach to emotions offers us the opportunity to study these phenomena, with the appropriate scientific rigor, yet without many of the challenges that accompany their study from the perspective of other disciplines. Furthermore, through the construction of models and mechanisms, including building robots embodied in the real world, we have the opportunity to test hypotheses and even question our deepest assumptions with respect to affective phenomena, their functions, and their underpinnings, all of which might ultimately prove useful in our quest for engineering intelligence.

1.1 Multiple Stages for Affective Learning

Based on evidence stemming from multiple disciplines that have studied affective phenomena from very different standpoints, we propose that affective learning occurs in a sequential set of events that take place when an organism is exposed to signals or cues that predict affectively significant events, which by our definition of affect, thus correspond to biologically significant events (i.e., events that are somehow related to fundamental life tasks). The proposed model is depicted in Figure 1-1. This model indicates: (a) the hypothetical psychological constructs that occur at each stage; (b) possible behavioral correlates; and (c) the computational components that are associated with, and support the events in each stage. Although the model suggests a sequential order, the fact that the behavioral components follow this order is not meant to imply that the brain mechanisms underlying these different behaviors and stages also function in a sequential order. In fact, quite the contrary occurs and depending on what is being studied, parallel processing (and even competition) occurs in the underlying brain mechanisms. The main issue that we want to point out with this model is that the formation of associations based on affectively significant events, produces predictable behavioral changes that are associated with multiple learning systems involving a variety of processes. Describing these processes is what will occupy the rest of this thesis.

First, however, a general description of the model is in order. The first stage shown in Figure 1-1 represents an "attention" (also often called "arousal") stage that involves the response to novel stimuli, most usually associated with "orienting responses" (ORs). In addition to the relational behavior associated with the ORs, novel stimuli also elicit complex autonomic changes (e.g., changes in heart rate and blood pressure, hormonal release into the bloodstream). If the eliciting stimulus is not of affective significance (either directly or because it has been associated with a

stimulus that is), the OR habituates until it is no longer generated. In contrast, the co-ocurrence of otherwise *neutral* stimuli with affectively significant ones, would elicit specific behaviors associated with the next stage. This second stage corresponds to the first step in the development of associative affective learning. In this stage, neutral stimuli acquire affective significance and thus become important to the organism, as they become reliable predictors of events of biological importance. The development of affective significance associations is related to the appearance of learned specific and non-specific responses, as will be discussed in more detail later. Non-specific responses have been referred to as *preparatory*, since they occur regardless of the nature of the learning contingencies and presumably in order to prepare the organism for the specific events that will follow. For instance, predicting the presence of a predator through signals in the environment, might trigger a set of preparatory responses that include accelerating the heart rate, releasing specific hormones such as adreno-cortisol, and sending blood to the limbs, all in preparation for escape. Specific responses, on the other hand, have been referred to as *consummatory*, as they end the preparatory phase of behavior and consist of actions that are specific to the affective event (e.g., escaping once the predator is actually detected). Thus this second stage is perhaps the most important stage in affective learning, as it is in here that stimuli are "coded" with affective value and meaning is ascribed, at its simplest level, to the events occurring in the world. In the third stage, this same kind of meaning is ascribed, but this time to actions. In this stage, flexible responses are learned based on the association of the outcome of an action (in affective terms), and the action itself. This stage comprises a set of highly complex events for which there is yet no complete understanding. However, we do know that the learning of more flexible responses starts to occur and when the events that led to this learning are repeated in a predictable manner, learning reaches asymptotic levels and the production of these responses become habitual. Finally, in the fourth stage, these habitual responses are organized into

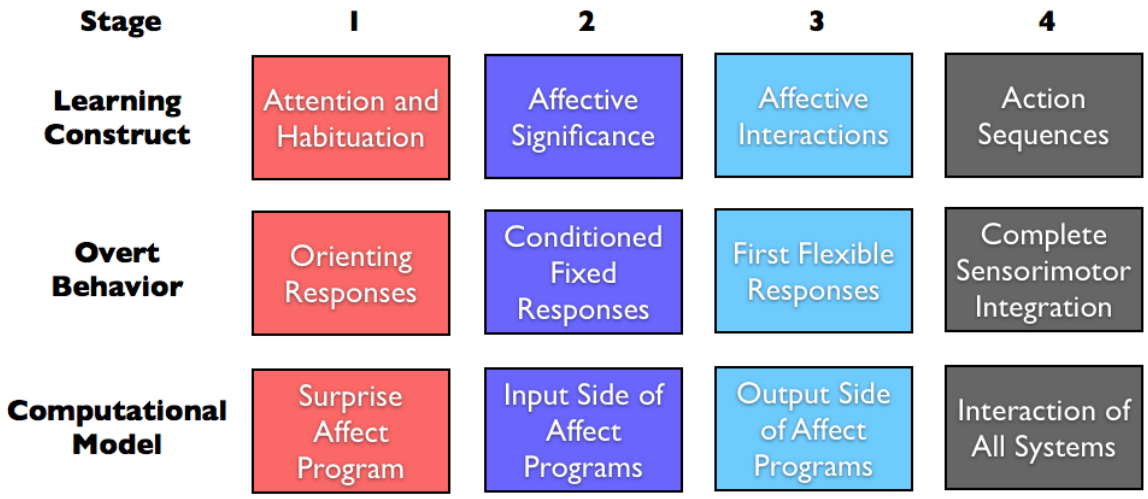

Figure 1-1: A multi-stage model of affective learning.

behavioral "chunks" composed of sequences of behaviors that represent the highest form of sensorimotor integration.

The scope of this thesis is limited to the first two stages of affective learning, and part of the third phase. The chapters that follow will describe these stages in more detail and will propose the use of the affect program abstraction as the primary construct to implement and understand these stages from a computational standpoint.

1.2 Contributions

This thesis makes the following contributions:

1. From a theoretical perspective, and perhaps one of the main contributions of this work, is a reconceptualization of the notion of emotion. We depart from traditional accounts that focus on the *experience* of emotion, which view emotions as states organisms can be in, and instead offer a perspective on emotion that describes these phenomena as functionally distinct processes which imple-

ment specialized solutions to prototypical situations that organisms (or robots) face regularly in their environments.

2. We present a unified computational framework for the study of emotion that accounts for different affective phenomena, including a variety of emotions and emotional behavior, as well as simple notions of moods and temperament. This framework further integrates these phenomena with several notions traditionally deemed to be integral components of intelligent behavior.

3. As the main component of this framework, we introduce a novel computational construct for an *affect program* as a biologically plausible abstraction for emotion. The primary function of affect programs is to mediate, control and synchronize the activities and interactions of several subprograms, including those that govern perception, attention, physiological regulation, goal selection, motor control, expressive social communication processes, action selection and learning, and so forth. Each of these affect programs establishes a mode of operation for the robot, which involves the coordinated adjustment and entrainment of these subprograms (responses) so that the whole system exhibits coherent behavior as a response to the confrontation with specific eliciting situations.

4. We present a model for incentive salience, which attributes motivational properties to stimuli and actions that signal the occurrence of events of emotional (biological) significance. An incentive salience approach contrasts with other views that propose that reward or incentive learning, as mediated by the brain's mesolimbic dopamine systems, is based upon global teaching signals that code for the errors in the prediction of reward. These views have found further acceptance in the neurosciences given that elegant computational counterparts, such as the reinforcement learning models, seem to work in a similar manner. However, an incentive salience approach, such as that proposed in this thesis,

provides an alternative explanation for the activity of these same brain systems and through simple and localized learning rules, together with the organizational principle that separates action into preparatory and consummatory behaviors, can account for some of the evidence seen in experimental paradigms. Something that reinforcement learning models, at least in their original form, cannot account for.

5. We propose an agent architecture that follows an affective-based decomposition and which provides a novel alternative to the control of intelligent robotic systems. In this approach we suggest that a different organization of action is pursued, one based not upon the desired external behaviors of the robots, but rather on the set of prototypical fundamental situations that the robot will encounter and view these as affective situations for which a set of coordinated responses dan be made available which deal with (and perhaps solve) such situations, much in the spirit of these biological schemas we have referred to herein as the *affect programs*.

6. Finally, we suggest a multi-stage model of affective learning that ties evidence from psychology and neuroscience regarding classical paradigms of associative learning and bridges these notions with a possible computational substrate, in the form of the affect programs abstraction.

1.3 Methodology

The primary goal of this research is to investigate affect from a computational perspective. We will argue throughout this thesis that robotics research can address scientific questions about the nature of affective processing in humans and animals. To this end, we presents a novel methodology for building robots that follows an affect-based

decomposition. This methodology, which extends previous work on behavior-based robotics (Brooks, 1986) and humanoid robotics (Brooks, Breazeal, Irie, Kemp, Marjanović, Scassellati & Williamson, 1998), stresses the use of computational models of affective processing to build and control intelligent systems that are capable of performing a variety of complex behaviors in the real world.

Underlying this research is the notion that affect is inherently intertwined with several attributes that we associate with intelligent behavior, such as multimodal sensory integration, natural social interactions, and development (Brooks et al., 1998). This idea is supported by surmounting evidence regarding the pervasiveness of affect in perception, memory, attention, behavior selection, and learning (Damasio, 1999; Gallagher & Chiba, 1996; LeDoux, 1996; Graybiel, 1998; Packard & Teather, 1998; Panksepp, 1998).

1.3.1 Integration

Work in this thesis argues for the need to build complete systems that go beyond shallow models of emotion, but rather include deep models of affect that act as the main programs that mediate perception, attention, motivation, behavior, learning, and motor control. The integration of such a variety of systems is without a doubt a difficult challenge to overcome. Thus, appropriate abstractions and encapsulation mechanisms are necessary. As suggested above, this thesis proposes the idea of an affect program as a useful abstraction that offers a natural decomposition for this task. Affect programs, which are defined in more detail in Chapter 2 and an implementation is described in Chapter 6, integrate a variety of sensory information and synchronize a number of functions in response to biologically significant events. Thus, it argues, they are well suited to act effectively as an integration mechanism by which activity in many different systems is bound together in a coherent manner.

1.3.2 Natural interaction

In a similar manner, this work argues that the expressive components of affect play an interesting role in communicating internal states and promoting natural interactions. By endowing our robotic systems with such expressive skills we can capitalize on people's natural abilities to support social interaction. This provides a more natural way for human-machine interaction as well as novel approaches for learning (Breazeal & Velasquez, 1998) and opportunities to understand the nature of affective signals (Velásquez, 1999). Supporting these ideas, Breazeal (2000) has made a compelling case in demonstrating the power of affect as a modulator of social interaction.

It should be noted however, that these communication signals and their applicability to social interaction, is but one of the many responses integrated by affective processing. While we consider this an interesting research problem, the work in this thesis rather focuses on deeper issues related to the computational problems that organisms must face when dealing with biologically significant events in their environments, and how affective processing is useful for integrating and coordinating a set of responses that deal with such situations.

1.3.3 Development

Development is an extensive and gradual process by which organisms acquire increasingly elaborate behaviors and new abilities. Recent work in robotics has begun to deal with issues of cognitive development (Scassellati, 1998). Affective development, on the other hand, is a new challenge that lies at the core of this thesis.

This work focuses on different learning mechanisms that allow the affect program abstraction to be extended and used as a building block for the construction of action repertoires (Graybiel, 1998). These learning mechanisms depart from the norm of traditional learning theory in the sense that they are not general-purpose, but rather

biased mechanisms that learn about emotionally significant aspects of the world that are relevant to any given affect program. It further argues that these constructions (basic affective processing with extended learning mechanisms) may account for some of the apparent cognitive-affective interactions that are believed to be part of the so-called higher order emotions.

To this end, this thesis investigates the use of different learning strategies, including both nonassociative and associative learning to promote development. It proposes an architecture that incorporates multiple learning mechanisms that are distributed in functionally meaningful ways across the affect programs framework. Of particular interest, as it is described below, is the use of affective learning schemes such as incentive learning and reward-based learning, which can focus the organism's attention and reduce the learning space by providing information concerning when to learn and what to learn.

1.3.4 Inspiration from Biology

This work draws upon ideas from different disciplines that have a longer tradition of studying affect, such as ethology, psychology, and neuroscience. Furthermore, it attempts to integrate and reconcile knowledge derived from each of these disciplines into computational models and abstractions that are biologically plausible and facilitate the understanding of affective processing in humans and animals.

In particular, it reviews and draws upon work concerning the possible neural substrates for some of the main issues considered in this thesis, such as the functional roles of the basal ganglia, hippocampus, amygdala, and other brain structures, in the processing of affect, behavior selection, and learning (Damasio, 1994; LeDoux, 1993; Graybiel, 1995; White, 1997).

However, it is important to note that while it is interesting to examine these

findings and models, the rapid progression of research in this area means that it is likely that many of these findings will be reevaluated or even dismissed. Thus, while this work is related to experimental studies of the nervous system, an effort has been made not to fall into the details of modeling particular structures or systems. In other words, the purpose of this work is not to provide a model of any of these brain structures, but rather to make use of our knowledge on these systems to support the ideas and models behind this research.

1.4 Studying Affect from a Computational Perspective

Emotion has traditionally been studied by trying to make experimental participants feel emotions or experience parts of emotion in the laboratory, and then measuring, through self-report and other methods, its different components, and indicators. However, given ethical guidelines and technological constraints, the induction of emotion in subjects is challenging (e.g., one must not exceed the intensity or kind of emotion that a subject would experience in everyday live). Thus, studying emotions, at the different levels of abstraction and at extreme ends of the affective range can pose difficulties which require new approaches.

1.4.1 Computational Models

For anyone who restricts the notion of emotion to that of the *experience* of emotion, it may seem strange and perhaps even impertinent that we commit to exploring the underlying computational architecture of emotions.

As Cosmides & Tooby (2000) have recently indicated:

> *It may strike some as odd to speak about love or jealousy or disgust in*

> *computational terms. "Cognition" and "computation" have affectless, flavorless connotations. In everyday language, the term "cognition" is often used to refer to a particular* subset *of information processing—roughly, the effortful, conscious, voluntary, deliberative kind of thinking one does when solving a mathematics problem or playing chess: what is often called "cold cognition".* — Cosmides & Tooby (2000, p. 98)

However, considering the definition of emotions put forward above (and detailed in Section 2.2.4), it should be apparent that studying emotion from a computational perspective is not only possible, but also appealing, given that by doing so, we should be able to delve deeper into the set of computational problems that organisms face while surviving in their environments, and devise new computational methods and models that contribute to the understanding of such important aspect of our lives.

Computational modeling of emotion, to the extent that it can allow us to explore questions that would otherwise be unethical or difficult to address in experiments with humans and animals[1], presents itself as an excellent tool and as the primary theoretical method for investigating the function and the mechanisms of emotion, and affective processing in general. These models allow us to capture the essential features of emotional systems at multiple levels of abstraction as well as at several spatial-temporal scales, from the inner workings of emotional appraisal, to the network coupling between emotional systems and the coordination and modulation of other systems be they motivational, attentional, learning or motor control.

By drawing inspiration, and tying together into our computational models the results and evidence gathered from multiple disciplines that have consistently studied

[1]Note, however, that recent approaches in the field of affective computing have introduced clever ways for elucidating the mechanisms of affective processing without causing harm, in the emotional sense, to the subjects that participate in such experiments (Picard, 1997). Likewise, advances in neuroimaging have also allowed for other, less intrusive, ways to study the neural underpinnings of such processing.

emotion over the years, such as Philosophy, Psychology, Neuroscience, and Ethology, we can speculate and test hypotheses that can be directly verified by past or current experiments in any of these disciplines. Furthermore, as it has been the case with computational modeling in other fields (e.g., Neuroscience), through a computational perspective we may be able to provide new perspectives to many of the problems that are being currently addressed, and to devise new methods and models that can be further confirmed through future experimentation.

1.4.2 Affective Robotics

In this thesis, we attempt to avoid the many general theoretical debates that have surrounded the study of emotion over the years and instead focus on understanding specific problems, such as understanding specific emotions and their conditions, what they are, how are they implemented, and how do they interact with other emotions and processes we usually attribute to intelligent behavior. To this end, it is not surprising that we adhere to the views that argue for the existence of discrete and biologically determined emotions, such as the Affect Programs theory described in Chapter 2, which views emotions as the mechanisms that allow organisms to deal with very fundamental life- and survival related tasks that have been recurrent throughout our evolutionary past. The main issue that stands out from this view is that it considers emotions to be functionally discrete, information-processing systems whose integrated mode of operation functions as a solution designed to take advantage of the particular structure of recurrent situations or triggering conditions to which the emotion corresponds (Cosmides & Tooby, 2000).

From this perspective, in order to understand specific emotions, however, we need to understand the class of problems faced by organisms as they are situated in their environments, evaluating events, interacting and selecting significant stimuli, and de-

termining solutions for the multiple contingencies they face. It is in this respect, that the use of robotic platforms provides an ideal experimental platform for a computational approach to the study of emotion. Robots are situated in the real world, which is often uncertain and dynamic, they have to deal with noisy sensors and actuators, and constantly face multiple and complex challenges, including the tasks of attending to and selecting relevant stimuli, determining the best course of action given any situation, and deciding when and what to learn about their environment in order to adaptively achieve (or maintain) some specific goals. These are precisely the class of problems that *constitute* emotions and represent their functional purpose. Thus, *Affective Robotics* may lead to a great number of scientific payoffs, as we will hope to describe some of them in the following chapters of this thesis.

1.5 Overview of the Thesis

The rest of this thesis is organized as follows:

Chapters 2 and 3 provide the conceptual background regarding the notion of affect and emotion and what their possible function is, as it has been described by researchers in the field. These chapters set up the stage and the conceptual framework upon which the rest of this thesis is based.

Chapter 4 reviews some of our current knowledge with respect to the neural mechanisms underlying emotional processing and affective learning.

Chapter 5 briefly describes the robotic platforms used as testbeds for all of the ideas that comprise the computational framework that is proposed in this thesis.

Chapter 6 presents our approach to engineering affect. It describes the main computational abstractions that comprise the Cathexis framework and introduces the notion of the affect program.

Chapter 7 builds upon the previous chapter in order to demonstrate how the

proposed framework can account for different affective phenomena, and how it can be used in the organization and control of behavior.

Chapter 8 describes our multi-stage model for affect learning, which includes both nonassociative and associative learning strategies that build upon the activity of affect programs to learn relationships between the different contingencies that organism face as they interact in their environments.

Chapter 9 demonstrates how affect can be used to modulate attentional processes, by means of the interactions between different affect programs and relying upon incentive salience properties as implemented in our computational framework.

Chapter 10 ends with a summary of the contributions of this thesis. It also examines related work in the synthesis of affect and affect-related models, especially those concerned with reward or incentive learning, and outlines possible extensions to the proposed framework as part of future work. Finally, in this chapter we speculate on a number of theoretical issues in the engineering of affect, and how the reconceptualization of emotion proposed in this thesis might lead to different views of the mind.

Chapter 2

The Nature of Affect

What characterizes the class of things we commonly refer to as *emotion*? What are their distinctive features? What elicits them and how are they produced? How do they differ from those of other affective phenomena such as moods and temperament? How can we study these phenomena? These are some of the questions posed by researchers of emotion and by those who have been interested in obtaining an integral understanding of the mind. Many different answers have been proposed, some of which seem to be in agreement with our folk use of affective terms, and yet others which seem ill-defined unless they are construed within the same specific stream of thought in which they were first produced.

A difficulty in reaching consensus in the characterization of affect and emotion is due, in part, to the ample scope of these affective phenomena, which opens up a definition space that can be explored through many different perspectives, and results in varied descriptions at multiple levels of abstraction. Ultimately, whatever characterization of affect is chosen, it usually shares the aim of understanding the phenomena and the conditions for their occurrence, but defining clear boundaries that distinguish one phenomenon from another has been, and still is, problematic. It is not uncommon to find researchers using the same terms (e.g., "emotion") to

refer to the study of very different classes of things, or to find others who provide new terms and definitions, when they are in fact studying the same kind of affective phenomena.

We would argue that the class of things that we commonly refer to as *emotion* comprises a set of disparate phenomena that have been grouped together mainly by our folk use of the term. However, this grouping is not necessarily due to the fact that all of these phenomena share distinct features, or can all be characterized in the same way, but rather because doing so is useful for our social communication as it offers explanatory parsimony with respect to our everyday understanding of ourselves (i.e., our "folk psychology"). This is not to say that *some* of these affective phenomena do not share distinct features at all. They do, and in fact the approach presented in this work is based precisely on the notion that some of the emotions can be characterized in such a way, and thus are amenable to study from a computational perspective.

This chapter aims to propose some answers to the aforementioned questions, thereby setting the conceptual framework upon which the work on this thesis is built. It briefly reviews some of the most common approaches to the study of emotion, focusing on the affect program theory, which, from our perspective, is the most compelling theory of emotion that resolves some of the issues involved in the characterization of the phenomena and provides an appropriate level of abstraction that is suitable for computational modeling. Readers amply familiar with the different models and theories of emotion, their explanatory power and limitations, may wish to skip this chapter.

2.1 Defining Affect and Emotion

There is very little agreement about the definitions and terminology used to describe emotion, especially as they arise from different disciplines such as Philosophy, Neu-

roscience, Psychology, and more recently Artificial Intelligence. As an example of this, Kleinginna & Kleinginna (1981) described the lack of consensus with respect to the definition of emotion, while they considered 92 different definitions given by researchers, together with 9 skeptical remarks—indicating that emotion is not a useful concept—and organized these according to different categories, ranging from their relation to physiological components or emotional/expressive behaviors, to definitions based upon motivational and adaptive views. For such a "commonly understood" and used term, the differences in the researchers' perspectives is simply astonishing!

To reduce the level of complexity of this definition space, researchers often decompose whatever notion they hold of emotion into basic features that are more suitable to study, and thus use these features as the indicators *sine quibus nons* of emotions and of how they should be measured. For instance, some researchers define emotions as feeling states, and thus measure these states by asking the subjects about the "level" of emotion they are experiencing (Scherer, 1984). Other common definitions of emotions include those based on the physiological reactions that they produce in the peripheral nervous system (Ekman, Levenson & Friesen, 1983) or those that consider the overt behavioral responses that they generate, including facial expressions and their feedback (Ekman & Friesen, 1986; Ekman, 1994*c*; Izard, 1971; Izard, 1994; Zajonc, 1985). Finally, the large majority of modern researchers of emotion, define them in terms of a set of cognitive appraisals, attributions and judgments which are also measured through self-report and introspection (Frijda, 1986; Roseman, Spindel & Jose, 1990). Suffice it to say, the approaches that rely upon the use of introspective measures and self-reports can be problematic, as they make the important assumption that the results of emotional processing are accessible to the individuals who experience them (i.e., they are conscious), and that these individuals can reflect upon them and quantify their intensity. Not to mention that there is no guarantee that the reflections being made correspond in fact to the occurring emotions, and not perhaps

to the results of other mechanisms, such as evaluative or language processes, to name a few possibilities. Notwithstanding, these are by far the most common approaches used in experimental and social psychology to measure emotions. Obviously, the approach is limited to humans, as it is easily seen how this would be difficult to replicate with other species (but see Section 3.3.1 for possible approaches to this end).

Another important difference among definitions of emotion lies upon the distinctions made with respect to its causes. As it will be described in section 2.2 below, some researchers believe that emotions are elicited by processes of evaluations (called *appraisals*) and attributions that relate environmental contingencies to the ongoing needs and goals of the organism making the appraisal (Frijda, 1986; Scherer, 1993). Other researchers believe that all organisms are biologically "prepared" to respond to specific objects and situations with specific emotional responses (Darwin, [1859]/1998; Izard, 1977; Ekman, 1992; Öhman, 1986). Finally, still other researchers see emotions as being elicited by a combination of these (Johnson-Laird & Oatley, 1992).

Considering that such markedly different phenomena can be labeled as emotions, and that many definitions exist for these, we will attempt to avoid confusion by providing some distinctions among the different terminology. In this thesis, we shall refer to"affect" in the most general sense. That is, we will use this term to imply that notion which encompasses *all* of emotional phenomena[1]. Therefore, emotions, moods, motivations and other affective phenomena will be classified herein as specific kinds of affect.

We still need to operationalize these terms further so that our attempts to describe these phenomena from a computational perspective have some grounds for our future discussions. As a general notion, we will define affect as *the set of processes that incite action, based upon the evaluation of biologically significant contingencies, and thus*

[1]Notice this usage departs from the colloquial use of the term, which usually refers to the feeling of an emotion.

are motivational. As we progress downward in our analysis, we will supplement this definition with other basic features that will help us further in our characterization of specific kinds of affect, and in particular of emotion.

2.1.1 Emotions, Moods and Temperament

Emotion, moods, and temperament are all terms that have been used to describe different features of affective phenomena. Often, however, these terms are used interchangeably, which is a common source of confusion. In this thesis we shall try to avoid that confusion by drawing some distinctions among them.

For many, the primary distinction between moods and emotion lies in their duration. Emotions are thought to be brief, while moods are seen as lasting much longer. This is certainly not the only criterion for distinguishing between these phenomena. Other distinctions include the notion that when compared to the component view of emotions, moods might not include some of these components, such as a prototypical facial expression. For instance, it would be common to observe multiple "angry" faces when someone is in an "irritable" mood, and not necessarily a specific one as it would usually be the case with an emotion. Also, when considering the eliciting causes of emotion described in the previous section, it is clear that emotions have specific eliciting conditions, and from an intentional stance, an object of interest as well. However, this is not necessarily so for moods, which, perhaps due to their duration, is more difficult to ascertain such causes, assuming they exist in the first place. In addition, there seems to be important interactions between emotions and moods. Emotions can lead to particular moods and moods seem to either lower the threshold, or potentiate the arousal, of compatible emotions. For instance, when a person is in an irritable mood, the person might become angry more readily than usual (Ekman, 1994*b*).

Davidson (1994), suggests that the main distinction between emotions and moods

might be a functional one. While emotions are directly linked to, and may serve the purpose of biasing action, moods might be more tightly related to cognition:

> *The primary function of moods, on the other hand, is to modulate or bias cognition. Mood serves as a primary mechanism for altering information-processing priorities and for shifting modes of information processing. Mood will accentuate the accessibility of some and attenuate the accessibility of other cognitive and semantic networks* (p. 52).

Although this will be described in more detail in Chapter 6, it should be noted this is precisely the view we ascribe to when distinguishing between emotions and moods. Moods are modeled as the tonic activation of the same mechanisms that, when phasically activated, would correspond to emotions.

Finally, temperament refers to those long-term trait-like constructs that modulate an individual's reactivity to affectively significant events. It is likely that temperament is associated with differences in the nervous system that can persist over longer periods of time, such as differences in the quantity of receptors for certain neurotransmitters, differences in reuptake mechanisms, or differences in the interactions between the neural mechanisms that support emotion and other neural systems with which they communicate (Davidson, 1994; Panksepp, 1994) . Whatever the case, it is clear that the notion of affective style, as temperament is often called, relates to the emotional traits of individuals that are thought to be largely biologically inherited, and as such, partially predetermined. Some emotional tendencies clearly have heritability components, and this has been shown many times in genetic research with animals that are bred for different emotional features. For example, breeding selectively for aggression, fearful reactivity or even higher curiosity is successful in just a few generations (Robertson, Martin & Candy, 1978; Peciña, Cagniard, Berridge, Aldridge & Zhuang, 2003).

2.1.2 Emotions as Processes

Given the prominence and frequency of emotions in our everyday lives, it is not surprising that our "folk psychology" has led us to label them using verbal expressions, mostly single words, that in most languages correspond to nouns. Thus, in describing the emotions we usually treat them as "things". These things have come to represent, from a traditional standpoint, a set of internal physical states of an organism that are usually accompanied by physiological and behavioral changes in the body. For instance, during a state of fear, you might feel tense and agitated, your face will change, your brows will be raised and drawn together, the skin around your mouth will tighten and your lower lip will be compressed and inverted, your vocal expression will also be modified to increase its pitch level and range, as well as increasing your speech rate, your heart will beat faster, your muscles will tense, your breathing will change, your perspiration will increase and your overall skin temperature might feel colder, as blood is sent to the limbs in preparation for a response.

As we have suggested before, labeling emotions as states is useful from a social communication standpoint, but the analysis of emotions at a functional level might be better off if they were treated as the variable results of active processes (Mandler, 1984). When we see emotions as processes, the relationships between the different components of emotion become more amenable to study. We can more easily describe how the different components are organized, synchronized and controlled. Think of the previous example with the emotion called "fear". Instead of focusing on the state of the organism, we can focus on the organization and synchronization of the different processes that control muscle tension, mediate blood pressure and heart rate, release hormones into the bloodstream, generate prototypical facial expressions, modulate vocal expression, and more importantly, motivate action while biasing attention and promoting learning. In addition, several basic questions can be addressed, such as

which processes are connected to other processes? Which of these connections are due to the same eliciting conditions, and which are simple correlations? Which are functionally dependent, and which are the effects of a central command system, as the view of affect programs suggests?

Furthermore, this perspective facilitates the distinction between emotions and moods suggested above. When we view emotions in terms of processes, rather than states, we avoid needless discussions about the different category boundaries between these affects. Thus, the question of whether something is a mood or an emotion is readily solved when we think of these affective processes as being graded in intensity. This turns the distinction between emotions and moods into a continuum of "emotionness" and "moodness" (Scherer, 1984; Frijda, 2000).

Finally, the view held by some that suggests that emotions *must* be conscious (Clore, 1994), which has been rightly challenged by many (see Zajonc (1994) and Winkielman & Berridge (2004) for examples), becomes a less relevant issue when emotions are considered as processes[2]. Let us clarify this idea further. When we restrict the notion of emotion to that of the *experience* of the emotion (commonly referred to as *feelings*), we discard all of its other components, which are by no means less important.

Consider for instance the component that corresponds to the emotional evaluation (i.e., the evaluation of stimuli significance). This process undoubtedly lies at the center of emotions as it forms the basis of the elicitation, control and the expression of emotion. Furthermore, there is ample evidence that suggests this process occurs non-consciously, and at least for certain events its neural underpinnings have been localized—mostly to structures that are not considered to play important roles in the

[2]This is not to imply that the conscious experience of an emotion is not an important component. It most definitely is, and it has even been proposed by some researchers that without these experiences or *feelings*, the learning that comes about as part of emotional processing would not be possible in humans, nor in other species (Panksepp, 2005).

kind of processes that cognitive scientists would deem as cognitive processes, such as the amygdala. Thus, emotional experiences are largely the result of unconscious emotional processing in brain systems specialized in mediating the various evaluation processes as well as the physiological and behavioral responses that are characteristic of each emotion. This appears to be a general principle of consciousness–that the content of consciousness is greatly determined by processes that are themselves not accessible to consciousness (Kihlstrom, 1987; Johnson-Laird, 1988). In this respect, LeDoux (1994) proposes:

> [...]*unconscious processing is the rule and conscious processing is what needs to be proven. If this view does nothing else, it shows how it is possible to study emotion similarly in organisms that have defensible consciousness (humans) and those for which consciousness cannot be proven (human infants and nonhuman animals)*(p. 292).

Aware of its controversial nature, we extend this notion to advocate that it is possible to study emotional processing from a mechanistic perspective, and thus we believe it is also possible to provide our robots with mechanisms that serve the functional purposes of emotion that have usually been reserved only to humans, without entering in unproductive (for now) discussions on the conscious nature of this processing.

2.1.3 The Structure of Emotion

What are the basic elements of emotional phenomena? Can we reduce emotions into something else? If so, how should we call these elements? In the search for irreducible elements of emotion, the most prominent accounts correspond to those that view the range of emotional phenomena as a set of discrete emotions, and those which take on the perspective that emotions can be further reduced into specific dimensions, such as valence (pleasantness) and arousal (activation). We will briefly review these accounts

in the sections that follow.

Dimensional Accounts

Many researchers argue that any given emotion, such as anger, is reducible to a small number of underlying dimensions. Earlier accounts proposed bivalent views of a single dimension, such as pleasure. Nowadays, most researchers who follow this approach take a two-dimensional view of emotion in which the two dimensions correspond to the degree of pleasantness and activation to which an emotion is experienced. Thus, an emotion is considered to be either *pleasant* or *unpleasant* (also labeled as *positive* or *negative*), and also experienced as *activated* or *deactivated* (also labeled and *aroused* or *sleepy*, and *engaged* or *disengaged*) (Russell, 1980; Reisenzein & Hofmann, 1990; Reisenzein, 1994; Barrett & Russell, 1999).

One of the main proponents of this view is Russell (1980) who argued for two independent bipolar dimensions that define the structure of all emotions. In his model, which results in the circumplex illustrated in figure 2-1, all emotions can thus be described as a combination of pleasure and activation. For example, an emotion such as sadness would be considered as highly unpleasant and moderately deactivated, whereas anger might be characterized as being moderately unpleasant, but highly activated.

It has been objected that a circumplex is only observed when many states that are affective, but not necessarily considered to be emotions are used in the analysis (e.g., "tired" and "relaxed" in figure 2-1) (Scherer, 1984). Another important criticism on these accounts is that most of the evidence supporting them is based upon self-reports and introspection. As we have suggested before, analyses based on self-report are not always reliable, as the individuals reflecting on their experiences might alter them as part of the process and thus generate differences in their reports. Furthermore, it is possible that what it is being modeled is not necessarily the structure of emotion

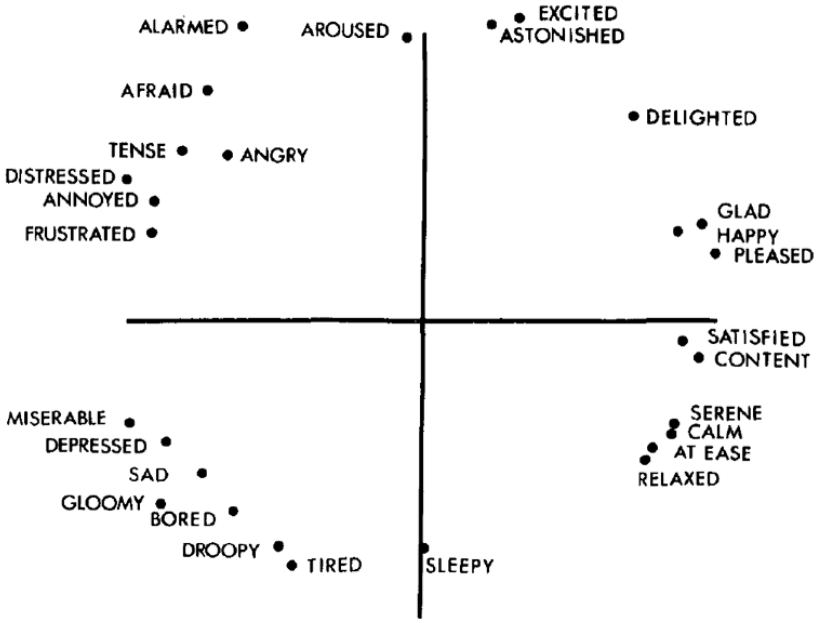

Figure 2-1: A multidimensional scaling solution for 28 affect words resulting in a circumplex. This figure illustrates Rusell's circumplex model based on two dimensions, valence and arousal, and the scaling that results after affective terms are modeled. Reprinted with permission from Russell, J. A., (1980). A Circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1161-1178.

per se, but rather the structure of evaluations, linguistic processes, or something even more basic than emotions, such as "core affect", which has been defined as "*a neurophysiological state that is consciously accessible as a simple, nonreflective feeling that is an integral blend of hedonic (pleasure–displeasure) and arousal (sleepy–activated) values*" — Russell (2003, p. 147).

In addition, given the level of subjectivity involved in the creation of such a model, the information provided by these dimensional models is rather vague at times. Whereas an emotion such as anger might be considered to be highly unpleasant by some (and reported as such), others report this level of activation to be quite pleasant. Where in this two-dimensional space should anger be localized then? Also, when thinking of the structure of emotion with the purpose of modeling it computationally,

which is the main objective of this thesis, one has to think at a much lower level of abstraction, which implies that at some point it is necessary to be able to relate these emotions and their localization in this two-dimensional space in terms of some measurable feature of the body, or reveal something about the kind of behavior that these emotional experiences result in, such as approach or avoidance. Given that a relationship is presupposed between pleasurable affect and approach, as well as between negative affect and avoidance, what should we think of emotions such as anger or fear? In these accounts these emotions are determined to have a negative valence (be unpleasant), and thus, presumably, they would instigate avoidance behavior. Such is the case of fear, in most cases, but not necessarily that of anger (e.g., approaching a conspecific in an attacking behavior). Given such distinctions, such characterization is problematic. How to explain the fact that we sometimes seek these emotions? For instance, when we see terror movies, or ride in roller coasters? The explanatory power of these accounts, at least with respect to our specific goal, is diminished. The reductionist view of all emotion into these dimensions results far too impoverished to explain the host of available data. An account such as this is coherent, however, with approaches that are primarily concerned with the expression of emotional states, such as Breazeal (2000), in which a computational model of emotional expression based upon these views was compellingly shown.

The Basic Emotions Account

It seems clear to many that there exist at least some range of paradigmatic emotions which would include instances that hold some correspondence to the English terms "fear", "anger", "surprise", "joy", "sadness", and "disgust". This is the set of so-called *basic emotions*[3], about which we will have more to say later on. For now,

[3]The specific list of elements varies somewhat, depending on the proponent. We will expand this set in Chapter 6

let us just mention that there are three main ways in which the notion of basic emotions has been used in the literature (Ekman, 1999; Ortony & Turner, 1990). The first use suggests that there are a number of *discrete* emotions that differ one from another in important ways. For instance, fear, anger, and joy differ in their eliciting conditions, as well as in their usually associated behavioral and physiological characteristics. This perspective posits a marked contrast to other views (described below) that treat emotions as being essentially the same constructs, which differ only in terms of activation or pleasantness.

The second meaning argues that these emotions are fundamental and sufficient elements to describe all emotional phenomena. The term *basic* applies to them in the sense that they constitute the building blocks for all emotions. By themselves, these emotions are descriptive of the most common emotional phenomena, and when combined, they can produce other more complex emotions—such as "guilt" or "shame"—much in the same way basic colors would be combined to produce composite colors (Izard, 1977; Plutchik, 1994). For instance, according to Plutchik (1994), one of the better known proponents of this view, fear and surprise would generate "awe" when combined, whereas joy and fear would result in "guilt". Figure 2-2 illustrates Plutchik's model of emotion and emotion blends. His three-dimensional circumplex model of emotions describes relations among the different emotions in a manner analogous to the way colors are represented in a color wheel. Folding the model into a cone, its vertical dimension represents the intensity of the emotion, whereas the circles represent degrees of similarity among the different emotions. According to this model, there are eight primary emotion dimensions defined as four pairs of opposing emotions. The emotions in the blank spaces correspond to blends of two of the primary emotions (Plutchik, 2001).

Finally, the third meaning of the basic emotions suggests that these discrete emotions have a biological basis, and as such they are *basic* for the organism. This view

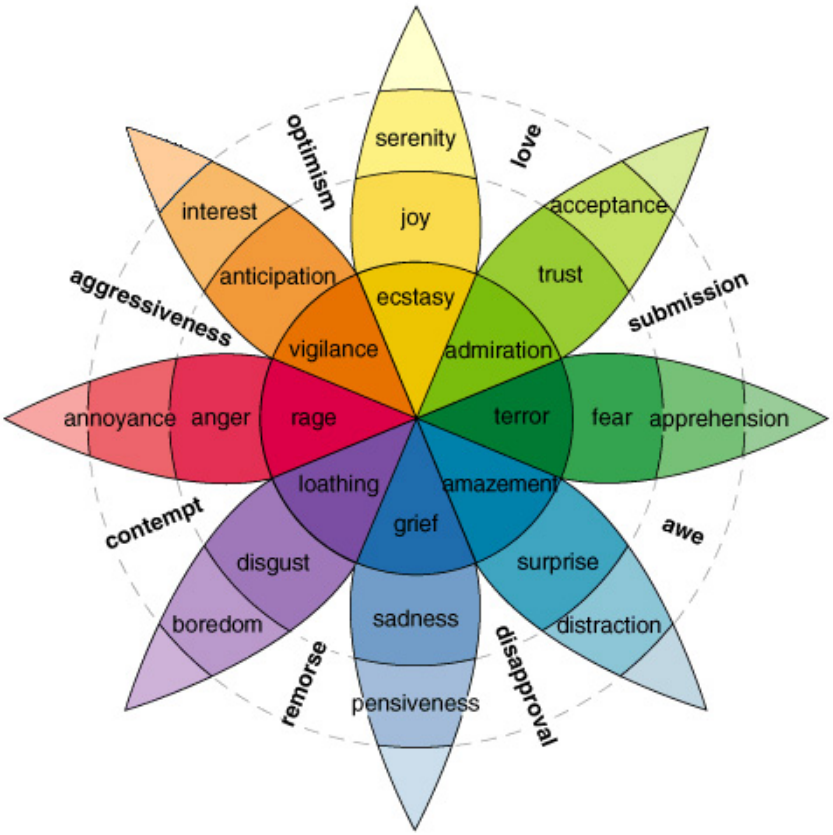

Figure 2-2: Plutchik's circumplex model of emotions and emotion blends. This three-dimensional circumplex model describes relations among emotions in a manner analogous to the way colors are represented in a color wheel. Folding the model into a cone, its vertical dimension would represent the intensity of the emotion, whereas the circles represent degrees of similarity among the emotions. According to this model, there are eight primary emotion dimensions. The emotions in the blank spaces correspond to blends of two of the primary emotions. Adapted from Plutchik (2001).

suggests that these emotions evolved due to their adaptive value in helping organisms deal with recurrent, fundamental life- and survival related tasks. Thus, the characteristics shared by these emotions are largely biologically determined (Tomkins, 1962; Tomkins, 1963; Izard, 1977; Johnson-Laird & Oatley, 1992; Ekman, 1992; Ekman, 1994*a*; Cosmides & Tooby, 2000). It is this meaning which has been more widely

used, but it is not without controversy.

Table 2.1: A Selection of Lists of "Basic" Emotions Based on Ortony & Turner (1990).

Reference	**Fundamental emotion**	**Basis for inclusion**
Arnold (1960)	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness	Relation to action tendencies
Ekman, Friesen & Ellsworth (1982)	Anger, disgust, fear, joy, sadness, surprise	Universal facial expressions
Frijda (1986)	Desire, happiness, interest, surprise, wonder, sorrow	Forms of action readiness
Gray (1982)	Rage and terror, anxiety, joy	Hardwired
Izard (1971)	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise	Hardwired
James (1884)	Fear, grief, love, rage	Bodily involvement
Mowrer (1960)	Pain, pleasure	Unlearned emotional states
Oatley & Johnson-Laird (1987)	Anger, disgust, anxiety, happiness, sadness	Do not require propositional content
Panksepp (1982)	Expectancy, fear, rage, panic	Hardwired
Plutchik (1994)	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	Relation to adaptive biological processes
Tomkins (1984)	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise	Density of neural firing
Watson (1930)	Fear, love, rage	Hardwired
Weiner & Graham (1984)	Happiness, sadness	Attribution independent

Not surprisingly, the specific list of emotions that are considered to be basic varies from researcher to researcher. Table 2.1 describe some of the lists proposed by theorists. Regardless of these differences, the important aspect of this definition is the reference to a biological determinism and a functional account, as Oatley & Johnson-

Laird (1996) declared:

> *The precise number of basic emotions is less important than the hypothesis that each kind of emotion has specific functions and that mechanisms that evolved to serve these functions map diverse events into a small set of emotional modes.* — Oatley & Johnson-Laird (1996, p. 365)

The account for basic emotions is appealing to most researchers, even if the third meaning of its usage (i.e., that they have a biological origin) is not appealing to others (Ortony & Turner, 1990; Turner & Ortony, 1992). As it will be described in the following sections, we follow a similar view of basic emotion programs in this work.

2.1.4 Nature Versus Nurture

Are affective phenomena completely due to the design, the constraints, and information processing styles of affective mechanisms provided by nature to organisms? To what extent is affective processing the result of individual learning about the different conventions, models, and roles provided by our cultural and societal environments?

Given the amount of compelling evidence for a neurobiological substrate (Panksepp, 1998), some of which is described in more detail in Chapter 4, it is difficult to dispute the idea that emotions have a biological basis. However, the exact nature of what these mechanisms remains an open question.

In any case, it should be clear that the capacity for emotion is deeply ingrained in many species. Furthermore, the processes of emotional evaluation, by which significant stimuli come to elicit emotions, are also thought to rest upon innate capabilities (LeDoux, 1996). There is also strong evidence that there exist innate dispositions related to specific emotions, or at least to forms of stereotypical behavioral tendencies given specific and recurrent situations such as facing an immediate danger, fighting for

resources or a mate, experiencing an irrevocable loss, working toward a goal, falling in love, escaping predators, and so on. Some of this evidence comes from neuropsychological findings (Panksepp, 1998), as well as the findings of universality of such action patterns (Cosmides & Tooby, 2000), including but not limited to facial expressions (Ekman, 1994*c*).

Biological and evolutionary explanations can go terribly wrong when they ignore development and cultural determinants, and we know that developmental and evolutionary explanations are essentially complementary. Still, it is not clear how biological dispositions interact with culturally determined traits. It is also unclear how emotions that have an important cognitive component, such as regret, relate to biological mechanisms and dispositions based on the basic emotions described abvove. How does a mechanism such as learning make use of what is being provided in the environment? Does what is given constrain what is learned? These questions have hardly been treated with any depth, in part because it is so difficult to do so from a traditional experimental perspective. A computational approach to emotions may provide the basis for such an endeavor. Starting from Chapter 6, and in those that follow, this thesis lays an initial foundation from which some of these questions can be explored.

2.2 The Quest for the Affect Abstraction: Looking at Emotion Theories

Providing a complete account of the major theories of emotion is beyond the scope of this work. We will however provide a general overview of each theory that will be sufficient for us to lay a basic conceptual framework useful for understanding the reasons behind some of the main design decisions considered in our computational

approach to emotion. For a much more comprehensive overview of the different theories of emotion, the reader is referred to Cornelius (1996).

Although no particular agreement exists upon a general strategy to investigate affect, most of these different theories do seem to fit into any of three main schools of thought concerning the study and categorization of emotion: *Cognitive appraisal theories*, *social constructionist theories*, and *evolutionary theories*. Depending on which theoretical approach is adopted, the specific methods to study emotion, the various levels of detail applied, the class of phenomena studied, the way in which emotions are though to be elicited, and the particular biases, goals and application of the findings of the research, the results will most certainly vary.

2.2.1 Cognitive Appraisal Theories

Research on the psychology of emotion has been dominated by cognitive perspectives, of which, appraisal theory is perhaps the most representative approach (Arnold, 1969; Ortony, Clore & Collins, 1988; Roseman et al., 1990; Lazarus, 1991; Johnson-Laird & Oatley, 1992).

The main idea behind cognitive appraisal theories is that emotions are generated as people evaluate or appraise the many different events that occur in their environments. This appraisal process requires some form of evaluative cognition, which depend on a person's attitudes, attributions, beliefs and desires, and which ultimately determines what emotion, if any, should be experienced on any given situation. Therefore, this means that different individuals might actually experience different emotions given the same event or stimulus. Consider for instance the event in which a vacation you had planned several months in advance to the sunny Caribbean is thwarted a couple of days before traveling due to a category F5 hurricane that hits the islands where your hotel was located. In such an event, some people might experience sadness for

not being able to travel, while others might experience anger, and still others might experience fear of the imagined conditions, even if the events did not touch upon them directly. This variability in the experience of emotion, which according to these views depends on the cognitive evaluations made by the individuals, was one of the main motivations for the development of these theories and became an argument against biologically relevant events and hence biological theories.

One of the main implications of cognitive appraisal theories is that emotions are not necessarily produced from hardwired constructs in response to certain, biologically relevant stimuli, but that the emotional significance of the events and objects depends on the goals and the coping strategies of each individual in any given situation.

In general, the main assumption of these theories is that it individuals, whether consciously or not, can assess the degree to which events are positive or negative, whether these events are in accordance to the individual's expectations, whether they obstruct or facilitate goals that the individual may have, whether they are under the individual's control or not, whether the events are familiar or novel, and whether the responses will be manageable or on the contrary overwhelming. All of these evaluations are then organized into taxonomies that reflect the different patterns of evaluation and thus the discrete emotion that the individual would experience.

For instance, Figure 2-3 illustrates the taxonomy proposed by Roseman (1984). In it, an emotion such as fear would be evoked after a cognitive appraisal has been made by the individual reflecting a pattern in which the event occurs under circumstances that are beyond the control of the individual, it would implicate the absence of a reward or the presence of a punisher, and the individual expected that the event is undeserved.

Given the amenable implementation of these taxonomies into production rules of the IF-THEN kind, it is not surprising that cognitive appraisal theories have become the favorite picks of most computational approaches proposed to date. However, an

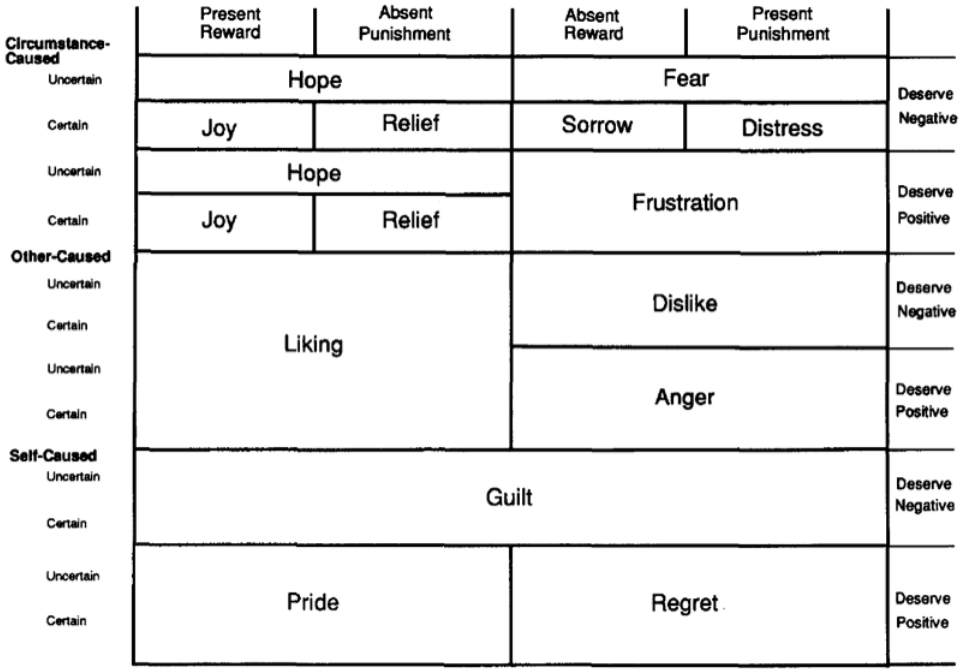

Figure 2-3: Appraisal Patterns According to Roseman. This figure illustrates Roseman's hypothesized appraisal constructs that elicit emotion. Reprinted with permission from Roseman, I. J., Spindel, M. S., and Jose, P.E. (1990). Appraisal of emotion-eliciting events: Testing a theory of discrete emotions. *Journal of Personality and Social Psychology*, 59, 899-915.

important assumption made by supporters of this perspective is that emotions can be explained by studying and understanding how people come to these evaluations about their world. To this end, researchers rely on self-report and introspective techniques as the main tools to develop these cognitive accounts and identify the relationships between different appraisal configurations and the category of emotion they elicit.

Considering the main purpose of this research, we would argue that symbolic representations of the qualitative observations of people's emotional experiences are useful to study *how people interpret and reason* about those experiences, but they do not necessarily provide much more information with respect to how the emotions are

actually processed and elicited in the first place, and much less about the underlying neural mechanisms.

In addition, and from a methodological perspective, it has been shown, both through experimental and clinical studies, that we often do not properly understand and verbalize the causes of our own behavior (Nisbett & Wilson, 1977; Gazzaniga & LeDoux, 1978). Thus, one must remain cautious of approaches that rely on introspection to provide information about causal knowledge (see Farthing (1992) for review). The main concern in this case is of *access* to the proper kind of information. These approaches assume that introspection on emotional experiences has access to the same kind of information used by the brain in producing those emotional experiences in the first place, which, again, has been shown this is not always the case (Gazzaniga, 1998).

2.2.2 Social Constructionism

The apparent differences in emotions among various cultures is one of the main motivations of the social constructivist perspective. Embedded within a broader social constructionism research program, this approach to emotion aims at understanding how different aspects of culture including social practices, moral structures, and language, influence and ultimately determine the different manifestations of emotional phenomena.

The analysis of cultural differences in language for emotion is one of the main influencers for the social constructionist approach (Lutz, 1988). As a general rule, this approach rejects the notion that there are biological truths in the sense described in section 2.1.4 and further on. To the contrary, it suggests that most human processes, conditions, and states are *constructions of society* that serve in the end certain goals of the same society it created them. Thus, from this perspective, emotions are products

of any given culture that are constructed by the culture and in essence to serve the culture (Gergen, 1985)

Averill (1980), one of the main advocates of the social constructivist approach, defines emotions as being:

> [...]*a transitory social role (a socially constituted syndrome) that includes an individual's appraisal of the situation and that is interpreted as a passion rather than an action.* — Averill (1980, p. 312)

Two main aspects of this definition, shared by other proponents of the social constructionism of emotion, are the notions of *social concepts* and *social roles*. The first notion, as in the cognitive appraisal theories, suggests that emotions are the product of a person's evaluations of events in the world. The main difference, however, is that the types of judgements performed by a person are not naturally, but rather, culturally determined. In other words, culture, mores, and social structures in general, shape the contents of those beliefs, attitudes and judgements. The second notion refers to the set of rules that allow a person to define what is the "proper" way of responding to a given situation.

According to this view, social concepts and social roles are not simply influenced by culture, they *are* the product of culture, thus explaining the wide variety of emotional phenomena and the characteristic patterns of behavior found in different societies (Averill, 1984; Armon-Jones, 1986). As a somewhat radical perspective, the social constructivist view has been intensely debated, especially by other researchers working on more biological and evolutionary approaches to emotion, which is the topic of the following section.

2.2.3 Affect Programs Theory

The affect program theory of emotion puts forward the idea that some emotions are pancultural "programs" enabled by biological capabilities acquired throughout our evolutionary past. The term "program" indicates that they are coordinated collections of complex biological responses that occur together in response to prototypical and recurrent situations for which adaptive solutions have been found.

Like other psychoevolutionary approaches, the affect program theory has its origin in Darwin's *The Expression of the Emotions in Man and Animals* (Darwin, [1859]/1998). In this seminal work, Darwin studied and reported on the various facial and bodily expressions concerning distinct emotions. Moreover, he suggested what their possible functions could be, and alluded to homologies to responses in other species.

Darwin's approach to emotion influenced most modern work on emotional expression (Izard, 1971; Ekman & Friesen, 1986). These recent theories, together with objective analysis of subjective human experiences, and recent evidence of specific brain systems (LeDoux, 1996; Panksepp, 1998), have led researchers to believe in the existence of a small set of discrete primary[4] emotional systems, or *affect programs*.

Affect programs can be defined as executive, operating systems that generate and coordinate short-term, stereotypical responses that allow organisms to deal with biologically significant events in ways that promote survival.

These responses involve a variety of elements such as facial and behavioral expressions, arousal of the Autonomic Nervous System (ANS), vocal expressions, modulation of attention, and affective feelings. It is important to note that no one of these elements constitutes the essence of an emotion. In particular, and contrary to other views, emotion feelings are no more central to the identity of a particular emotion

[4]Primary or *basic*, in the sense of being fundamental and biologically determined, not in the sense that it can be combined with other emotions to form secondary ones.

than is its characteristic facial expression.

Different theorists have various reasons to classify affect programs as *basic*. These include different arguments such as the fact that some of these response patterns are pan-cultural and homologies can be found in related species (Ekman, 1992; Izard, 1994), and ample neuroscientific evidence suggesting the existence of several intrinsic emotional systems in the mammalian brain, including those mediating fear, anger, separation distress, interest, play, sexual lust, and disgust (Flynn, 1967; LeDoux, 1996; Panksepp, 1998).

In preliminary work, described in further chapters, we have found the affect program paradigm to be both a compelling argument on the nature of affect, and a useful abstraction to implement and examine affective processing from a computational perspective.

2.2.4 Revisiting the Definition of Emotion

In the previous sections we defined affect as the set of processes that incite action, and thus are motivational. At the same time, we reviewed some of the basic issues involved in the analysis of emotion, including the ideas of discrete emotions versus dimensional accounts of emotion, states or processes perspectives, and biological determinism versus cultural development.

Building upon these ideas, we can further characterize our view on emotion, basing our definition on the affect program theory and in that of Panksepp (1998), who is one of the main proponents of this view, and a pioneer in the recent field of *Affective Neuroscience*. From this field's perspective, we define emotions (in the affect programs sense) as systems that:

1. Are biologically predetermined and designed to respond unconditionally to stimuli arising from major life-challenging circumstances.

2. Organize diverse behaviors by activating or inhibiting, and synchronizing sensorimotor subprograms and concurrent autonomic-hormonal changes that have proved adaptive during our ancestral past

3. Change the sensitivities of sensory systems and modulate attention in a manner relevant to the behavioral sequences that have been activated

4. The activity arising from these programs outlasts the circumstances that elicited them in the first place

5. Can come under the conditional control of emotionally neutral environmental stimuli through local learning systems

6. Have reciprocal interactions with the brain mechanisms that elaborate higher decision-making processes and consciousness.

7. These emotion circuits, constitute local learning systems that bias the organism to pay attention to and learn about the environment whenever the eliciting circumstances have activated the circuits. In other words, these circuits, when active, bias the organism for learning, indicating *what* to learn and *when* to learn it.

In the following chapters, we describe how, through a computational approach, we take the explanatory power of the affect programs theory and convert it into a computational abstraction of what an emotion is and how it is implemented.

2.3 Summary

This chapter introduced some of the most commonly debated issues related to the study of emotion, including the contentious definition of this construct, its structure and the different levels of abstraction at which it can be studied.

We briefly reviewed the main approaches proposed by different theorists with regards to the study of emotion, and focused on a psychoevolutionary approach that provides the most compelling theory of emotion, from our perspective, while at the same time addresses many of the issues involved in the characterization of the phenomena and provides an appropriate level of abstraction that is suitable for computational modeling.

Finally, we introduced the notion of *affect programs*, as the primary theoretical constructs for investigating the function and the mechanisms of emotion, and which correspond to biologically predetermined operating systems that generate and coordinate short-term, stereotypical responses that allow organisms to deal with biologically significant events in ways that promote survival.

Chapter 3

The Function of Emotion

In the previous chapter, we alluded to the notion that all organisms possess a set of discrete emotions that serve specific purposes. What are these purposes? Why do we have emotions? In this chapter we address these questions from a functional and operational perspectives. At the functional level, and following the affect program theory, we propose that emotions (affect programs) exist due to the need that all organisms have to deal with fundamental life–and survival–related tasks, such as avoiding danger, finding mates, securing shelter, seeking resources and solutions to satisfy physiological needs, guarding these resources, and so forth. Of course, dealing with complex tasks such as these, and doing so in dynamic and uncertain environments, must involve the activity of a variety of mechanisms that operate at various levels. We suggest that the coordination and synchronization of these mechanisms is, at an operational level, the main reason for emotion.

In the sections that follow, we will describe some of the mechanisms that are coordinated, modulated and synchronized by the set of affect programs described before. These interactions will set up the stage for the computational models that will be described in further chapters.

3.1 Emotion and Cognition

Since early times, emotion has always been separated from cognition. When William James wrote and published his seminal paper on 'What is an emotion?' (James, 1884), his descriptions were certainly such that this separation between cognitive processes and emotions seemed more than reasonable. From a Jamesian account, we are afraid because we escape from danger, or sad because we cry. Our perception of all bodily changes involved in the processing of affect, according to James, determine how we feel, and thus the emotion that we experience. James and Lange's ideas led theorists to set out with the task of understanding the set of bodily changes and processes associated to emotion. One consequence of such line of thought was that emotion was kept a separate entity from mind processes, as they referred only to cognition. Emotion was thus reduced to visceral sensations, and further separated from thought and rational processes. This was further exacerbated with the dawn of behaviorism, which almost completely removed the topic of emotion from the studies of the mind. Whereas cognitivism implicated the thoughts and beliefs of an organism as determinants of its actions, behaviorists dealt with objective and observable measures only, which had no place for emotions and other psychological constructs that were "unmeasurable" from an objective standpoint.

Much time has passed since then and many advances have elucidated new avenues of thought in this respect. In particular, recent work in the neural substrates of emotion have allowed insight into otherwise "obscure" mind processes and as a result it has increased the interest in undertaking the study of emotion, which directly or indirectly, involve examining the relation between emotion and cognition.

The work by Damasio (1994), for instance, has been paramount in raising interest in understanding emotion, and affective processing in general, from a mind perspective. His work argues that contrary to popular belief, emotion and reason go together,

hand in hand, in the processes that generate coherent and intelligent behavior. Based on studies with patients that had lesions in the frontal lobes, particularly in the ventromedial and orbitofrontal cortices, he suggested that emotion was more than just a disruptor of rational thinking, and that in fact it was a necessary influence for rational decision-making and other processes we tended to associate only to higher-level cognition. The work of Damasio, together with the pioneering work by (LeDoux, 1996) in elucidating the specific mechanisms for fear conditioning, have certainly brought emotion and cognition to the forefront, and most importantly, back together.

Although reviewing the whole spectrum of interactions between affect and cognition is beyond the scope of this work, it certainly is important to recognize the fact that affective and cognitive processes are intertwined in many interesting and unknown ways. Some of these interactions include the emotional modulation of memory and learning, the relationships between emotions and the planning and execution of goal-directe behaviors, the influence of evaluative processes in early information processing, the influences and modulation of emotion in perceptual fluency, and the relationship between emotions and the darlings of cognitivism: beliefs and goals. Interactions between emotion and cognition are aplenty! The following sections focus on just two of these interactions, as they are an integral component of this work: affect and motivations and goals, and the interactions between affect and attention.

3.2 Emotion and Motivation

Perhaps one of the main interactions that need to be discussed is that between emotions and incentives. Incentives or motivations refer to those events and objects in the world that elicit or have a tendency to elicit actions, as they are of biological significance to organisms.

The relations between emotion and motivation constitute a great area for debate.

As it is the case for the term "emotion", "motivation" suffers from a similar polysemy. As Minsky (2006) suggests, these are "suitcase" words that can be used to refer to many different situations and phenomena. One can view motivation as a cause of emotions, as one of its major aspects, and as one of its consequences. Here, we will attempt to make distinctions between these terms in order to avoid confusion. In principle, we will argue that a specific kind of affect program, the *Seeking* system, is the main emotional system that deals with the incentives of the world.

3.2.1 A Brief History of the Study of Motivation

The concepts related to incentive motivation came about as the notions related to drives, and drive reduction theories (Hull, 1943), fell from grace in the 1960s. Multiple advances in the understanding of brain function led theorists to reject the views that posited how drive and drive reduction theories explained the way incentives came to promote behavioral responses and learning. Based upon the arguments that seeking and learning about rewards could not operate under these principles, new theories were developed, and some of these theories incorporated the notion of incentive motivation (Bolles, 1972; Bindra, 1978; Toates, 1986).

In particular, Bolles (1972) proposed an account for these phenomena based on incentive expectancies. He suggested that organisms were motivated, and learned about rewards, not by any reinforcement produced by the reduction of a particular drive (e.g., hunger or thirst), but rather because they had learned to expect the occurrence of particular hedonic rewards. In other words, he argued that organisms learned associations between a neutral stimulus (S), such as a light or sound, and a hedonic reward that followed (S*), such as a palatable food. He called them S-S* associations to suggest that the S would elicit an expectation for the S*, which was of motivational significance before the association was formed (Bolles, 1972). He also

suggested that organisms learned R-S* associations, namely that their own response (R) became useful predictors for the occurrence of the S*.

This expectancy relationship was questioned by others with respect to its motivational value. That is, why would an expectancy association cause motivation? To overcome these shortcomings, Bindra (1978) extended Bolles' account, but otherwise rejected his notion that it was this expectancy alone the main cause for incentive motivation. Bindra suggested that a stimulus that predicts a reward (the S in Bolles' terms) does not simply create an expectation, but rather it also elicits a motivational state of the hedonic reward itself (the S* in Bolles' terms). Thus, the learned association implies not only a cognitive expectation of the reward, but also causes the organism to perceive the stimulus (S) as if it were the hedonic reward (S*). In other words, the S takes on, or is attributed, specific affective and motivational properties that otherwise would normally belong to the S* itself.

Bindra's claim that neutral stimuli become the same as incentive stimuli as a function of their association did not go without criticism, however. Critics questioned that if neutral stimuli simply became incentives because of learning, then they would always elicit responses, much as incentives do, regardless of the physiological states that an organism would be in (Gallistel, 1978). Physiological mechanisms do regulate the motivation for incentive rewards, and one normally seeks out food when hungry, water when thirsty, or a mate when sexually aroused. Thus, some link was missing which would connect the ideas of physiological states to those of incentives.

Incorporating these concepts further into incentive motivation, along came Toates (1986), who extended the Bolles-Bindra accounts and incorporated the notion that physiological drives and regulatory mechanisms could mediate the affective and motivational value of stimuli. One of his most interesting propositions was that physiological or drive mechanisms did not need to drive motivation directly, but they could have multiplicative effects on the hedonic value and the incentive value of rewards.

Likewise, by the same mechanisms suggested by Bolles and Bindra, they could also enhance the affective and incentive value of the stimuli that predicted these rewards. Thus, an interesting set of interactions was now accounted for, which would consider the effects of physiological mechanisms, the affective and motivational properties of incentives, and the learned associations between predictive stimuli and rewards.

An important addition to these concepts was Toates' position with respect to cognitive expectancies. He suggested that Bolles' cognitive expectancies and Bindra's basic incentive motivation processes might be active simultaneously and would act in different ways to control the behavior of an organism via two different pathways: a stimulus-bound one and a goal-directed one (Toates, 1986).

These ideas were later taken up most notably by Berridge and colleagues, and by Dickinson and Balleine, among others, and developed further into what is currently known as *Incentive Salience* and *Incentive Learning* theories which constitute the foundation for the motivational accounts included in this thesis (Berridge & Robinson, 1998; Berridge, 2000; Dickinson & Balleine, 2002).

3.3 Emotion and Rewards

To many, the term reward is often used as a substitute for conscious pleasure. That is, something that is consciously liked in the hedonic sense. However, is conscious pleasure really necessary for an incentive to be considered as rewarding? While this might be an obviously positive answer for some, it turns out to be incorrect. Despite the many accounts that argue for the impossibility of affect without awareness (Clore, 1994), there is evidence to the contrary suggesting that unconscious affect is possible, as shown by the experiments of Winkielman & Berridge (2004). What does this mean for those theories that account for reward only with respect to their hedonic or affective value?

Reward has been treated as a unitary concept in psychological terms, however recent evidence suggests that this notion in fact does not correspond in reality to a monolithic process, nor to a unique underlying brain mechanism. Recent studies have shown dissociations of multiple reward components, as well as several different neural circuits that may very well code for and process reward-related information(Berridge & Robinson, 2003; Schultz, 1998).

As it will be described shortly, the notion of reward corresponds to multiple psychological components (see Figure 3-1), and not only to the affective component that reigned in psychological traditions. Hence, when we discuss about reward it is important to explicitly refer to which of these components we are referring to in order to avoid unnecessary confusion.

A first component clearly identified by the experiments of Schultz and colleagues (Schultz & Romo, 1990; Schultz, Dayan & Montague, 1997; Schultz, 1998; Schultz, 2002; Schultz, 2006), relates to learning. That is, considering rewarding events, how can we learn about relationships among stimuli and about the consequences and outcomes of our actions? A second component is of an affective nature: obtaining and consuming rewards can produce affective reactions in the hedonic sense. Finally, how do we choose a course of action? What events and objects in the world are significant for us to pursue and how do we obtain the impulse or motivation to act upon them and learn from them? In other words, a third component of reward relates to its motivational aspects. Thus, at the very least, there are three interrelated components of reward that must be considered when studying and discussing about reward. Although the interactions between these components is still an open question, as it is the precise identification of how brain circuits mediate these constructs, it is clear that we should be more precise when referring to reward and reward-related processes.

It is important to note that the affective and motivational components of reward,

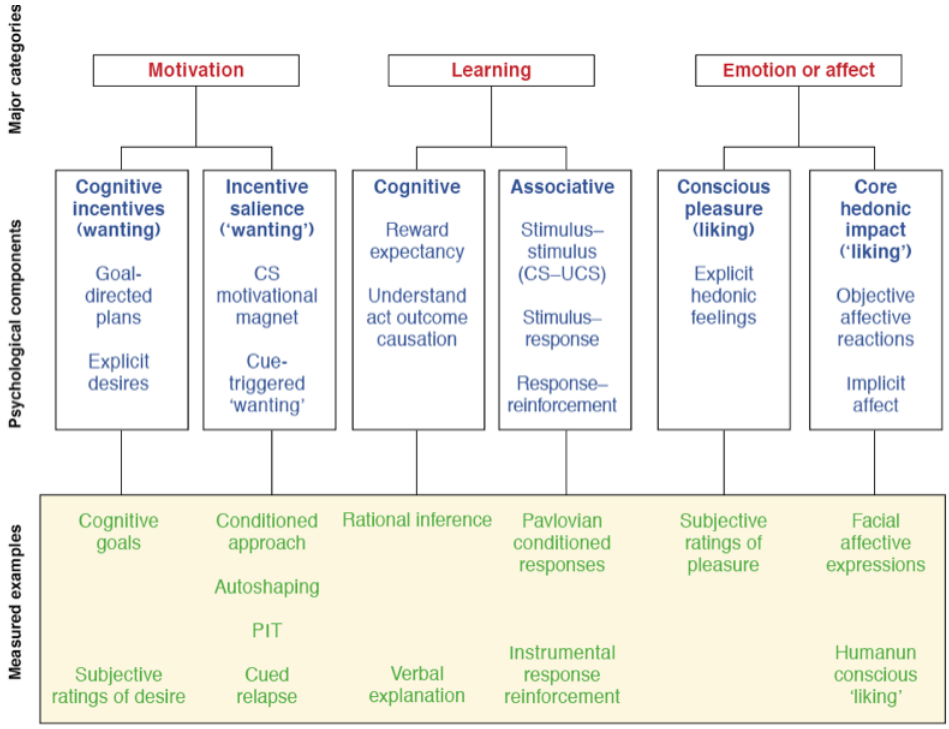

Figure 3-1: Multiple Components of Reward. This figure illustrates the different components of reward: learning, motivation and affect. Each of these components can be consciously experienced, but implicit accounts for each of these components do exist as well. Reprinted with permission from Berridge, K. C. and Robinson, T. E. (2003), Parsing reward, *Trends in neurosciences*, 26(9), 507513.

as described in Figure 3-1, can exist even without conscious awareness. These implicit processes are interesting to the study of reward because if core reward processes can be dissociated from their subjective feeling, they might be better suited to objective measurement in brain manipulation experiments (Berridge & Robinson, 2003).

Let us consider each of these implicit components separately. First, recall the Bolles-Bindra-Toates multi-theory reviewed above. This theory suggests that learned incentive stimuli have both affective and motivational consequences. That is, they can be both 'liked' and at the same time 'wanted'. The use of quotes surrounding these terms was proposed by Berridge and colleagues, to refer to the objective measures

of each construct. Hence, when referring to conscious pleasure as the subjective affective feeling the term is used without quotes (liking), but an objective measure of an affective reaction would thus correspond to what is referred to as 'liking'. Similarly with respect to 'wanting'. When speaking of the conscious and subjective notion of desire, the term is used without quotes (wanting). However, the objective motivation process, which we will describe later on as *incentive salience*, would be referred to as 'wanting'.

Liking and wanting are fairly standard terms that are sometimes used interchangeably. While some causality might be implied, this is not necessarily so, but most people would think that incentives (rewards) whenever they are liked they are also wanted. Berridge (2000) and colleagues have taken these ideas further and suggested that these concepts actually correspond to two very different incentive processes as it will be described in the following section.

3.3.1 'Wanting' versus 'Liking'

Incentive salience or 'wanting' follows the Bindra-Toates rules for incentive learning but identifies separable brain substrates for 'liking a reward versus 'wanting the same reward. The incentive salience model was proposed by Berridge & Robinson (1998) as a way to reconcile the different and competing theories about brain dopamine function (which will be described in more detail in Chapter 4). In particular, the model intended to reconcile the fact that dopamine sometimes seemed to mediate sensory pleasure, like Wise and others suggested (Wise, 1996; Wise & Rompre, 1989; Wise, 2002), when in fact this might not be the real case.

'Liking' is essentially hedonic impact or affective value of a reward—the brain reaction underlying sensory-pleasure triggered by immediate receipt of reward such as a sweet taste, as measured in objective terms. While these reactions are usually

triggered by so-called unconditioned stimuli, 'liking' can also be produced by predictive, conditioned stimuli, following the same principles suggested by Bindra, and described above (Bindra, 1978). On the other hand, 'wanting' or *incentive salience*, corresponds to the motivational incentive value, rather than the affective value, of the same reward (Berridge & Robinson, 1998).

Why did brains evolve separate 'wanting' and 'liking' mechanisms for the same reward? As Berridge & Robinson (2003) suggest, perhaps 'wanting' evolved after more primitive 'liking' so as to provide a common currency shared by all rewards which could be useful in order to compare and decide between the certainly different 'liking' choices for food, sex or other incentives. Through some clever and elegant brain manipulations, Berridge and colleagues have shown dissociations of these two incentive processes, which as we mentioned before, normally function and go together, but which can be split apart (Wyvell & Berridge, 2000; Berridge & Robinson, 1998; Berridge & Robinson, 2003).

'Liking' can be produced without eliciting 'wanting', and vice versa, 'wanting' without 'liking'. 'Liking' without 'wanting' occurs when brain manipulations are performed in such a way as to suppress the neurotransmission of mesolimbic dopamine. For instance, disruption of mesolimbic dopamine systems, through neurochemical lesions of the dopamine pathway that projects to the Nucleus Accumbens (NAS) or by receptor-blocking drugs, dramatically reduces incentive salience or 'wanting to eat a palatable reward, but does not reduce affective expressions, including orofacial expressions such as tongue protrusions, of 'liking' for the same reward (Berridge & Robinson, 1998; Berridge, 2004). Such disruption of dopamine systems leaves the individuals nearly without any motivation for any incentive at all, whether conditioned or not, and be it food, water, sex, and even drugs (Brauer, Goudie & de Wit, 1997; Smith, 1995; Berridge, 2006). Strikingly, however, 'liking' of these same incentives, remains intact, as it can be assessed by either affective facial expressions or

subjective ratings in the case of humans (Peciña, Berridge & Parker, 1997; Berridge & Robinson, 1998).

Conversely, 'wanting' without 'liking' can be produced by several brain manipulations, including electrical stimulation of the lateral hypothalamus (Berridge & Valenstein, 1991; Panksepp, 1998), which makes rats 'want' more of any incentive that is presented to them. Similarly, 'wanting' without 'liking' can be triggered by microinjections of amphetamine that activate dopamine neurons in the NAS. The presentation of a Pavlovian CS for food causes rats to work even harder than normal in order to obtain food. The enhanced 'wanting' obtained by the microinjections of amphetamine into the accumbens is not accompanied by increased 'liking', as evidenced by the microinjections failure to increase positive hedonic patterns of behavior elicited from the rats by a sweet taste (Wyvell & Berridge, 2000). Likewise, mutant mice whose brain receptors receive more dopamine than normal due to their genetic mutation also show excessive 'wanting' of sweet rewards. Interestingly enough, these same rats seemed to 'like' sweet taste less than normal mice do (Peciña et al., 2003).

many have suggested that drug addiction in humans may elicit similar processes, as drug addicts often report drug cravings even when they do not derive much pleasure from them. Consider for instance tobacco related addiction. Nicotine generally does not elicit substantial pleasure in most people, but can still be quite addictive.

So what is 'wanting' if it is not 'liking'? According to Berridge and colleagues, 'wanting' is an incentive process mediated by mesolimbic dopaminergic systems that can essentially "tag certain stimulus representations in the brain. When incentive salience is attributed to a reward stimulus representation, it makes that stimulus attractive, attention grabbing, and a target for action in the Bolles-Bindra-Toates sense described above (Berridge, 2006).

'Wanting' and 'liking' are both necessary for an incentive to be perceived as a normal reward. 'Wanting' without 'liking' would constitute a "fake" reward, devoid

of any affective pleasure. Perhaps not unlike what alexithimia produces in certain people. Notwithstanding, 'wanting' is an essential component of reward, without it, 'liking' would simply be an affective reaction. For an incentive to be experienced as a reward, both 'wanting' and 'liking' need to be present. The process of incentive salience, which forms the foundation for the model described in Chapter 8, is the one that makes a specific stimulus or an action the object of desire, and that tags a behavior as the rewarded response (Berridge, 2004).

3.4 Emotion and Attention

Affect certainly influences attention, whether at a low level, by promoting saliency to stimuli that are of affective significance, and thus reducing the attentional space to only those stimuli that are of biological interest to the organism, or at a higher level by promoting coherence in behavior, helping reduce dithering between behavioral responses when motivational conflicts arise.

In this work, we will operationalize the notion of attention to mean *the collection of processes that serve to provide coherent control of action.* Organisms regularly face a variety of stimuli, each of which may trigger different responses which may compete for the ultimate control of the organism's resources. Affective behavior, as we will describe in this work, depends on the ability to selectively use elements of these stimuli, while ignoring others. This is known as selective attention, and as we will see later on, this process proves essential in the ability to produce coherent and adaptive behavior.

Recent research suggests that neural structures, such as the amygdala, which have long been implicated in the coordination of emotional behavior and emotional information processing, are also implicated in attentional processes. First, affective processing, through the amygdala, has been implicated in the modulation of orienting

responses by associative learning; and second, the enhancement of the associability (increased likelihood that a neutral stimulus will be learned and associated to an incentive) of particular contingencies (Holland & Gallagher, 1999).

Based upon evidence relating amygdala functioning to the modulation of attentional processes, this work will propose a mechanism by which the *Surprise* and the *Seeking* affect programs, interact and mediate affective processes, essentially acquiring the ability to interrupt ongoing stimulus processing when novel or significant stimuli are detected and to habituate to those that have no affective significance, thus promoting coherent behavior.

3.5 Summary

In this chapter we have described some of the most important interactions between affect and many other constructs and mechanisms we have considered to be part of intelligent behavior, and which have mostly been ascribed to higher cognitive processes.

We argue, however, that in order for affect programs to implement many of the solutions required to promote and sustain life, its activity is necessarily intertwined with these other processes, in fact, coordinating, synchronizing and regulating them as part of the functional mode of operation each affect program promotes.

We have discussed two such interactions, namely, that of affect and motivation and affect and attention. Further chapters build on these ideas as we attempt to design and build deep models of affect.

Chapter 4

Neural Substrates of Emotion

In Chapter 1, arguments were made questioning traditional assumptions of intelligence, which have resulted in a set of methodological principles that have guided the construction of our robotic systems and, in general, our approach to understanding intelligent behavior (Brooks et al., 1998). This chapter complements these arguments from an affective perspective and provides a synopsis of the lessons learned regarding the neural substrates of emotion that all other levels of analysis in emotion research may have to consider if they are going to fully comprehend the nature of emotions.

4.1 No Monolithic Emotional System

Undoubtedly, there are appraisal mechanisms in the brain, which assess the biological and affective significance of events as processed by perceptual systems. Contrary to these views, however, the notion of a single, one-for-all system of emotions is not supported by evidence from neural science. Recent brain studies suggest there is no single "motivational" system in the brain, and there is no monolithic "emotional" system either. In fact, the evidence rather indicates that there are many brain circuits, mediated by numerous neurochemistries, that coordinate and mediate

the set of processes we have referred to as the emotions (Panksepp, 1993). Furthermore, it is well known that these systems, be they the dopaminergic, the noradrenergic, cholinergic or serotonergic systems, influence and modulate each other (Panksepp, 1986; LeDoux, 1996; Introini-Collison, Dalmaz & Mcgaugh, 1996; Gill, Sarter & Givens, 2000), thus resulting in a set of distributed systems, perhaps interacting with each other, and at moments orchestrating the many different internal and external responses we have come to know and define as affective.

4.2 Emotional Systems of the Brain

This section reviews some of the aforementioned evidence regarding the existence of specific circuits situated in intermediate areas of the brain that have been conceptualized as sensorimotor emotional command systems, based upon earlier and excellent reviews by Panksepp (1998) and Panksepp (2000).

4.2.1 The Fear System

The fear system is perhaps the most studied emotional system in the brain. Interest in the neural substrate for the fear emotion emerged from studies of the Klüver-Bucy syndrome, a complex set of behavioral changes caused by damage to the temporal lobe and related structures in primates (Klüver & Bucy, 1939). Animals that have such kind of lesions suddenly lose their fear to stimuli that were previously considered to be threatening. They also exhibit an increased frequency of oral and sexual behaviors (i.e., they eat many different things that "normal" animals would find unpalatable and they attempt to copulate even with members of other species). Subsequent studies by Weiskrantz (1956) showed that lesions confined to a specific area of the brain called

the amygdala [1] produced the emotional aspects of the Klüver-Bucy syndrome.

A very basic circuit has now been identified that extends from the amygdala through the anterior and medial hypothalamus to the lower brain stem (through the periaqueductal gray (PAG)—a large structure in the midbrain, consisting of small to medium neurons surrounding the aqueduct of Silvus, otherwise known as the cerebral aqueduct and which has been thought to be involved in protection and defensive reactions, notably distress calls and affective defense), and then to specific autonomic and behavioral output components of the lower brain stem and spinal cord (Davis, Rainnie & Cassell, 1994). Fear behaviors can be elicited by artificially activating this circuit, and conditioned fears can be developed by pairing neutral stimuli with unconditioned stimuli that normally evoke this system. It is likely, however, that this is only one of the many circuits mediating fear.

The neurochemistry that controls this system includes the excitatory neurotransmitter glutamate and a variety of neuropeptides, including corticotropin releasing factor (CRF), neuropeptide Y, NPY, and the endogenous enzodiazepine-type system each of which may eventually be found to modulate slightly different kinds of fears (Fanselow, 1994).

Most of the current work on the fear emotional system has focused on the paradigm of fear conditioning, analyzing how learned inputs from the convergence of thalamic and cortical inputs into a nucleus of the amygdala called the lateral nucleus (LA), which then sends information concerning the cues that predict aversive events to the central nucleus (CeN) and down to various integrative and output components in the centromedial diancephalon and mesencephalon (LeDoux, 1996). It is well known that other forms of information, such as the contextual cues associated with aversive events enter the system through the hippocampus (Kim & Fanselow, 1992) and yet

[1] This set of subcortical nuclei was named the *amygdala* due to its almond-like shape

others from other higher areas of the brain.

4.2.2 The Anger System

It makes good evolutionary sense for the systems mediating fear and anger to be intimately related, for one of the functions of anger is to provoke fear in competitors, and one of the functions of fear is to reduce the impact of angry behaviors from threatening opponents. Thus, the brain circuits mediating this type of emotion are not surprisingly located close to those mediating fear (see 4.2.1. Some of the main circuits identified run from medial zones of the amygdaloid complex, through the lateral hypothalamus (LH) and down through the PAG, where further organization onto the output components of the brain stem and spinal cord is observed.

4.2.3 The Lust System

Sexuality, as an emotional system, has been widely neglected in emotion research. Nonetheless, many emotion theorists agree that this is a short-sighted view of the general concept of affect, which surely should involve the situations and responses related to sexual behavior.

Recent evidence from behavioral neuroscience suggests the existence of different neural circuits that mediate male and female sexuality in all mammalian species. These systems are dissociated, and in the case for male sexuality, they involve the preoptic areas of the hypothalamus. In the case for female sexuality, the same structure is implicated, namely the hypothalamus, only its ventromedial components (Becker, Breedlove & Crews, 1992) as cited by (Panksepp, 2000). Each of these systems, as it is with most of the systems organization throughout the neuroaxis, is modulated by higher circuits, especially those that stem from the medial amygdala, septal area, and the bed nucleus of the stria terminalis.

Likewise, the neurochemistries of the *Lust* male and female systems have a variety of components, including many hormonal regulators that we are just starting to elucidate and comprehend. However, it is well known that male sexuality is principally mediated by the hormone vasopressin, whereas female sexuality is mostly regulated by the hormone oxytocin and the leuteinizing hormone-releasing hormone (LH-RH) systems (Moss & Dudley, 1984). In terms of the orgasmic component that is at a center stage of the *Lust* system, and perhaps an important contributor to the learning of sexually related responses, many question still remain, but preliminary evidence suggests that when this component occurs and is experienced, it is regularly accompanied by release of oxytocin as well as endogenous opioid release within the brain (Petersen, Caldwell, Jirikowski & Insel, 1992).

4.2.4 The Care System

A very important system, especially as it relates to social organisms is that which mediates responses directed at caring for each other. It has been proposed that these affective urges evolved from preexisting sexual circuits (Panksepp, 2000). Thus, the hormone oxytocin, which as we reviewed plays an essential role in regulating female sexuality (see Section 4.2.3), is also a key player in the initiation of maternal behaviors (Panksepp, 1998). Oxytocin also plays a fundamental role by mediating the circuits that help deliver milk to babies when they are nursing. The stimulation of the mother's breasts is relayed as somatosensory signals that involve neurons in the hypothalamic paraventricular nucleus, which connect to the pituitary gland, thus releasing the hormone (Panksepp, 2000). This is not the only oxytocin system in the brain that regulates the *Care* system. In fact, multiple other systems exist, some of which have already been elucidated. As Panksepp (2000) describes:

For instance, if first-time mothers (at least of the animal variety) cannot

feel the surges of their brain oxytocin systems, then they do not rapidly develop maternal competence (Petersen et al., 1992). Indeed, the ancestral form of this hormone—namely, vasotocin—already helped deliver babies long time before mothers cared for their offspring. — Panksepp (2000, p. 148).

It is interesting to note that the same hormones that control different aspects of giving birth, as well as those related to feeding the young, are also essential for generating "care" feelings in the mother. How exactly these hormones control the affective feelings and responses associated to the *Care* system remains an open question, but knowing that there are specific brain circuits mediating social bonds provides further evidence for the existence of basic affect programs.

4.2.5 The Distress System

This system has been referred to as the *Panic* system by others (Panksepp, 1998; Panksepp, 2000), as it relates to the panic that comes about from separation distress or anxiety. The circuits for separation distress have been mapped in the brain by identifying those sites that produce the cries that young animals make when isolated from social companionship (Jürgens, 1976; Panksepp, 1998).

In the forebrain, the circuits are located in the bed nucleus of the stria terminalis, dorsal preoptic area, and ventral septal area. In the diencephalon, the circuits are concentrated in medial thalamic areas, and in the mesencephalon, in the dorsal parts of the rostral PAG. If one continues to descend caudally through the PAG, the areas that generated distress calls begin to produce cries of pain, which might suggest that separation distress emerged from more primal pain mechanisms (Panksepp, 1998).

In the cortex, the anterior cingulate has been implicated in the higher integration of social feelings (MacLean, 1990). Presumably, this system is involved in generating

a special psychic energy to social motivation. It is as if it psychologically hurts to be alone.

There are several interactions among these circuits, many of which are not yet known, but whose neurochemistries have been mostly identified. These systems can readily be activated by injections of glutamate and blockade of the same neurotransmitter receptors significantly reduce the calls produced by separation distress (Panksepp, 1995; Panksepp, 1998).

Another important elicitor of this affective system is CRF, while many other peptides have been implicated in the reduction of distress calls, especially opioids, oxytocin, and prolactin, which seems to make sense from an evolutionary perspective, given that these mediate social bonds as it was described in Section 4.2.4 (Panksepp, 2000).

4.2.6 The Play System

Although play systems have not been generally accepted in neuroscience or psychology as affective systems, we do briefly describe them here as they might be the precursors for systems mediating what we would refer to as *Joy* (Panksepp, 1993).

There is substantial evidence that systems mediating what could be called "rough-and-tumble" play exist in the mammalian brain. Many of the neural aspects behind this system are not well understood, but the evidence so far seems to indicate that critical components are directly related to the separation distress systems (see Section 4.2.5), which is reasonable from an evolutionary standpoint. The processes that may constitute basic systems for what we commonly refer to as *Joy* and *Sadness* should be closely intertwined in the brain.

These circuits have been located in the parafasicular area—which integrates somatosensory signals that promote play—and ventrorostral areas of the PAG, where

positive affective responses can be generated with brain stimulation. Conversely to separation distress systems, opioids increase the activity of *Play* circuits (Panksepp, 1995).

Several other systems exist. In fact, Ikemoto & Panksepp (1999) have described an important one which deals with issues of seeking out the resources that organisms need, in order to survive. This system has been called the *Seeking* system, and we will defer its discussion for later chapters, as it forms the basis for our notions on incentive salience (see Chapter 8).

4.3 Affective Strategies for Learning: Multiple Systems

In the previous section we reviewed some of the known neural substrates involved in affective processing, as emotional circuits of the brain akin to the affect programs we described in Chapter 2. Let us now focus on some of the things we know about the neural substrate for learning.

It is now a common view to identify a variety of aspects of learning and memory involving independent and perhaps parallel neural systems in the brain. For instance, a system involving the hippocampus is thought to be necessary for tasks that require the use of information about relationships between stimuli. A separate system involving the amygdala seems to mediate the formation and selection of behaviors based on the association of neutral stimuli with biologically significant stimuli that have affective connotations. This system is often seen as a simple associative learning system that acquires associations of the type stimulus-stimulus (S-S). Finally, a system including some of the different structures of the basal ganglia (e.g., ventral and dorsal striatum) is thought to mediate incentive learning and the formation of reinforced

stimulus-response (S-R) associations. The latter is also considered to be an associative learning system in which neutral stimuli come to release specific behaviors due to the repeated strengthening of the association between these stimuli and behaviors that have rewarding consequences (McDonald & White, 1993; Packard, Cahill & McGaugh, 1994; Packard & McGaugh, 1996).

A very interesting and open question relates to how these learning systems interact, working in parallel, in normal organisms. In an effort to obtain a better understanding of some of the issues involved, experimental studies have been developed for the analysis of information processing in these neural systems. Most of these studies involve the use of different approaches and tools, including electrolytic or neurotoxic lesions, transgenic techniques, as well as imaging technology. In contrast, very few, if any, have relied on the use of computational approaches.

The work proposed in this thesis attempts to contribute toward this end. Specifically, it focuses on the interactions among multiple learning systems, especially those with parallel associative learning schemes that are triggered by, and require the intervention of affectively significant (or unconditioned) stimuli. This includes those types of learning commonly referred to as *incentive learning* and *habit learning*, but excludes other types of learning that do not seem to require the presence of unconditioned stimuli, such as those described above which are mediated by the hippocampus and related structures.

4.4 Reward and Incentive Learning

Many stimuli and events in the environment have affective significance as they may bring beneficial or harmful consequences for an organism. According to the behavioral response they evoke, these stimuli have been labeled *appetitive* (rewarding) or *aversive* (punishing).

As elaborated by Thorndike (1911), and more recently by Schultz (1998), appetitive or rewarding stimuli have three important and distinct functions: (1) they interrupt behavior and elicit approach and consummatory behaviors; (2) they act as positive reinforcers, increasing the frequency of behaviors leading to them, and maintaining these learned behaviors by preventing extinction; and (3) they induce subjective feelings of pleasure (hedonia) and positive emotional states. Conversely, aversive stimuli elicit avoidance responses, act as negative reinforcers by increasing and maintaining avoidance behavior, and induce subjective feelings of displeasure and/or distress (Schultz, 1998).

An essential aspect of associative learning relates to the abilities of certain sensory stimuli to predict the occurrence of affectively significant stimuli. Predictions provide information about future events before they actually happen, allowing the organism to choose and prepare appropriate behavioral responses (e.g., stopping current behavior and initiating approaching or avoidance responses) in order to deal with the stimulus in an effective manner.

Many of the details on how this "affective tagging" process occurs are not completely understood, but it is known to involve the contingent presentation of a stimulus, just prior to the presentation of unconditioned stimuli. Notwithstanding, new findings from neuroscientific studies are providing evidence suggesting some of the neural substrates that mediate the prediction and processing of rewards.

4.4.1 Neural Substrates: The Dopamine System

The dopamine (DA) system appears to be one of the main actors involved in mediating incentive and reward-based learning. This system originates primarily from DA-containing neurons, henceforth referred to as dopaminergic neurons, in the substantia nigra (SN) and the ventral tegmental area (VTA), and projects to frontal

cortex and various structures of the basal ganglia, including the nucleus accumbens (NAS) and the dorsal striatum (DS). Supporting these ideas are the results of a variety of experimental studies that include selective lesions of the dopamine system, electrical self-stimulation, selective use of DA receptor agonists and antagonists, and self-administration of major drugs of abuse, such as cocaine, amphetamine, and alcohol (Wise & Rompre, 1989; Robbins & Everitt, 1992; Wise, 1996; Schultz, 1998; Chiara, 1999).

In the sections that follow, we will take a closer look at the DA system and review some related hypotheses of its function, as they provide interesting and useful ideas with respect to learning that have influenced and motivated some of the work proposed in this thesis.

4.4.2 Characteristics of Dopaminergic Neuron Responses

Dopaminergic neurons in SN and VTA of monkeys exhibit short, phasic responses to a variety of rewarding and reward-predicting stimuli (Romo & Schultz, 1990; Ljungberg, Apicella & Schultz, 1992; Mirenowicz & Schultz, 1994). These responses are fairly homogeneous regardless of the type of appetitive stimuli. In other words, dopaminergic neurons do not discern between various appetitive stimuli, such as food or liquids, nor between different sensory modalities, or between primary rewards and conditioned appetitive stimuli.

Responses of dopaminergic neurons to aversive stimuli have not been fully characterized. Mirenowicz & Schultz (1996) have shown that dopaminergic neurons do not respond in a similar and consistent way to aversive stimuli (Mirenowicz & Schultz, 1996). Other research seems to indicate they do (Trulson & Preussler, 1984; Abercrombie, Keefe, DiFrischia & Zigmond, 1989; Young, Ahier, Upton, Joseph & Gray, 1998). These contradictory results suggest that more research is

needed before reaching a conclusion in this regard.

Dopaminergic neurons are also activated by novel and salient stimuli (Horvitz, 2000). This response, however, habituates rapidly when the stimulus is repeated without any emotional and behavioral consequences. That is, the stimulus neither corresponds to, or predicts, a reward or punishment. An interesting aspect of this case is that activity of dopaminergic neurons seems to be correlated with overt behavioral-orienting responses (Schultz & Romo, 1990; Strecker & Jacobs, 1985).

The responses of dopaminergic neurons appear to be very dependent on the unpredictability of stimuli. If the presentation of a primary reward is repeated in a predictable manner, dopaminergic neurons do not exhibit a response at the time of the presentation of reward, instead, their activity is gradually transferred to the earliest reward-predicting stimulus (Figure 1A). In contrast, when a fully predicted reward does not occur, the activity of dopaminergic neurons exhibit a reliable depression that occurs at *exactly* the time of the expected reward delivery (Figure 1B).

In addition to the responses described above, dopaminergic neurons respond to stimuli that do not predict rewards, but which resemble reward-predicting stimuli in the same context (Ljungberg et al., 1992; Schultz, 1997). Schultz and colleagues interpret these responses as a stimulus generalization property of dopaminergic neurons. Usually, these responses are lower in magnitude and are followed by an immediate depression that is further ensued with activity after the occurrence of reward (Figure 1C), or no response in its absence (Figure 1D).

4.4.3 Hypotheses on DA Function

A variety of hypotheses that account for some of the experimental data relevant to DA function have emerged recently. Three of the most influential yet different interpretations of these data, all of which are relevant to this thesis, are briefly reviewed

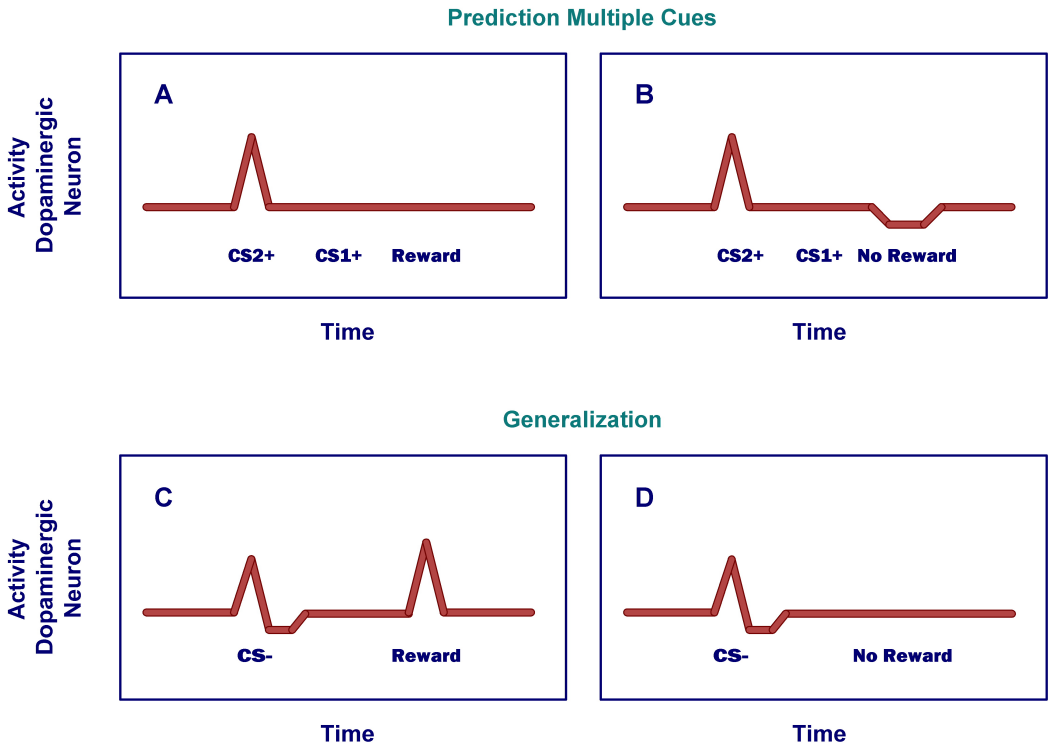

Figure 4-1: Responses of Dopamine Neurons

in the following sections.

The Effective-Reinforcement Hypothesis

A recent hypothesis that relates DA activity to formal theories of learning (Rescorla & Wagner, 1972) and to computational approaches in machine learning (Sutton & Barto, 1981; Sutton, 1988) has been proposed by Schultz and colleagues (Schultz, 1998; Schultz, Romo, Ljungberg, Mirenowicz, Hollerman & Dickinson, 1995; Schultz et al., 1997). According to their view, DA activity encodes expectation of reward, or more precisely, an error in the prediction both in time and magnitude of reward.

Reflecting on the characteristic DA responses summarized above in 4.4.2, they note that dopaminergic neurons emit a positive signal (activation) when an event

is better than predicted, a negative signal (depression) when an event is worse than predicted, and no response when the event is fully predicted. Therefore, dopaminergic neurons in the substantia nigra and VTA, suggest Schultz and colleagues, report primary rewards relative to the difference (or error) between the actual occurrence and the prediction of reward. By doing so, they may act as a global teaching or effective-reinforcement signal that is sent to other neural mechanisms, such as the striatum and prefrontal cortex, where it can mediate associative learning, as well as the consolidation of reinforced behavior learning and subsequent action selection.

This hypothesis has received increased attention in the last several years. Schultz and colleagues have provided supporting evidence, and have linked it to computational approaches in reinforcement learning. In particular, they note that the temporal-difference (TD) learning algorithms (Sutton, 1988), resemble in all major respects the responses of DA neurons and, to some extent, the basic anatomy and connectivity of the basal ganglia (Schultz, 1997).

Nonetheless, and based on the same data reviewed in 4.4.2, objections to this effective-reinforcement hypothesis have been raised. Of special interest are the findings relative to the responses of dopaminergic neurons to novel and salient stimuli, as well as the less characterized responses to aversive stimuli. In the case of the former, it is reasonable to assume that a novel event may be appetitive (involving reward), aversive (involving punishment), or it may be insignificant (no affective consequences). Considering that according to this view dopaminergic neurons encode an error in prediction of reward, why would all novel stimuli be evaluated as initially rewarding? That is, better than predicted? Moreover, this also means that behaviors that are active at that moment, or some time before that, will be reinforced or maintained, which might not be appropriate.

Along these same lines, Redgrave, Prescott & Gurney (1999*b*) make a compelling argument against this hypothesis when they note that the response of dopaminergic

neurons, presumably corresponding to the evaluation of the affective significance of a stimulus, occurs before or during the saccadic response that is made to foveate the stimulus for further analysis (Redgrave et al., 1999*b*). This observation is significant because it means that any computations made by dopaminergic neurons signaling the evaluation of a novel object would have to be performed even before the object has been "looked at".

Now, with respect to the case involving responses to aversive stimuli, if DA neurons encode information about effective reinforcement would it not be also reasonable to expect that their activity show some depression when aversive contingencies occur?

In view of these issues, we will consider other hypotheses that have been suggested with respect to the involvement of DA neurons in reinforcement, incentive, and habit learning.

The Switching Hypothesis

A different interpretation of the data has been presented, which suggests that the dopamine signal is an essential component in the processes that are involved in reallocation of limited attentional and behavioral resources necessary to deal with unexpected, salient (biologically significant) stimuli (Oades, 1985; Redgrave, Prescott & Gurney, 1999*a*).

Underlying this view is the notion that for an organism to deal with a rewarding stimulus in an effective manner, it must first interrupt ongoing behavior and switch attentional and behavioral resources. Thus, it is suggested that the basal ganglia act as a central selection device evolved to resolve conflicts between different systems competing for limited cognitive and motor resources. In the appropriate contexts, the basal ganglia would implement their selection processes by disinhibiting sensorimotor connections of winning systems and inhibiting those of losing ones (Redgrave et al., 1999*a*). At the core of such implementation would be the activity or depression of the

DA signal, which would promote switching in the first case, and focus on currently selected resources in the second one.

Furthermore, it has been suggested that the DA signal may also be useful in binding salient stimuli to a selected action. Once this link is created, it may be strengthened or weakened depending on subsequent signals indicating outcome. According to this, dopamine activity would have a more general role in associative learning and it would not be limited to rewarding contingencies as it is the case with the effective-reinforcement perspective.

The Incentive Learning Hypothesis

A final theory on DA function to be considered in this thesis corresponds to that of Ikemoto & Panksepp (1999), who argue that ascending meso-accumbens DA systems (i.e., projections from VTA dopamine neurons to the NAS) constitute a general purpose system that is important in sensorimotor integrations that facilitate flexible approach responses to a variety of salient stimuli (Ikemoto & Panksepp, 1999).

The functional role of NAS DA signaling is divided into two main phases: *activation of flexible approach-seeking responses* and *incentive learning*. This refers to the notion that NAS DA may first be involved in facilitating exploratory activities (approaching stimuli for investigation) and more gradually in signifying importance of novel stimuli because of their association to opportunities for consummatory behavior (incentive learning).

It should be noted that from this perspective, incentive learning is not limited to appetitive events, but also includes aversive ones. In such contexts, NAS DA facilitates avoidance responses, only these are reformulated as approaches toward "safety". Thus, the same flexible approach systems are at play in both appetitive and aversive contingencies.

Another interesting aspect of this view is the recognition of two separate types

of approach responses: a flexible response system, as described above, which operates when organisms are learning about incentive contingencies, and a habit response system that operates based on over-learned incentive responses. This habit system allows organisms to acquire and maintain procedural performance. The nigrostriatal DA system appears to be an important structure involved in habit formation and stimulus-response learning (Knowlton, Mangels & Squire, 1996; Graybiel, 1998; Hikosaka, 1998).

From the three theories discussed above, this approach-seeking view seems to account for the most evidence with respect to some of the possible roles of DA signaling. In addition, it fits perfectly well with the idea of affect programs described earlier: the approach-seeking system integrates and coordinates attentional, perceptual, affective, behavioral and learning processes needed for adaptive approach toward goals.

4.5 Summary

In this chapter we have briefly reviewed some of the emotional brain circuits that exist in the mammalian brain. We have also reviewed current knowledge regarding the neural substrates for learning, emphasizing on the mesolimbic dopaminergic system, which has been widely implicated in the process of incentive salience.

Several theories regarding the function of this dopamine system were reviewed. Work on this thesis suggests that these theories are not necessarily fundamentally contradictory, and depending on the issues considered nor do they need to be mutually exclusive. The following chapters describe the implementation of affect programs that include many of the ideas represented in each of these views, including separate sensorimotor pathways for preparatory and consummatory behaviors, systems for the detection of novel and salient contingencies, an organization in which the detection of natural (unconditioned, innate) stimuli automatically promotes learning, and a

system in which the predictions of affectively significant stimuli play an essential role in the specification of contexts and the gradual occurrence of stimulus-response or habit learning.

Chapter 5

Experimental Platforms

Earlier we described our overall goal as one of understanding emotions from a computational perspective. We have argued that in order to achieve this goal, we should focus not on the general theoretical debates that have traditionally occupied the field of emotions, but rather on understanding specific problems, such as understanding specific emotions and their function as solutions to particular situations that organisms face in their environment. In particular, we are interested in understanding how affect programs act as the main glue that bind together, in a coordinated fashion, a variety of subprograms that govern perception and attention, motivational priorities, action selection, learning, and motor control.

To this end, it is necessary to be able to understand the computational problems faced by organisms situated in their environments, evaluating events, interacting and selecting significant stimuli, and determining solutions for the multiple contingencies they face. In recent years, we have developed a number of robotic platforms, both physical and simulated, that allow us to accomplish this. We have created several robots and situated them in the physical world (or a high fidelity simulation of it) where they face an uncertain and dynamic environment, dealing with noisy sensors and actuators, and constantly facing multiple challenges, including the tasks of at-

tending to and selecting relevant stimuli, determining the best course of action given any situation, and deciding when and what to learn about their environment in order to adaptively achieve (or maintain) some specific goals.

This chapter describes these robotic platforms and the world they inhabited. Both these robotic platforms and their environments constituted our experimental platforms where different scenarios, which are described in further chapters, were instantiated and evaluated, in our efforts to achieve our primary goal of understanding the underlying computational architecture of emotion.

5.1 Yuppy, an Ugly Pet Robot

Over the last several years, the Humanoid Robotics Group at the MIT Computer Science and Artificial Intelligence Laboratory has been building a variety of robots that have humanoid forms or that are otherwise human friendly in the sense that they posses an increasingly complex behavioral repertoire that includes human-like competencies. One of the first robotic platforms used to experiment with the notion of affect programs was a mobile robot named Yuppy, shown in Figure 5-1. This robot was designed as a proof of concept to test the validity of some of the concepts behind the use of computational models of emotion as the basis for robotic control.

This robot was intended to be a personal pet robot, albeit an ugly one. Yuppy was used as the initial testbed for the computational model of emotion called Cathexis (Velásquez, 1996) and some of the ideas involved in emotion-based robotic control (Velásquez, 1998*b*), that will be discussed with more detail in Chapter 6.

Yuppy, which has been since then decommissioned, had a number of onboard sensors, including two color CCD cameras, one of which was mounted on a two degrees of freedom head that was used for both navigation and for "looking" at objects of interest. It also included an active stereo audio system consisting of two

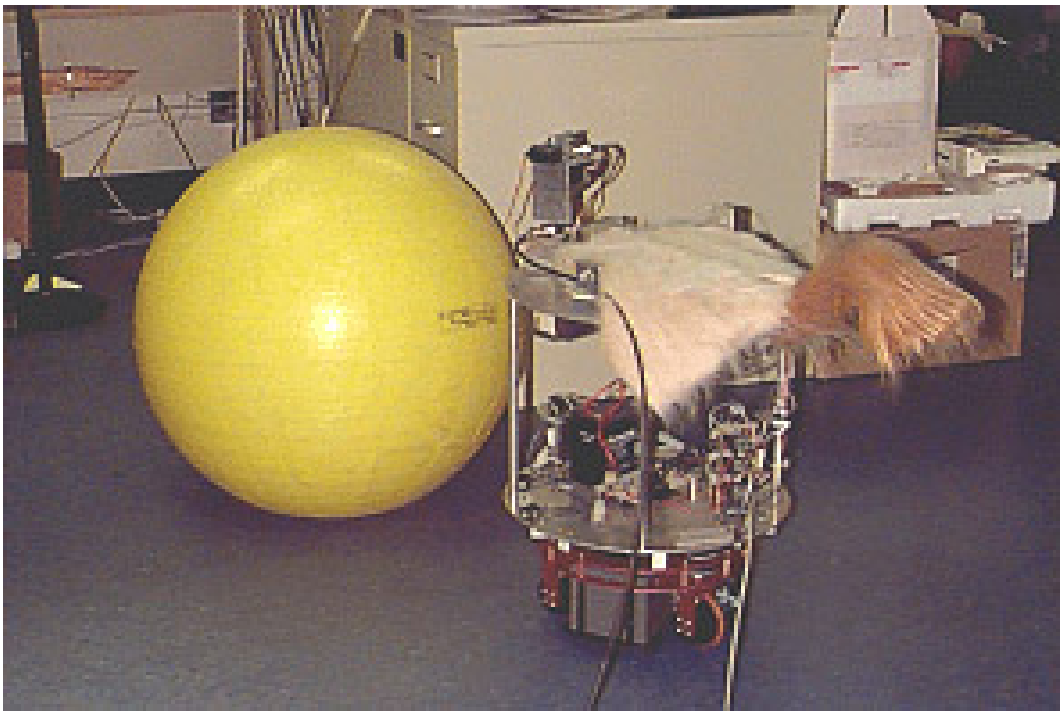

Figure 5-1: Yuppy, an affective robot built at the MIT Computer Science and Artificial Intelligence Laboratory

microphones mounted on Yuppy's ears (which, in our commitment to ugliness, decided to place in the robot's body, instead of its head). Yuppy's sensory modalities also included eight IR proximity sensors mounted around the robot's body. Furthermore, a pressure sensor that simulated a patch of touch-sensitive skin was placed on top of its body. Likewise, a pyro-electric sensor used to detect changes in temperature due to the presence of people was mounted in Yuppy's head. Finally, a system that sensed changes in head and body orientation and alignment was used as a simple proprioception system.

5.1.1 Computational Platform

The robot used a commercial mobile platform (RWI B12) which has a synchronous drive system that allowed it to translate and rotate. The drive system included three wheels that are kept parallel all the time. These three wheels are all driven, providing appropriate traction for locomotion.

The B12 platform includes an on-board NEC 78310 microcontroller. This microcontroller controls the motors and the battery system. The motor control for steering and driving uses optical feedback from the motors' encoders and pulse width modulation (PWM) to power them. The microcontroller also reads the voltage and current from the batteries. Communication to the base is achieved through an RS-232 serial interface using ASCII commands. The command format is given by a two letter mnemonic command and a hexadecimal number. The commands allow the robot to read the status of the base and to change the motors' positions, velocities, and accelerations. Information about the commands can be found in The B12 platform includes an on-board NEC 78310 microcontroller. This microcontroller controls the motors and the battery system.

5.2 Coco, a mobile baby gorilla robot

An aesthetically more pleasing robotic platform was later built to continued developing and testing the ideas presented in this thesis. This robot was named Coco, and was modeled after the morphology of a baby silverback gorilla. The robot, illustrated in Figure 5-2, was intentionally designed to be small (approximately 0.5 meters in length weighing 9.1kg). In order to maximize both mobility and the possibility of approaching humans in a fairly realistic manner, the robot was quadrupedal. The characteristic long forelimbs and shorter hind legs of the gorilla were accentuated in Coco to provide duality of usefulness in walking and the possibility of generating ex-

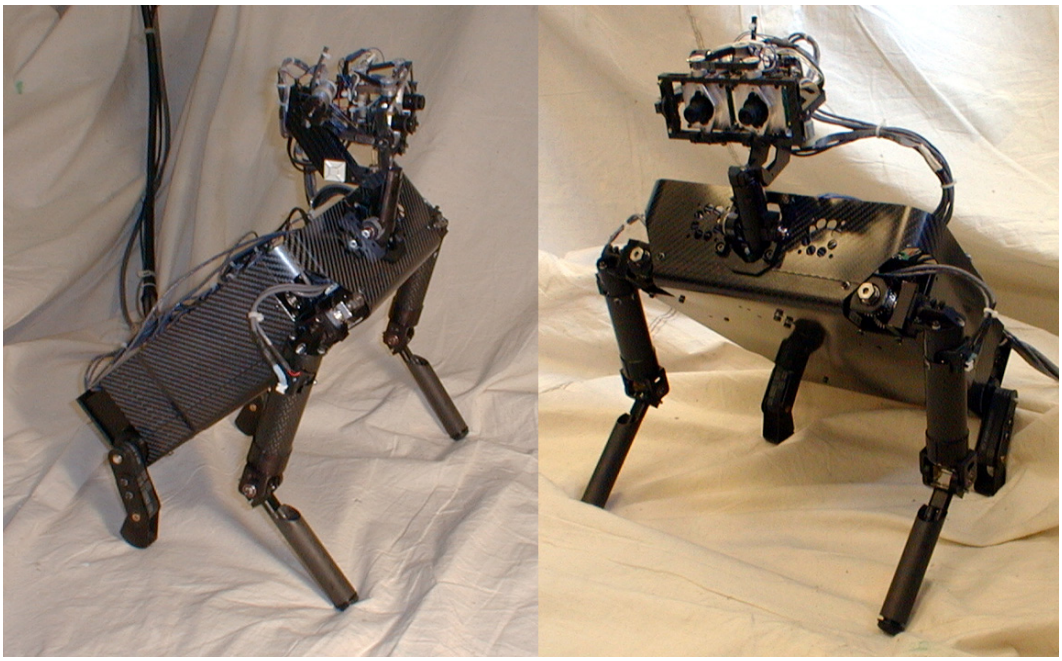

Figure 5-2: Coco, a baby gorilla robot

pressive gestures. Coco had two degrees of freedom in its shoulders and one degree of freedom in each hip, knee, and elbow. Coco also had a five degree-of-freedom vision head. The robot's design took into account the need for on-board DSP control that actuates servos for each degree of freedom. A serial link provided high-level position commands to the robot from an off-board computer array. Coco was designed to be modular so improvements to actuators and sensors could be made incrementally. The limbs and head bolt into a monocoque body chassis that housed the motor-control electronics.

Vestibular System

The human vestibular system plays a critical role in the coordination of motor responses, eye movement, posture, and balance. The human vestibular sensory organ consists of the three semi-circular canals, which measure the acceleration of head

rotation, and the two otolith organs, which measure linear movements of the head and the orientation of the head relative to gravity. To mimic the human vestibular system, Coco used a three-axis inertial sensor from Intersense (www.isense.com). The sensor consisted of a single integrated remote package and a processing module. The remote sensor is mounted on the robot's back in a position that allowed it to move with the robot's head but remain stationary when the eyes were moving (similar to the positioning of our own vestibular organs). The sensor delivered both the angular accelerations in roll, pitch, and yaw and an absolute angular measurement in two dimensions with respect to the gravity vector. The sensor processing module communicated through a standard serial RS-232 interface to the main processing system.

5.2.1 Computational Platform

Coco's computational platform included a network of off-the-shelf PC computers. As part of the experiments described here only 10 processors were used, ranging in speed from 600 MHz to 1 GHz, but the cluster was expandable to many more nodes. Processors were interconnected by a 100 Mbps ethernet. Each processor ran Windows NT as the main operating system. MPI, a message passing standard, was used to create a flexible C++ code base that provided robust interprocess communication over the network. MPI has been used by the high-performance computing community to write parallel applications for large cluster of computers (Forum, 2002). The robot was connected to this computational platform through commercial video digitization boards and through commercial motor control boards (MAX2000^{TM} 3-axes distributed motion controllers from Agile Systems, www.agilesys.com).

5.3 Marvin, a Simulated Robot

All of the experimental scenarios that were instantiated and tested with the two robotic platforms just mentioned were also replicated in a simulated environment. A robot named Marvin, was developed in a computer graphics 3D world, using a physics based engine named ODE and a variety of simulation tools, based upon the Gazebo simulation engine.

Gazebo is part of the Player and Stage projects that originated in the Robotics Research Labs at the University of Southern California. Player is a networked device server, and Stage is a simulator for large populations of mobile robots in 2D environments. Gazebo, created by Koenig & Howard (2004), is a high fidelity outdoor simulator that provides a realistic environment for robotic simulations in a 3D world. Similar in nature to its 2D counterpart, Stage, Gazebo can simulate a population of robots, objects and sensors. Since both Gazebo and Stage are compatible with the Player environment, client programs written using one simulator can usually be run on the other with minor modifications (Koenig & Howard, 2004)

Gazebo, was designed to be able to accurately reproduce the dynamic environments a physical robot would encounter in the real world. It is capable of providing realistic sensor feedback, and simulate several of the physical properties of the robotic models it uses. All simulated objects have mass, velocity, friction, and other attributes that allow them to behave realistically when pushed, pulled, knocked over, and so forth. All robots in this simulated world are dynamic structures composed of rigid bodies connected via joints. Forces, both angular and linear, can be applied to surfaces and joints to generate locomotion and interaction with an environment. The world itself is described by landscapes, extruded buildings, and other user created objects. Almost every aspect of the simulation is controllable, from lighting conditions to friction coefficients (Koenig & Howard, 2004).

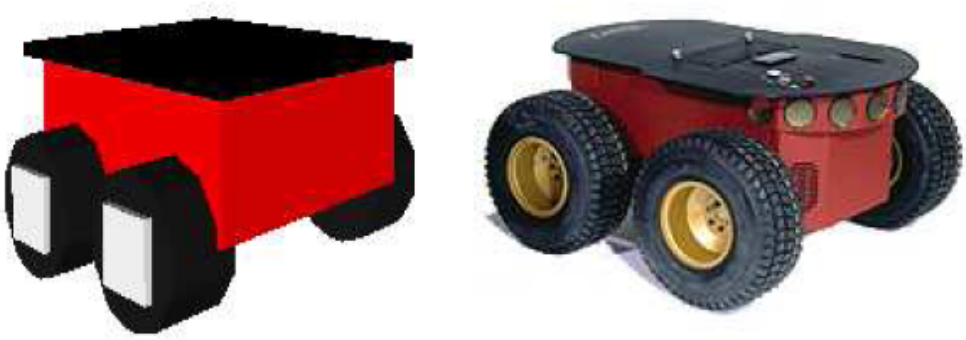

Figure 5-3: A Pioneer 2DX Robot in the Gazebo Simulator. Gazebo is a high fidelity robotics simulator for 3D environments that is capable of providing realistic sensor feedback as well as an accurate dynamic environment, such as those that physical robots would face in the real world. This figure illustrates how a physical robotic base such as the Pioneer 2DX system can be simulated in the this engine. Adapted from (Koenig & Howard, 2004)

Marvin was modeled as a Pioneer 2DX robot, by using some of the features available in Gazebo and ODE. The robot model, illustrated in Figure 5-4, was equiped with several sensory systems, including a SICK LMS200 scanning laser range-finder, and a Stereo Vision Head, which was used to obtain visual information of the simulated world. A variety of algorithms for visual processing, such as blobfinding, and stimuli saliency, were implemented by using open-source tools including OpenCV and CMVision.

5.3.1 Marvin's World

The world Marvin inhabits, depicted in Figure 5-5, is an open-ended simulated world with blocks and objects that represent different objects in Marvin's environment, ranging from resources, to predator-like agents, or simple obstacles.

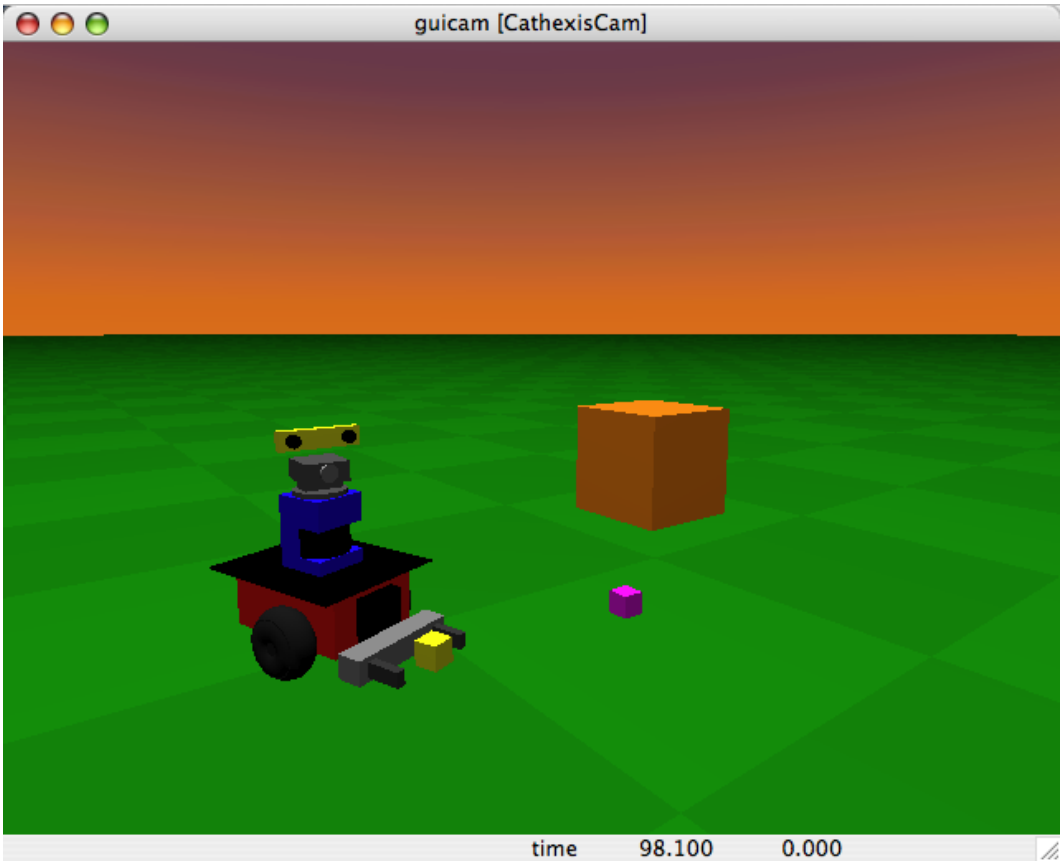

Figure 5-4: Marvin the Affective Robot. Marvin is a simulated robot that has been endowed with affective systems in the form of affect programs.

5.4 Summary

Several robots were built and simulated at the Humanoid Robotics Group at the MIT Computer Science and Artificial Intelligence Laboratory as the main experimental platforms to explore and address some of the issues related to the involvement of affect in intelligent behavior, and which go beyond the regulation of social interaction through emotional expression. Yuppy was primarily used to explore the involvement of affect in behavior regulation. On the other hand, both Coco, with a gorilla-like morphology, and some human-like sensorimotor systems, and Marvin, a simulated robot, were useful platforms for exploring a more integral and comprehensive compu-

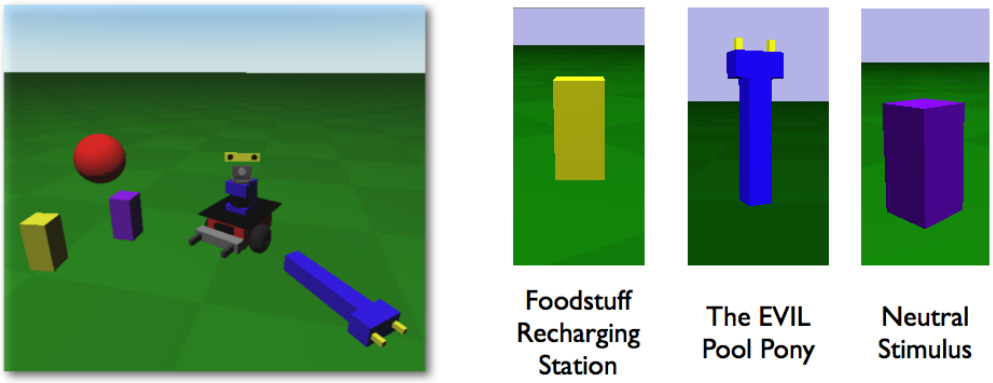

Figure 5-5: Marvin's World. The world Marvin inhabits is a simulated 3D world with open spaces where every so often it may encounter a variety of objects, some of which are of significance to the robot, some others which are simply obstacles or objects that have no pre-determined significance. This figure illustrates some of these objects, including the main affective stimuli which correspond to Marvin's predator (the Evil Pool Pony) and Marvin's recharging station (the Yellow Block, which represents foodstuff). Grabbing this Yellow block will increment Marvin's battery level to some determined value. Likewise, other neutral stimuli, such as the purple blocks or the red spheres, also form part of this simulated world.

tational model that places affect as the cornerstone in regulating behavior, attention, perception and learning. All of these robots were used to support the work described in the following chapters.

Chapter 6

Engineering Affect: The Cathexis Framework

> *The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions. I suspect that once we give machines the ability to alter their own abilities we'll have to provide them with all sorts of complex checks and balances.*
> — Marvin Minsky (1986, The Society of Mind)

As one of the main contributions of this thesis we propose a unified computational framework for the study of affective phenomena which attempts to capture the essential features of emotional systems at multiple levels of abstraction, as well as at several spatiotemporal scales, from the inner workings of emotional appraisal, to the network coupling between emotional systems and the coordination and modulation of other systems be they motivational, attentional, learning or motor control. This chapter describes the main issues and design decisions involved in the implementation of this computational framework, as well as its implementation details and some of its first applications in the control of robots. The framework, named Cathexis[1], extends

[1]From the Greek *kathexis* meaning holding or retention. The term was introduced into the literature as a translation of Freud's term "besetzung", which connotes a concentration and transfer

previous work (Velásquez, 1996; Velásquez, 1997; Velásquez, 1998*a*), and has been used to create a variety of affect-based systems that have proven useful in providing insights about the inner workings of emotional systems and their interaction with other psychological constructs we tend to associate with intelligent behavior.

6.1 Scope and Design Principles

Our main challenge focused on the development of a computational framework that could be used as a tool to explore and study different theories, concepts and models related to affect. Our main motivation for such endeavor highlights our interest in understanding affect from a computational perspective, as we expressed in our introductory remarks (See Section 1.4). Consequently, this thesis focuses on the problem of building the computational (and in some cases the physical) infrastructure needed to support these sorts of tasks.

Greatly influenced by the work reviewed in the previous chapters, which involved different perspectives on affect from evolutionary psychology, ethology, and the neural sciences, this work integrates theories and concepts from these diverse viewpoints to build a synthetic model of affect that observes the following design principles[2].

6.1.1 Deep Model of Affect DP 1

As we have reviewed in the previous chapters, affect involves a wide variety of complex phenomena, most of which is not completely understood. Considering the design and evaluation of a computational model that synthesizes affect, this poses a problem

of emotional energy onto an object or idea.

[2]These design principles have been captioned with DP Number. Throughout this thesis, whenever a description of the implementation or evaluation of the Cathexis model relates to any specific design principle this will be made clear by including the design principle's caption in such description.

with respect to deciding what makes such a model "good" or "appropriate". In Section 10.3, we review several different models and architectures related to this work which have been proposed for a variety of applications. Clearly, affect can be studied at many different levels, and all approaches offer different insights into this complex set of phenomena. Undoubtedly, there is no "silver bullet" with respect to models of affect. Different models will work best in different domains, and thus, how appropriate they are depends greatly on their specific goals and purposes.

Notwithstanding, given our main goal of contributing work that will get us one step closer to understanding affect from a computational perspective, our main design principle is to build a *deep model of affect* that is as comprehensive as possible. By *deep* we mean a framework that delves farther down into many of the computational issues that individuals (or robots) face while situated in their environments, attempting to maintain any particular set of goals. If you recall, these were the set of problems that *constituted* the emotions as we defined them in Chapter 2. Thus, our main goal is to contribute toward a computational framework that allows us to explore many of these computational issues, attempting to understand the sorts of representations (at various descriptive levels) of the prototypical situations or problems that the individual (or the robot) will address, and the different programs of specialized solutions, that will be implemented as the synchronization of appraisal mechanisms, action arbitration strategies, attention modulation algorithms, learning models and motor control subprograms. In relation to this, Picard (1997) suggests that a comprehensive model would account for some of the following phenomena:

Emergent Emotions and Emotional Behavior DP 1.1

A computational model might not explicitly synthesize emotions. That is, it might not have explicit abstractions representing emotions or any other kind of affective phenomena. Notwithstanding, such a model might be able to generate behavior that

appears to arise from emotion processes or is otherwise perceived as emotional. An example of such kind of mechanism can be seen in Braitenberg's vehicles[3] (Braitenberg, 1984). This notion of emergent affect is a very interesting one, and one that is widely exploited in most systems that use emotions as the main mechanisms to mediate social interaction. These emergent emotions rely heavily not on the underlying emotional mechanisms, but rather on the individual observers of the "emotional" behavior. As such, much of the power comes from the observer, and not from the models themselves.

Primary Emotions DP 1.2

Besides emergent emotions and emotional behavior, a comprehensive computational model should be able to produce fast, emotional responses to specific situations. These emotions correspond to the more primitive, innate or pre-organized kind of emotions, which some refer to as the *primary emotions* (Damasio, 1994). From our theoretical standpoint, we defined this kind of emotions in Chapter 2 as the *affect programs*. Thus, a computational model for this kind of primary emotions should account for a set of domain specific programs, each functionally specialized for solving different adaptive problems (e.g., avoid danger or seek out resources to solve impending internal motivational problems) and which become active by a different set of environmental situations.

[3]Braitenberg's vehicles are machines—described by Valentino Braitenberg in his book: *Vehicles: Experiments in Synthetic Psychology*—in which the direct coupling of sensors to actuators (e.g., direct connections, cross-wired or inhibitory), produces "creatures" that are extremely simple, but which, to an observer, give the appearance of exhibiting emotions, such as "fear", "aggression", "love", or "affection".

Secondary Emotions DP 1.3

Primary emotions do not describe the full range of emotional phenomena that complex organisms exhibit. Some emotions involve much more detailed cognitive processing of the different contingencies that arise in an organism's environment. These more cognitively generated emotions may involve an evaluation of the event especially as it concerns the organisms's goals, beliefs, attitudes, and expectations. Emotions of this kind have been referred to by many as *secondary emotions* (Damasio, 1994), and correspond to the lines of thought reviewed in Section 2.2.1 (Ortony et al., 1988; Roseman, 1984; Lazarus, 1991; Johnson-Laird & Oatley, 1992)

Emotional Experience DP 1.4

A system that synthesizes emotion can also address the many issues related to the emotional experience per se, which may involve awareness at many different levels including an awareness of the cognitive and physiological changes produced by an emotion episode as well as the corresponding subjective feeling.

Addressing this issue would involve addressing the many complex and open questions related to consciousness, which despite of their interesting and important nature lie outside of the scope of the current work[4].

Integration With Other Phenomena DP 1.5

From our perspective, affect is believed to be a central point of coherence for the integration of the many different mechanisms that help an organism exist and survive in its environment, obtain energy, reproduce, and pass on its genes onto the next generation. More specifically, affect is inherently intertwined with many different

[4]It is interesting to note, however, that some researchers have recently claimed that emotional mechanisms might also be responsible for some of the aspects involved in what we would call consciousness (Damasio, 1999; Panksepp, 2005).

processes and mechanisms involved in the generation of intelligent behavior. A deep model of affect should address or at least point to possibilities for addressing some of these interactions, which include, but are not limited to:

- Differences between emotions and other affective phenomena, such as moods and temperament
- Affect-mediated action selection and the generation of goal-directed behavior
- Affect and regulatory mechanisms (e.g., motivational systems such as hunger or temperature regulation; interactions with immune systems)
- Affective Learning
- Affective-Cognitive interactions
  - Affect-mediated attention
  - Affect modulation of memory
  - Influences of affect in perception
  - Affect and high-level decision-making (as opposed to low-level which we consider to be related to the choice of actions)

One of the main ideas put forward with this work is precisely that affect can be an appropriate *"glue"* that ties together and coordinates many processes and constructs involved in generating intelligent behavior. Thus, as it will be explained in the following sections, the proposed model interacts with and regulates other systems that mediate perception, attention, behavior, learning and motor control

6.1.2 Applicability to Robot Control DP 2

Aside from being useful to explore concepts, models and theories of affect, we believe that affect is a good abstraction to decompose the problem of high-level control of

an agent, be it a software agent or a physical one, such as a robot. In other words, affective processing is well suited to act effectively as an integration mechanism by which activity in many different systems is bound together in a coherent manner. This has been defined elsewhere as *Emotion-Based Control*—the control of autonomous agents that relies on, and arises from, emotional processing (Velásquez, 1998*a*).

6.1.3 Biological Feasibility DP 3

As a final design principle, and considering our main interest of understanding affective processing in humans and animals, we approached this work with the goal of developing a computational model of affect that was biologically plausible. In particular, we were interested in the possible neural substrates for some of the main issues considered herein, such as the functional roles of the basal ganglia, hippocampus, amygdala, and other brain structures, in the processing of affect, behavior generation and selection, motivation, and learning. For that we drew inspiration from much of the evidence stemming from different disciplines, and especially from research in Psychology and Neuroscience (LeDoux, 1993; Graybiel, 1995; White, 1997; Panksepp, 2000).

6.2 The Affect Program Abstraction

What is an appropriate computational abstraction to represent an emotional system? As alluded to before, the answer to this question depends greatly on the kind of application and purpose of the model that would involve such an abstraction. Keeping the aforementioned design principles and issues in mind, our approach to this end follows a psychoevolutionary perspective to the study of affect, based upon the notion of *affect programs* defined in Section 2.2.3 (Darwin, [1859]/1998; Izard, 1971; Ekman & Friesen, 1986; Panksepp, 1998; Cosmides & Tooby, 2000).

We defined affect programs as executive, operating systems that generate and

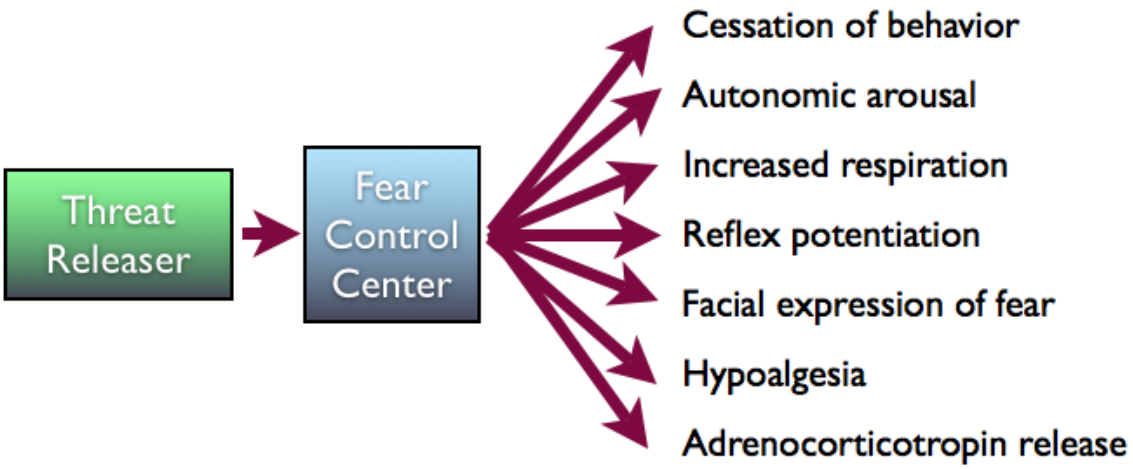

Figure 6-1: A Fear Affect Program. This figure illustrates an instance of a fear affect program as thought to exist in the mammalian brain. Different brain structures are involved, most notably the amygdala, which is thought to mediate the evaluation of the affective significance of threatening stimuli and the further coordination of various responses associated with the brain emotional system mediating fear. A threatening stimulus is detected by perceptual systems. This information is sent to the fear control center (roughly corresponding to the lateral (input) and central (output) nuclei of the amygdala), which then activates and coordinates a variety of responses mediated by different brain centers. Adapted from LeDoux (1996) and Iversen et al. (2000).

coordinate short-term, stereotypical responses that allow organisms to deal with biologically significant events in ways that promote survival. These responses involve a variety of elements such as facial and behavioral expressions, arousal of the Autonomic Nervous System (ANS), vocal expressions, modulation of attention, and affective feelings.

As described in Chapter 4, several affect programs of this kind have been suggested to exist in the mammalian brain, including circuits that mediate paradigmatic emotions such as *fear*, *anger*, *surprise*, *joy*, *sadness* and *disgust*, and other not so paradigmatic such as *incentive seeking*, *maternal care*, and *separation distress*.

Figure 6-1 illustrates an example of the fear affect program which is thought to

Figure 6-2: An abstraction for affect programs. This figure illustrates the main abstraction of the Cathexis framework: the Affect Program. An Affect Program couples sensors and actuators through affective processing. It consists of a set of filtering mechanisms, named Releasers, which provide information about the different stimuli in the robot's environment, an Affective Evaluation Unit that assesses the affective significance of said stimuli and coordinates appropriate actions, named here as Preparatory and Consummatory Behaviors. Information through this affective processing mechanism flows from left to right.

be mediated by the amygdala. Affect programs of this kind constitute the primary theoretical constructs in our approach to understanding affect from a computational perspective.

Drawing inspiration from these theoretical constructs, we propose the Affect Program[5] as the primary abstraction that comprises the computational framework described in this thesis. This abstraction is illustrated in Figure 6-2.

An Affect Program couples the robot's sensors and actuators through affective processing. It consists of a set of filtering mechanisms, named Releasers, an Affective Evaluation Unit that assesses the affective significance of the perceived stimuli and coordinates the issuing and control of an associated set of specific responses, name here as Preparatory and Consummatory Behaviors.

[5]Throughout this thesis, we use a different notation in order to distinguish among the different abstractions and elements of the framework, and the theoretical constructs that carry the same name. Consequently, an Affect Program would refer to the model's abstraction, whereas an affect program (no formatting) would refer to the concept

Each of these components will be described in more detail later, but for now let us mention the general flow of information for an Affect Program:

1. Releasers filter sensory data and detect relevant contingencies that are relevant to the Affect Program they are associated with.

2. Information about the detected contingencies is sent to the Affective Evaluation Unit which assesses the affective significance of such contingencies (i.e., determine whether a stimulus is biologically significant—or of importance to the agent's goals).

3. If the contingency is distal (in space or time) the Affect Program's control determines the response that would prepare the agent to deal with it. This is done through Preparatory Behaviors which are described in more detail later.

4. If the contingency is proximal the Affect Program's control determines the specific response or Consummatory Behavior that would deal with it appropriately, based on the sensory-specific properties of the event.

5. The appropriate behavior is selected (based on the two immediate points above) and executed.

6.2.1 Relation to the Affect Program Concept

The Affect Program abstraction has a direct correspondence to the theoretical concept defined in Section 2.2.3. Thus, Releasers correspond to perceptual systems in the brain that form part of the affect program circuits, and which provide information regarding the different contingencies an organism experiences in the world. The Affective Evaluation Unit corresponds to the executive, command centers that evaluate the biological significance of events and coordinate the appropriate responses, if any. Finally,

Figure 6-3: An instance of a Fear Affect Program. This figure illustrates a specific instance of the Affect Program abstraction. The Fear Affect Program showed here includes a set of Releasers that detect a threatening stimulus and determine the level of the threat depending on whether it is distal or proximal. The Fear Affective Evaluation Unit, depending on the significance of the stimuli would coordinate the associated behaviors, which in this case correspond to a Freezing Behavior, a Fearful Expression, and the internal mediation of mechanisms for Reflex Potentiation

the Preparatory and Consummatory Behaviors, correspond to the set of short-term and stereotypical responses that are commonly associated with emotional episodes, and which would involve not only relational behavior, such as emotional expressions and specific actions, but also internal responses and influences on other systems, including the release of hormones in the bloodstream, the modulation of mechanisms of attention, or the interactions with other affect programs (see Section 6.3.3).

6.2.2 Instances of Affect Programs

Figure 7-3 illustrates a specific instance of the Affect Program abstraction. The Fear Affect Program shown in the figure has a set of Releasers that include the detection of a threatening stimulus and determine the level of the threat depending on whether it is distal or proximal). Given all available information, the command center that is part of the Fear Affect Program evaluates the affective significance—with respect to the fear affect program—of the perceived stimuli. If the stimuli are significant, then a

coordinated execution of all responses (Preparatory and Consummatory Behaviors takes place.

As with this Fear Affect Program, several other affect programs representing discrete emotions can be instantiated and modeled as part of the Cathexis framework. Each instance will thus have its own control and evaluative mechanism, together with its efferent and afferent connections to Releasers and Behaviors, respectively. Thus, each Affect Program will receive information from its own set of Releasers, process it in its own style, and influence behavior. These differences in the *information processing style* of each of the affect programs is at the essence of the model.

As part of the three robotic systems that were described in Chapter 5, we instantiated several variations of the following affect programs:

1. **Seeking**: An affect program that mediates appetitive motivational situations. It is primarily concerned with leading organisms (in our case the robots) to pursue goals in their environment. The term *Seeking* was coined by Panksepp (1998) to reflect the idea that organisms "seek out" the fruits of their environment. In this work, we restrict the use of the *Seeking* affect program to the notion of seeking solutions to impending goals, such as physiological needs mediated by regulatory mechanisms. In Chapter 10, however, we speculate further with this notion and suggest that a mechanism such as this might be the foundation for seeking solutions and alternatives to other, more abstract, situations as well.

2. **Surprise**: This affect program deals mediates the robot's attentional resources. As such, it responds primarily to novel and sudden stimuli and regulates the primary active attention responses, which include head and body Orienting Responses (ORs).

3. **Fear**: An affect program that synchronizes a variety of subprograms, in response

to dangerous contingencies or threatening stimuli. Depending on the specific robot, these threatening stimuli include dark environments where the robot's vision sensors are of little use, and more material stimuli such as predators (e.g., the "evil blue pool pony" described in Section 5.3.1).

4. **Joy**: This affect program mediates all responses to rewarding stimuli and situations.

5. **Distress**: An affect program that deals primarily with punishing stimuli and separation distress contingencies, when referring to social contexts.

6. **Anger**: This affect program deals primarily with those situations that restrict freedom of action for the robot or that impede access to its resources.

6.3 A System-Level View of the Framework

Theoretical constructs can be viewed and described at multiple levels. For instance, traditional approaches to building control systems for robots decompose the problem into a set of functional subsystems. Even more recent approaches that are greatly influenced by behavior-based approaches, such as the pioneering work of Brooks (1986), and which should then imply a behavior-based decomposition, exhibit nonetheless an overall architecture that is decomposed not into behaviors, but rather into a series of functional subsystems that include, in most cases, a perceptual system, a motivation or drive system, a behavior system, and a motor system (Tyrell, 1994; Blumberg, 1996; Breazeal, 2000).

If we consider the main components of the Affect Program abstraction, which will be described in more detail in the following sections, one could also view the Cathexis framework from a similar perspective, as illustrated in Figure 6-4. However, such a functional systems view would only have explanatory purposes as the subsystems

described in this figure *do not really exist* as part of the implemented framework. Instead of a traditional functional decomposition, the Cathexis framework is implemented as a set of Affect Programs that interact with each other, and which reunite all the functionality that is associated to each of these systems. In other words, the problem of high-level control of a robot follows an affect-based decomposition approach in which Affect Programs are used as the main entities that define a complete control system. This is akin to Brooks' *Subsumption Architecture* (Brooks, 1986), and other behavior-based approaches (Maes, 1989; Arkin, 1990; Tyrell, 1994; Blumberg, 1996), with an emphasis on the following notions:

1. The decomposition is not behavior-based, but rather affect-based. This seemingly trivial difference actually has considerable implications. As an abstraction for emotion, Affect Programs have a specific information processing style, which constraints the kinds of evaluations based on the input to each Affect Program and the kind of outputs it produces. Thus, the organizational principles of the Affect Program abstraction are at the essence of the computational model. This difference in organization further suggests that the decomposition is not based upon the desired external behaviors of the system, but rather on the set of prototypical fundamental situations that the agent will encounter, together with the set of coordinated responses available and necessary to face such situations, all of which is captured by the affect program theoretical construct as described in Chapter 2 and its possible neural underpinnings reviewed in Chapter 4.

2. No layered control is implied by the architecture, although Affect Programs can build new layers of functionality by adapting their structure both on their input (adding new Releasers) and output sides (adding new connections to Behaviors) as it will be described in further chapters.

3. Parallel processing is a specific property of the model that implies that all affect

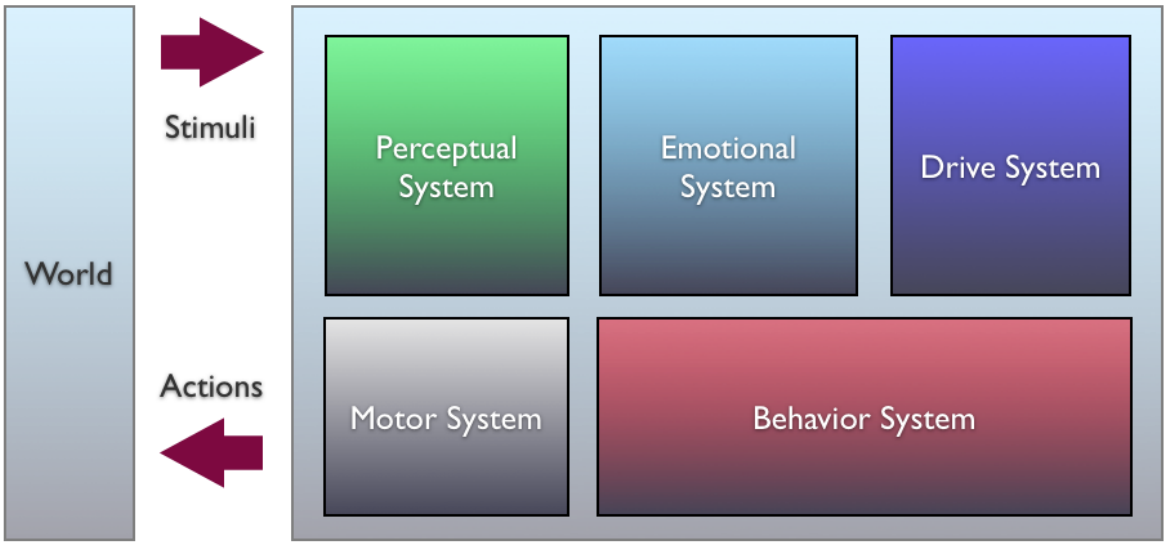

Figure 6-4: A Hypothetical Functional Decomposition of the Cathexis Framework. This figure illustrates a hypothetical functional decomposition of the Cathexis framework. However, this functional decomposition has but explanatory purposes as the subsystems described in this figure *do not really exist* as part of the implemented framework. Instead of following a functional decomposition, the Cathexis framework is implemented as a set of Affect Programs that interact with each other, and which reunite all the functionality that is associated to each of these systems. Thus, it would be more correct to think that *each* Affect Program internally implements these functional systems as part of its information processing style.

> programs have access to all sensors the agent possesses as well as to all of its actuators. Naturally, only those sensors and actuators that are relevant for the affect program solution would require access. This notion furthermore suggests that Affect Programs compete for control of the overall system in a manner that will be described in Section 7.1.

6.3.1 The Systems Concept

It is still useful to describe an architecture from a system-level perspective. As we hinted above, the main systems in the Cathexis framework are not the subsystems

usually found in traditional decompositions (e.g., perceptual system, behavior system), but rather correspond to the set of modeled affect programs. In this framework we make the basic assumption that normal behavior involves a continuous flow of information through each of these independent systems (i.e., affect programs). The systems process (i.e., filter, combine, associate, and even alter) this information. The resulting output ultimately controls and regulates behavior, either directly or indirectly. It may be the case that under certain situations some part or parts of an affect program may be changed by the information being processed, and this change will alter the processing of similar information on future occasions, which results in a corresponding change in the output of the system. When these changed outputs result in observable modified behavior, this modification is attributed to the process of adapting or "learning." We will defer discussion on the ideas of learning an memory to Chapter 8.

6.3.2 Parallel Processing

Traditional models of affect have been parsimonious in their descriptions of the affective functions of the brain and are mostly based on a single set of emotional concepts involving a single processing style (i.e., a single emotion system that generates—mostly through cognitive appraisal—all possible emotions) (Arnold, 1969; Ortony et al., 1988; Roseman et al., 1990; Johnson-Laird & Oatley, 1992). As we reviewed in Chapter 4, there is no such thing as a single monolithic emotional system in the brain, but rather several distinct systems, each with their own processing style. More than one affect program continually processes affective information and influences behavior. This idea leads to the concept of parallel processing. As illustrated in Figure 6-5, this concept means that several independent systems mediate affective information processing simultaneously and in parallel. These are the systems corresponding to

Figure 6-5: Affect Programs as systems in the Cathexis framework. This figure illustrates a more appropriate system-level view of the Cathexis framework. In this type of decomposition, all Affect Programs are seen as systems that receive information and influence behavior. Each of these systems has full access to all of the robot's sensors and actuators. Despite the vertical organization of the figure, no hierarchy is implied, but the notion of *parallel processing* does apply: All systems have access to the same information and may process it simultaneously, but each system has its own *processing style* as described in the text.

specific instances of Affect Programs. In the brain, these would correspond to the neural circuits described in Section 4.2.

6.3.3 Affect Programs Interactions

The notion of parallel processing suggests that Affect Programs compete for control of the overall system. Depending on the specific contingencies that the robot encounters in its environment, specific Affect Programs will become active and will synchronize and coordinate an appropriate set of responses to deal with any particular situation. It may be possible, however, that multiple Affect Programs become simultaneously

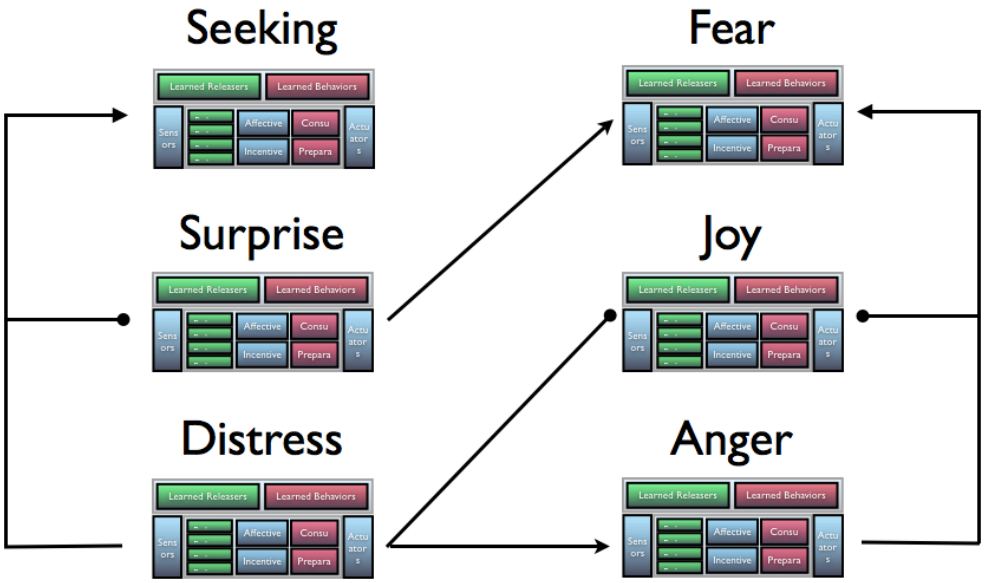

Figure 6-6: Interactions Between Affect Programs. This figure illustrates a hypothetical set of interactions (inhibitory and excitatory inputs) between Affect Programs The framework is general enough that any interaction of this kind can be modeled. Thus, the *Distress* Affect Program might contribute to the activity of *Anger*, whereas it might inhibit *Joy*, for instance.

active. To deal with any conflicting situations (e.g., when conflicting affect programs cannot be co-activated) an arbitration mechanism has been set in place as it will be described in Section 7.1. Furthermore, Affect Programs can interact by providing inhibitory or excitatory input to one another. Thus, it is possible to mediate these conflicts by including inhibitory connections between Affect Programs that could not be co-activated, such as between the *Joy* and *Distress* affect programs for instance, or even promote other interactions as suggested in Figure 6-6.

6.4 A Network of Basic Computational Units

Let us lower our descriptive level a bit and consider for a moment the details of the implementation. The Affect Program abstraction is implemented as circuits of nonlinear computational units that correspond to the basic processing element of the framework. Figure 6-7 illustrates the notion of a basic computational unit. Each unit has a set of inputs, an integrative mechanism that may filter, combine, and alter such inputs, and a set of outputs that influence other units in the system.

Each computational unit has specific properties that are shared (and in some cases refined) by all components of an Affect Program. In other words, Releasers, Affective Evaluation and Control Units, and Behaviors are all implemented by the same computational elements. Thus, an Affect Program is composed of several of these elements, connected into specific, functional circuits (e.g., a fear circuit composed of perceptual systems to detect dangerous contingencies, a command node to evaluate the affective significance of such events, and several responses (behaviors) to act upon to them).

The general form of the response or activation of each basic unit is a nonlinear function of its inputs and the strengths of their connections, or weights, as described in Equation 6.1

$$A_i(t) = f(\sum_k (S_{ki}(t) \cdot W_{ki}(t))) \tag{6.1}$$

Where $A_i(t)$ is the activation of system i at time t; $S_{ki}(t)$ is the value of input k and $W_{ki}(t)$ is its associated weight at time t, where k ranges over the set of inputs for system i. Finally, f is a limiting function, such as the standard ramp and logistic functions, which limit the activity of these units to be within certain ranges.

The following sections provide detailed descriptions for each of the elements that compose the Affect Program abstraction and review the main ideas behind the work

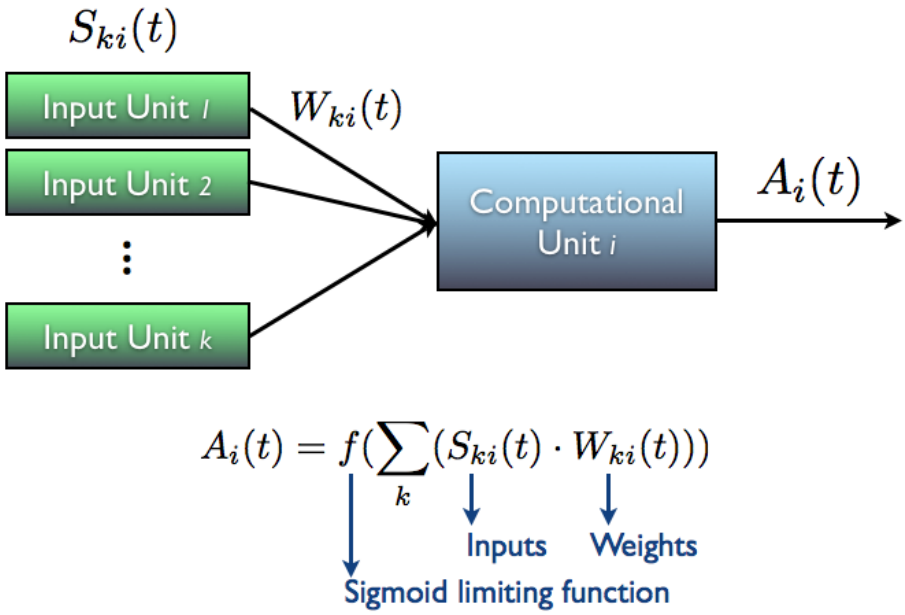

Figure 6-7: Basic computational element. This figure illustrates the basic processing units that compose the Affect Program abstraction. The same units are used to implement all components of the abstraction. The main computational properties of these units are shared by all components. In some cases, however, specific components may specialize the information processing style of the element.

that influenced and motivated such design decisions. But before delving into details, let us consider a sample scenario to illustrate the main ideas related to the affect programs construct.

6.5 Gorillas in Our Midst: A Sample Scenario

Imagine for a moment that you are a well-known primatologist studying the natural habitat of the Silverback Gorilla in the East African savanna and its surrounding forests. On a perfect day, you have stationed your safari vehicle behind a set of trees where an adult Silverback sits down close to two other young gorillas, who are playfully pushing and shoving each other in a manner typical of "rough and tumble"

play. Suddenly, a loud crackling noise is heard up above in the tree branches, and both you and all gorillas direct their attention toward the tree branches in anticipation of some explanation for the sound. Sure enough, soon after a leopard is revealed through the foliage. Its cover blown by the crackling sound of broken branches, the leopard hesitates to jump to the ground and stays up above in the tree, lurking in the branches awaiting for the right time to go after its prey. In an instant, the two young gorillas have initiated an escaping response, while the adult Silverback has assumed a defensive, yet aggressive posture. You feel shaky, your heart is beating faster and an urge to escape the situation is all that occupies your mind. Yet, despite the fact that your body might be telling you otherwise, this urge is inhibited by the realization that you are safe inside of your vehicle, where you finally stay to capture the moment.

Back to reality. In a scenario like this, a variety of affect programs might become active at different times and their different components process information from the environment and coordinate and execute appropriate responses. For instance, a *Play* affect program, such as that described in Section 4.2.6 might be the active system controlling the behaviors exhibited by the two young gorillas at the beginning of the scenario. When the loud noise occurs, however, activity in this system might be inhibited by a *Surprise* affect program that would attempt to gain control of the animals' attentional resources to identify the nature of the sound in order to determine whether it was affectively significant or not. The detection of the predator threat would likely activate a *Fear* affect program in each of the characters of our scenario, which would synchronize a variety of responses. In the case of the two young gorillas, a set of responses preparing them to escape would ensue, whereas in the adult Silverback a different set of responses which would include a fighting stance would have been active. Likewise, a different set of responses would have been coordinated in your case, given our hypothetical scenario.

We will continue exploring this scenario in further sections, as we describe the

different components of the Affect Program abstraction. Let us start with the notion of Releasers.

6.6 Releasers: A Window to the Robot's World

Nikolaas Tinbergen (1951) and Konrad Lorenz (1973) suggested possible mechanisms that trigger stereotypical behavior in organisms[6]. An important element of their work involved the notion of an *Innate Releasing Mechanism (IRM)*, which essentially corresponds to a hypothetical filter-trigger complex that initially filters a stimulus to determine if it is a sign stimulus—an external signal that evokes a particular action—and then triggers the corresponding behavior.

The main idea behind the IRM notion is that a stereotypical behavior occurs only under specific contexts. Thus, when an organism is not under the appropriate circumstances, the "energy" to motivate the behavior builds up inside its nervous system. The more it builds up, the easier it is to elicit the behavior. According to Tinbergen (1951), this energy is held in check by a central neural mechanism that is under constant inhibition from another center. This other center is the IRM. When the organism perceives an appropriate stimulus, its corresponding IRM is activated, which in turn "releases" the inhibition on the central neural mechanism, allowing the behavior to be executed.

Drawing upon these ideas and building upon previous work (Velásquez, 1997; Velásquez, 1998*b*), we implemented the Releaser[7] abstraction which corresponds roughly to an IRM. Releasers may be associated with any other computational elements, such

[6]This work gave Lorenz and Tinbergen, together with Karl von Frisch, the 1973 Nobel Prize in Physiology or Medicine for their discoveries concerning organization and elicitation of individual and social behavior patterns

[7]Releasers were previously referred to as "Sensors" or "Elicitors" in Velásquez (1996) and Velásquez (1997).

as the Affective Evaluation and Behavior units (see Sections **??** and 6.9). In essence, Releasers filter sensory data and identify special conditions that provide excitatory (positive) or inhibitory (negative) input to the system they are associated with. Releasers thus act as a mechanism that provides an assessment of perceptual information detecting specific conditions in particular stimulus which would prove of affective significance and therefore should motivate behavior.

To illustrate these ideas further, let us think back to our gorillas scenario. In this hypothetical example, the gorilla's sensory systems would collect data from the environment, which in this case would very likely include data regarding the crackling sound, the leopard, and several other stimuli occurring at that time. Let us further suppose that the gorillas have a set of Releasers that filter these sensory data and determine whether sign stimuli are present or not. In this scenario, a Loud Sound Releaser and a Predator Releaser might be part of the gorillas' Releaser repertoire, and would determine whether the sensory data regarding the crackling sound and the shape coming out from the trees actually correspond to the sound and morphology of a predator, respectively, in which case this information would be fed into the Affective Evaluation and Behavior units of different affect programs (e.g, *Surprise* and *Fear*) where further processing would ensue in order to determine what responses would be appropriate under the circumstances.

6.6.1 Kinds of Releasers

As we have discussed above, each Affect Program has associated a set of Releasers that constantly check for the appropriate conditions that would evoke an affective response. In contrast to other computational models proposed to date that emphasize the notion of cognitively generated emotions, we consider both cognitive and non-cognitive kinds of Releasers for the different instances of Affect Programs. Influenced by Izard's multi-

system for emotion activation (Izard, 1993), these Releasers have been divided into four different theoretical groups:

1. **Neural**: Considers the effects of neurotransmitters, brain temperature, and other neuroactive agents that can lead to an affective response. Many of these are regulated and can be affected by hormones, sleep, diet, and environmental conditions. For instance, there is a great deal of evidence that shows that decreased levels of norepinephrine and serotonin are associated with depression (Meltzer & Lowy, 1987). Similarly, it is clear that several chemical agents, such as carbon dioxide, yohimbine, and amphetamines produce anxiety in humans by activating the noradrenergic system (Charney & Redmond, 1983).

2. **Sensorimotor**: This type of Releasers includes sensorimotor processes, such as facial expressions, body posture, muscle action potentials, and other central efferent activity, which are not only important in regulating ongoing affective responses, but can also elicit them in the first place. Some evidence supporting this type of elicitors comes from neuropsychological studies in which experimenter-directed manipulation of facial muscles, composing a specific emotional expression, produces the subjective feeling corresponding to that emotion, as well as emotion-specific patterns of ANS activity (Ekman et al., 1983).

3. **Motivational**: This system includes all motivations that lead to emotion. In this model, motivations include the set of regulatory mechanisms discussed later on in Section 6.7 and which would represent physiological needs and internal goals. 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 & Davidson, 1986), pain or aversive stimulation causing distress, or the levels of physiological variables such as blood sugar, temperature, sodium levels and so forth.

4. **Cognitive**: This system includes all type of cognitions that activate emotion, such as appraisal of events, comparisons, attributions, beliefs and desires, memory, and so on. Previously, Velásquez (1997), defined this set of Releasers to correspond to one of the taxonomies popular in the cognitive appraisal theory view of affect (Roseman et al., 1990). In an effort to adhere to the design principle of biological plausibility (see Section 6.1.3), the present work does not explicitly model Cognitive Releasers as a mapping of such a taxonomy, but rather suggests that these can be acquired as the robot interacts with its environment and starts making associations between the stimuli it encounters and the different affective responses they might release. A proposed first step toward the acquisition of this kind of Releasers relates to the idea of incentive learning, which is the main topic of Chapter 8.

Natural and Learned Releasers

One final distinction we make with respect to Releasers, regardless of the theoretical group they belong to, is that they can be innate and hard-wired (*Natural Releasers*), or they can be learned (*Learned Releasers*). All Releasers discussed so far (and throughout this chapter) are natural releasers. We use the notion of learned releasers to represent stimuli that tend to be associated with, and are predictors of, the occurrence of natural releasers. This type of Releasers and the mechanisms and processes involved in their generation will be discussed in detail in Chapter 8.

6.6.2 Habituation of Releasers

Nonassociative learning is an interesting, and often overlooked, type of learning in which an organism acquires information about the properties of a single stimulus by being exposed to it repeatedly.

All Releasers in the Cathexis framework posses short-term memory and thus can habituate to stimuli. To achieve this, we have implemented a model of feedforward habituation based on the work by Dragoi (2002). This model allows us to produce suppressive and facilitatory habituation effects depending on the nature of the stimulus. For instance, repetitive presentation of a stimulus that is not affectively significant decreases the strength of the output response of the Releaser (habituation effect), whereas the evaluation of a stimulus as being of affective significance acts as a facilitatory effect, following the initial presentation of the stimulus, which increases the output response strength (recovery or dishabituation). Much like the work of Staddon (1993), this model explain rate sensitive properties of habituation and dishabituation, both of which serve our purposes of modeling the influences of affect in attentional mechanisms, as it will be described in Chapter 8, where this property of Releasers (habituation) play a very important role in attention and coherence. As we will demonstrate, habituation allows organisms (robots in our case) to become accustomed to initially distracting stimuli, and to learn to perform more effectively in an otherwise noisy environment.

The model works by processing each stimulus of interest to the specific Releaser via two different pathways: one inhibitory and one excitatory. Figure 6-8 illustrates this idea, with a stimulus X being fed into two different processing units: I and E, which correspond to these different processing pathways, and the output R (Releaser's strength), which is sensitive to the difference between inhibition and excitation of these two units.

The specifics of the model are described by the following set of interrelated equations describing the activity of both inhibitory and excitatory units and the output response.

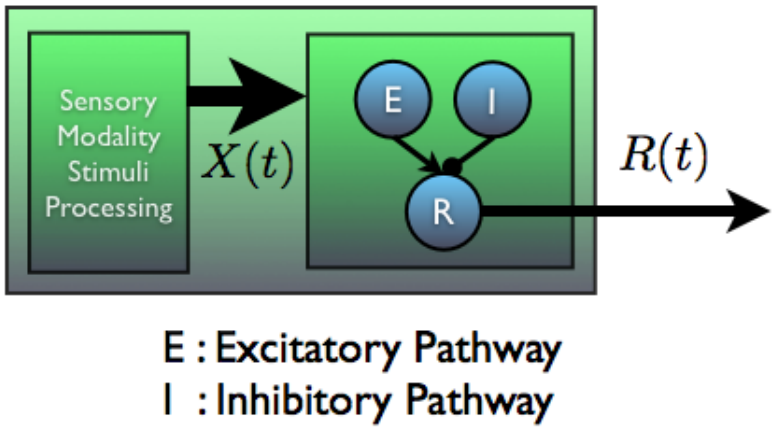

Figure 6-8: Releasers' Habituation Mechanism. This figure illustrates the model of habituation implemented for all Releasers in the framework. This habituation model processes a stimulus X through two different pathways: an inhibitory one (that goes through neuron I) and an excitatory one (that goes through neuron E). The response R is determined by the set of equations described in the text, and accounts for both suppressive and facilitatory habituation effects.

$$I(t) = I(t-1) - \alpha_1 I(t-1) + \alpha_2 X(t)[1 - I(t-1)] \tag{6.2}$$

$$E(t) = E(t-1) - \beta_1 E(t-1) + \beta_2 X(t)[1 - E(t-1)] \tag{6.3}$$

Where $X(t)$ represents the input stimulus, α_1 and α_2 are the decay and increase rate constants associated to the inhibitory neuron I, and β_1 and β_2 are the decay and increase rate constants associated to the excitatory neuron E.

The strength of the releaser's output R is controlled by the difference between the excitatory and inhibitory units as described in the following equations, where F is a sigmoid function used to limit the response, where q controls the slope of the sigmoid.

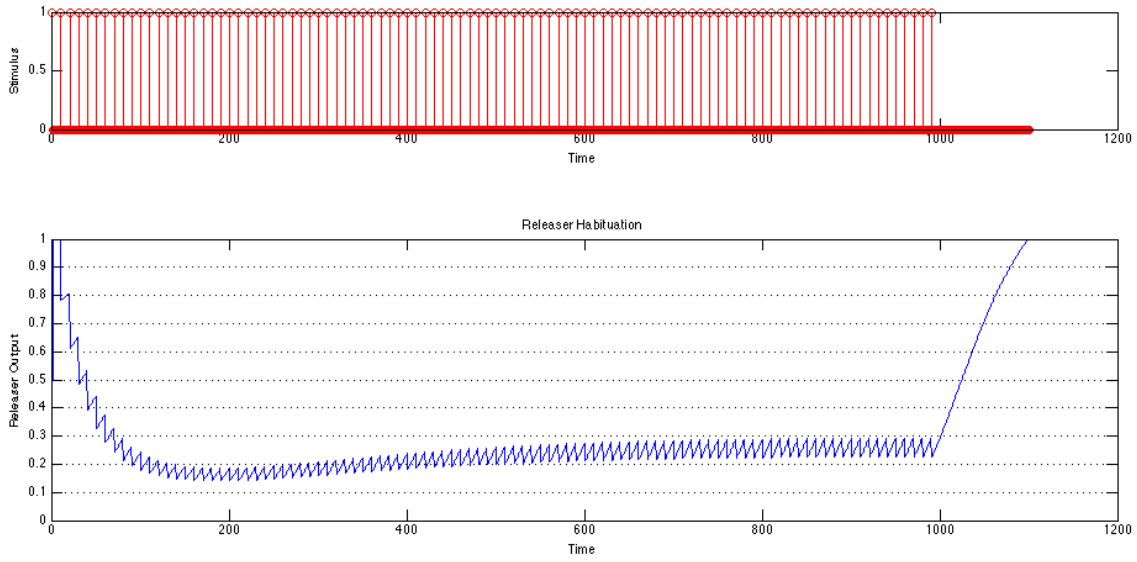

Figure 6-9: Results of Releasers' Habituation. This graph illustrates the results of rate sensitive habituation and recovery when a train of stimuli is presented for any particular Releaser (upper graph). The Releaser's output response R is plotted in the lower graph. The following parameters where used: $X(t)$ is set to 1 when the stimulus is present, and 0 otherwise. $\alpha_1 = 0.1$, $\alpha_2 = 0.3$, $\beta_1 = 0.05$ and $\beta_2 = 0.08$.

$$R(t) = 1 - F(max[0, I(t) - E(t)]) \tag{6.4}$$

$$F(x) = \frac{1 - e^{-qx}}{1 + e^{-qx}} \tag{6.5}$$

Figure 6-9 illustrates the results of the rate sensitive habituation and recovery model incorporated into the processing of every Releaser. In this example, a train of stimuli is presented for the Releaser to process (upper graph). The Releaser's output response R is plotted in the lower graph. The following parameters where used: $X(t)$ is set to 1 when the stimulus is present, and 0 otherwise. $\alpha_1 = 0.1$, $\alpha_2 = 0.3$, $\beta_1 = 0.05$ and $\beta_2 = 0.08$.

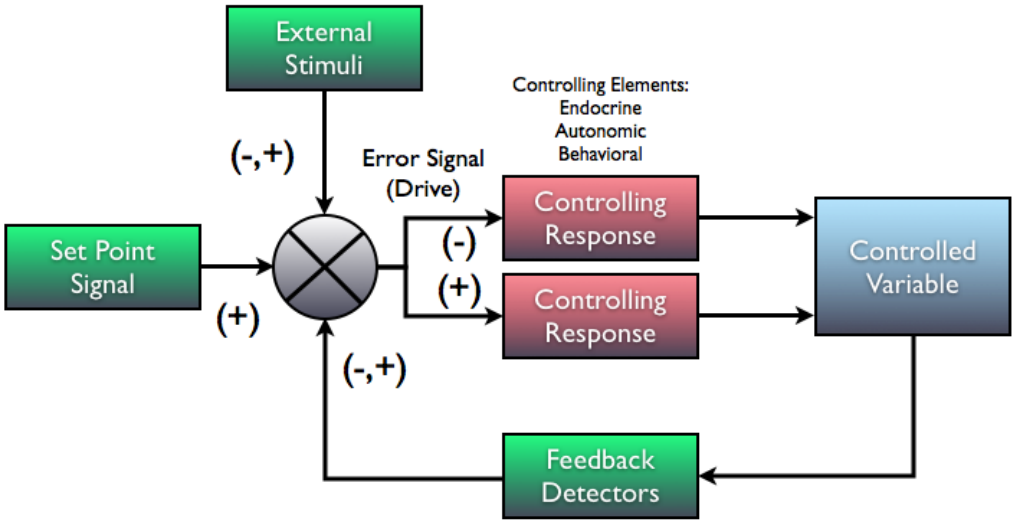

Figure 6-10: Regulatory Mechanisms. This figure illustrates the concept of drives or regulatory mechanisms seen from a control systems perspective. A specific control system regulates a controlled internal variable (e.g., temperature). When feedback detectors signal that the variables is above or below the set point an error signal is generated, which in turn facilitates appropriate regulatory responses. Adapted from Kupfermann et al. (2000, p. 1000).

6.7 Regulatory Mechanisms

Most models of affect that have been embedded in robot (or agent) control systems include the notion of *drives* as part of the main elements of a motivation subsystem (Blumberg, 1996; Cañamero, 1997; Breazeal, 2000; Konidaris & Barto, 2006). Breazeal (2000), for instance, relies heavily on homeostatic regulation mechanisms to regulate social interactions with her robot.

In contrast to these models, we explicitly differentiate between emotions and drives, not only in terms of their motivational value, but also in terms of how they are implemented in the framework.

With respect to their motivational value, we follow an approach similar to that of Tomkins (1962), which suggests that drives are cyclical in nature and are associated

with and satisfied by a relatively restricted range of stimuli. Emotions, on the other hand, are not cyclical, they can be related to an enormous variety of phenomena, and can motivate an equally wide range of cognition and actions. Thus, although drives (seen as error signals) may appear to ultimately result in controlling responses (e.g., behavioral, autonomic, or endocrine) that correct the error, they do so not because they have intrinsic "motivational power", but rather because they are associated with specific affect programs that regulate this kind of affective processes.

In Chapter 8, we will argue, based on existing evidence that suggests a functional role for specific emotional brain systems, that regulatory mechanisms such as those described here and commonly referred to by many as drives, are mediated by a specific kind of affect program: The Seeking Affect Program.

To implement drives we follow a control systems approach and model them as a particular kind of internal Releasers called Drive Releasers. Figure 6-10 illustrates the notion of a drive seen as an error signal within a regulatory mechanism. In such a mechanism, a controlled internal variable is first measured through some of the robot's internal sensors and then compared to a desired value or set point. If its value does not match the set point, an error signal (the drive signal) is produced. Since this signal is computed as part of the information processing performed by a Drive Releaser, it can be sent to whatever system the Drive Releaser is attached to, where it can be combined with the activity from other Releasers.

6.8 Affective Evaluation

The Affective Evaluation unit is at the core of the notion of affect programs. It corresponds to the main processing center that evaluates the affective significance of events and coordinates and synchronizes appropriate responses that deal with such contingencies and ultimately lead the robot to exhibit robust and adaptive behavior.

The evaluation of affective significance is also related to the notion of emotion intensity. This is undoubtedly an open question in the field of emotion research. In this thesis, we propose that the affective value of events is directly related to the intensity of the emotional episodes they generate. Thus, we believe these notions to be equivalent. The evaluation of affective significance of events, for each affect program, is performed through the same type basic computational units described in Section 6.4. However, their processing style is extended with functionality that considers the excitatory (positive) and inhibitory (negative) input from other Affect Programs, as well as temporal decay to model how the intensity of an emotional episode decreases over time. The activity pattern of the Affective Evaluation Units follows the description of Equation 6.6

$$A_i(t) = f(g(A_i(t-1)) + \sum_k (R_{ki}(t) \cdot W_{ki}(t)) + \sum_l (\mu_{li}(t) \cdot A_l(t))) \qquad (6.6)$$

Where $A_i(t)$ is the activation of affect program i at time t; $A_i(t-1)$ is its activation at the previous time step; g is the function that controls the temporal decay of the activation of affect program i; $R_{ki}(t)$ is the value of Releaser k and $W_{ki}(t)$ is its associated weight at time t, where k ranges over the set of releasers for affect program i; $\mu_{li}(t)$ is the strength of the excitatory (positive) or inhibitory (negative) input from affect program 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.

With respect to our gorillas scenario, the contingencies detected by the gorillas' Releasers would be assessed by the Affective Evaluation Units of specific affect programs, such as *Play*, *Surprise* or *Fear*, as described in this scenario. For instance, the presence of other young gorillas would be affectively evaluated as significant by the *Play* affect program. In turn, when the loud crackling noise occured, the *Surpirse*

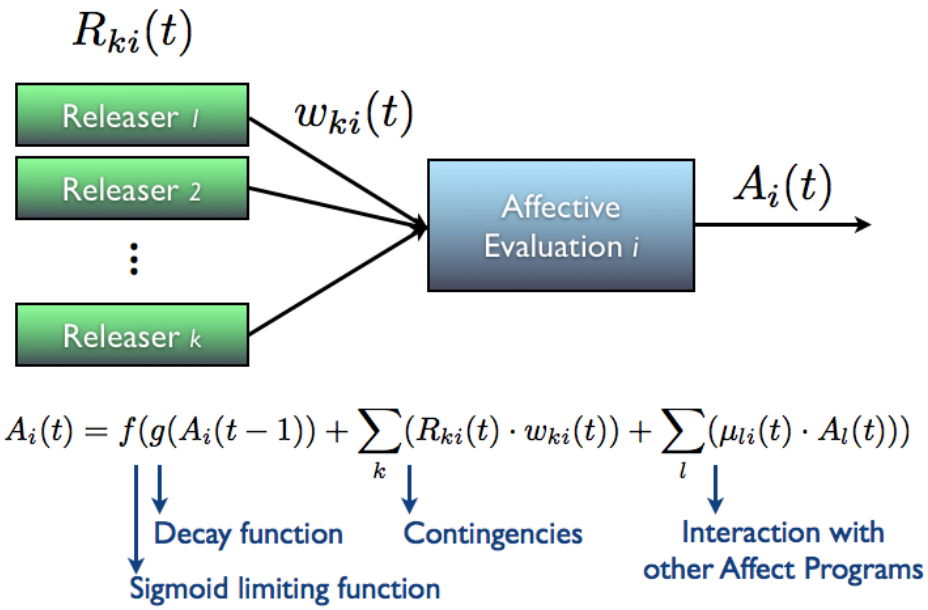

Figure 6-11: Computing the Affective Value of Events. This figure illustrates how the Affective Evaluation Unit computes the affective value of different contingencies detected by the affect program's releasers. This affective value is directly related to the level of activation of the affect program, and thus to the notion of emotional intensity. The specific activity pattern follows the description of Equation 6.6 in the text.

affect program's level of activity would be directly related to the affective value assigned to this event by its Affective Evaluation Unit. Likewise, the detection of the predator would be evaluated by the *Fear* affect program's Affective Evaluation processing, which would assign an affective value that would be appropriate to the level of the threat.

The Affective Evaluation and Control Unit component of Affect Programs bears 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 affective value to different stimuli, the activation of affective responses, and affective learning (Damasio, 1994; LeDoux, 1996; Panksepp, 1995).

6.9 Behaviors: Responding to Contingencies

Up until now we have described the input components of the Affect Program abstraction, and its evaluative and control component. In this section we describe their output components: Behaviors.

From an ethological perspective, Behaviors correspond to the set of fixed responses or behavioral repertoire associated to a particular species. These behaviors are common to all members of the species and are as characteristic of the species as their structural features would be. Lorenz (1973) and Tinbergen (1951) called these patterns of behavior *Fixed Action Patterns* (FAP), which, once activated, are performed in a stereotyped way unaffected by external stimuli (e.g., a frog's prey-catching tongue flick is performed in the same way whether or not the prey is caught).

Integrating these ideas to an affective perspective, Behaviors in the Cathexis framework are implemented as the highly stereotyped and precise responses that are associated with a particular affect program, such as the freezing behavior produced by fear or a courting behavior regulated by the sexual affect program described in Section 4.2.3. It should be noted, however, that Behaviors do not only correspond to overt responses like those mentioned previously, but may also involve internal responses such as activity in the sympathetic and parasympathetic divisions of the ANS, increased reflexes, or the regulation of stress responses. Figure 6-1, for instance, illustrated some of the responses known to be associated with the activity of the brain systems mediating fear. Clearly, in the case of a robot many of these responses are not yet possible[8]. However, the Cathexis framework does allow for the modeling of such type of responses, even if they only exist as simulations[9]. For instance, the regulatory

[8] At least not at the present time, since we have not endowed our robotic systems with real regulatory (and related) mechanisms that would be analogous to systems such as the reproductive, immune, or autonomic nervous system. It may be the case, however, that some of these systems might actually be necessary in order to replicate and better understand living creatures.

[9] In fact, from an affect perspective, it is the author's opinion that not contemplating these

mechanisms that model physiological needs rely on this type of simulated responses to modify the state of the controlled internal variables.

As in many other architectures (Brooks, 1986; Arkin, 1990; Maes, 1991; Blumberg, 1994), the Behavior construct corresponds to actions that represent goal-directed activities an agent can perform given a particular situation. Unlike these architectures, however, Behaviors in this framework depend more heavily on affect to provide a motivational context that determines their relevance and arbitrates possible conflicts with other responses.

Through its associated Releasers, a Behavior obtains the necessary information supporting the low-level motor response that is to be issued when the response becomes active. Although we will defer discussion of this distinction to the following chapter, it should be noted that Behaviors can either be Consummatory or Preparatory, which would correspond to stimulus-specific and affectively-general responses, respectively. Thus, the selection and execution of a Behavior such as "chewing", would depend on sensory-specific information about food (e.g, is food present in the mouth of the organism, what are its physical properties?), whereas a Behavior such as "approaching" the food would merely depend on the general affective information about the food (i.e., is it a positive or a negative stimulus).

Going one last time to our gorillas scenario, the different gorillas (and you as a primatologist), had a variety of responses available as part of their behavior repertoire. Depending on the specific affective evaluation, the "flight" or "escape" behavior became active in the case of the two young gorillas, whereas the "fight" behavior was the selected response for the adult Silverback. The selection of one response over another depends directly on the information provided by the affect program's Releasers

issues (even in robotic systems) is a gross oversight of most existing models of emotions, especially since many of the elicitors and responses involved in affective processing are related to such type of mechanisms.

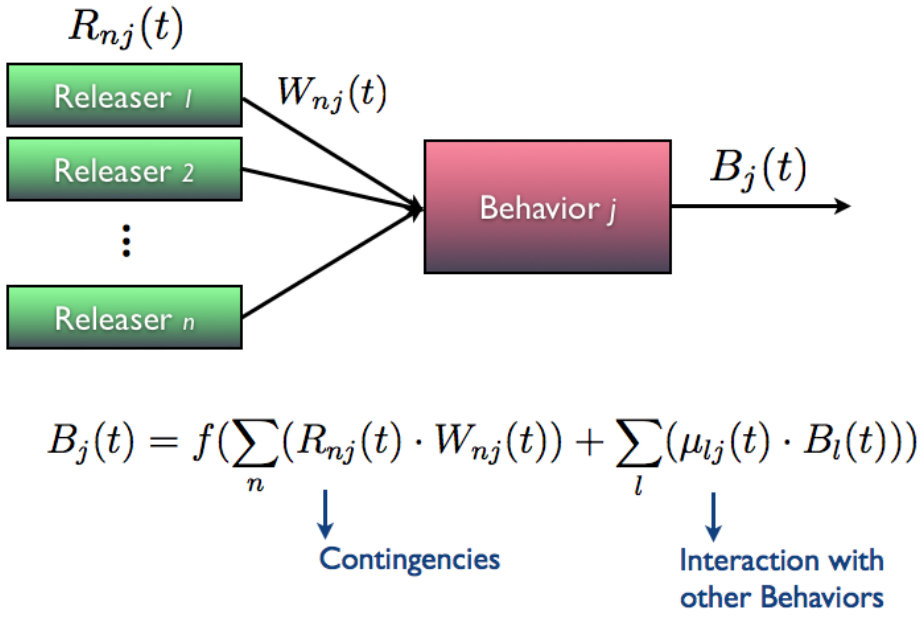

Figure 6-12: Computing the Value of Behaviors. This figure illustrates how the value of Behaviors are computed in the Cathexis framework. The value of the different Releasers is combined with the inhibitory or excitatory input from other Behaviors to compute the value of each Behavior unit. This value will be used in an arbitration mechanism to select the appropriate Behavior in response to any particular contingency. The specific activity pattern follows the description of Equation 6.7 in the text.

and the affective value given to the detected contingencies. Considered together, this constitutes the value of a Behavior as indicated next.

The activation value of Behaviors follows the general pattern of basic computational units described in Section 6.4, but extends it to incorporate functionality related to the possible interactions between different Fixed Responses as described in Equation 6.7

$$B_j(t) = f(\sum_n (R_{nj}(t) \cdot W_{nj}(t)) + \sum_l (\mu_{lj}(t) \cdot B_l(t))) \tag{6.7}$$

Where $B_j(t)$ is the value of Behavior j at time t; $R_{nj}(t)$ is the value of Releaser n

and $W_{nj}(t)$ is its associated weight at time t, where n ranges over the Releasers for Behavior j; $\mu_{lj}(t)$ is the strength of the excitatory (positive) or inhibitory (negative) input from Behavior l, where $B_l(t)$ is its activation value at time t.

The value of Behaviors is of special interest as it will provide the basic means to arbitrate and select between different possible responses within an active affect program. This arbitration or action selection mechanism will be the topic of the next chapter.

6.10 Summary

This chapter presented a unified computational framework, named Cathexis, for the study of emotion and affective phenomena in general, based upon the construction of computational models that integrate several concepts and mechanisms that have been traditionally deemed as integral components of intelligent behavior.

This approach is based on the notion of *Affect Programs*, which we defined in Chapter 2 as adaptive biological schemas that have proven useful, throughout our evolutionary past, in helping us deal with life and survival-related fundamental situations. The Cathexis framework provides a computational counterpart to this theoretical construct through the Affect Program abstraction, which consists of a set of filtering mechanisms, named Releasers, an Affective Evaluation Unit that assesses the affective significance of events, and a set of Behaviors that are synchronized and controlled by the Affect Program in response to specific contingencies.

The following chapters will instantiate a variety of these Affect Programs, and through different scenarios, we will illustrate how they serve many different purposes, ranging from providing the motivational context that synchronizes and controls the execution of different behaviors and maintaining their relevance and coherence (Chap-

ter 7), to the regulation of learning processes by which the robot can learn from past emotionally significant contingencies and modify its behavior accordingly (Chapter 8), and finally, to the modulation of attention through different sensory modalities (Chapter 9).

Chapter 7

Affective Behavior

This chapter demonstrates how behavior can be generated and controlled by the activity of the affect programs described in the previous chapter. We will also introduce a scheme for the organization of behaviors that is based upon earlier ideas and observations from psychology which divide action into two different sensorimotor pathways: one that is concerned with distal contingencies and promotes preparatory responses to deal with them, and one that deals with proximal events which require stimulus-specific responses that are generated according to the sensory and affective properties of the stimulus. As we will describe shortly, we believe that these organizational principles that are widely spread in biological organisms can be useful when applied to agent architectures that account for the selection and control of actions.

7.1 Arbitration and Action Selection

The concept of parallel processing described in Figure 6-5 implies that information about all contingencies and ongoing contexts in the robot's world, reaches and activates appropriate Releasers, and in turn, may activate several Affect Programs simultaneously. How is then coherent activity formulated and coordinated in such

situations? This question relates to the issue of action selection, which is a fairly standard research problem in agent architectures. Within the proposed model, arbitration takes place at different levels, following a two-stage activation system as illustrated in Figures 7-1 and 7-2.

First, all Affect Programs compete with each other for the activation of their Behaviors. This is regulated through the levels of activity of each of the Affect Program's Affective Evaluation Units as described in Equation 6.6 (Figure 7-1). Essentially, the process is as follows: For all instances of affect programs, their affective value is computed, which depends on the input of the affect program's releasers as they represent the detection of prototypical contingencies in the agent's world. The affect program with the highest affective value is selected for execution, which simply means that this affect program takes control over the agent's resources, as it is about to select the appropriate set of responses that would deal with the contingencies that activated the affect program in the first place.

Once an affect program has been activated, which means that its level of activity has surpassed an activation threshold, an emotional episode is said to occur, and a set of responses must be now be coordinated and synchronized. This control process corresponds to the second stage of the action selection process (Figure 7-2), and it develops as follows: For all behaviors associated to the active affect program, their value is computed, which ultimately depends on the input of the affect program's releasers that provide sensory-specific information about the contingencies that activated the affect program, and which prove useful in deciding which behaviors and responses to select and execute. The behavior with the highest affective value is selected for execution, which simply means that this behavior takes momentary control over the agent's actuators. This second stage of the action selection process will be repeated as long as the affect program is active, which means that multiple behaviors can become active at various moments during that time period.

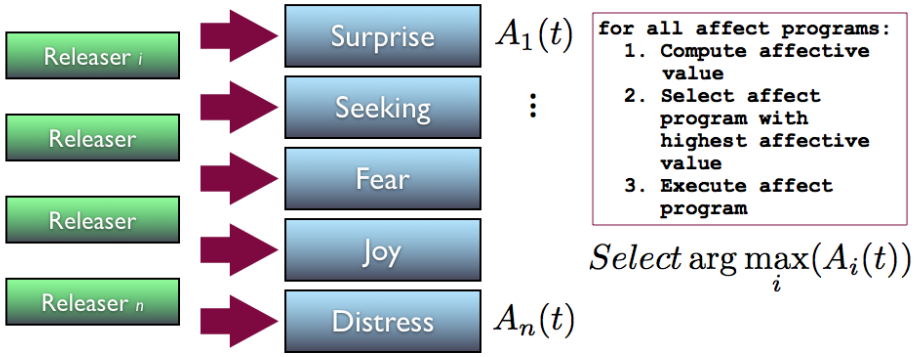

Figure 7-1: Selection of an Affect Program—First Stage in Action Selection Process. This figure illustrates the first stage of the action selection process. For all instances of affect programs, their affective value is computed, which depends on the input of the affect program's releasers as they represent the detection of prototypical contingencies in the agent's world. The affect program with the highest affective value is selected for execution, which simply means that this affect program takes control over the agent's resources, as it is about to select the appropriate set of responses that would deal with the contingencies that activated the affect program in the first place.

When dealing with more complex robotic systems that have many degrees of freedom and richer sensing abilities, it may be possible however for the robot to engage in more than one activity at the same time. Thus a separate level of arbitration takes place at the motor system level as well (not depicted here). Losing Affect Programs may still issue Behaviors, but their execution depends on the specific actuators these behaviors compete for. Thus, non-conflicting Behaviors are allowed to be executed simultaneously as long as the sensory and motor systems they depend on are separable.

7.2 Scenarios for Affective Behavior

To illustrate these ideas, consider the following evaluation scenario as implemented with our simulated robot. We instantiated Marvin's affective repertoire with multiple

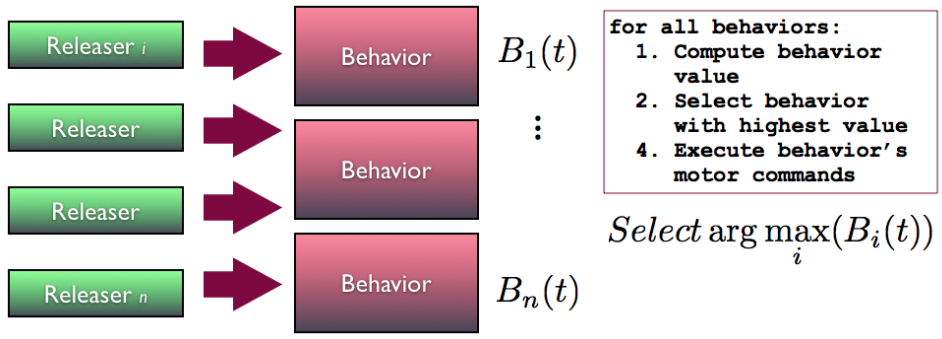

Figure 7-2: Selection of a Behavior—Final Stage in Action Selection Process. This figure illustrates the second stage of the action selection process. For all behaviors associated to the active affect program, their value is computed, which ultimately depends on the input of the affect program's releasers that provide sensory-specific information about the contingencies that activated the affect program, and which prove useful in deciding which behaviors and responses to select and execute. The behavior with the highest affective value is selected for execution, which simply means that this behavior takes momentary control over the agent's actuators. This second stage of the action selection process will be repeated as long as the affect program is active, which means that multiple behaviors can become active at various moments during that time period

instances of the affect programs described in earlier chapters. In particular, we created instances for the *Surprise*, *Seeking*, *Fear*, *Joy*, and *Distress* systems. Let us focus for a moment on a particular instantiation for Marvin's *Fear* affect program.

Figure 7-3 illustrates an example of a *Fear* affect program instantiated in our simulated robot. This affect program's main purpose was to detect dangerous contingencies as represented by two main events: the presence of Marvin's predator (i.e. the evil blue-pool pony) and the detection of dark environments. To this end, we implemented a set of releasers which detected these two different contingencies and provided sensory-specific information related to the distance or range of the robot with respect to these threatening stimuli. In a similar fashion, we implemented three main behaviors that represented the set of responses necessary to deal with these threats: a *Flight* response, a *Fright* response and a *Fight* response. The so-called three "Fs"

Figure 7-3: An Instance of the *Fear* Affect Program. This figure illustrates an instance of the *Fear* affect program implemented in the simulated robot described in Section 5.3.

represented different possible responses that would deal with the prototypical situation of facing a dangerous event, based upon the sensory-specific information of this event.

In this particular evaluation scenario, Marvin's *Fear* affect program would select the flight behavior if the robot's distance to the predator, whenever detected by its releasers, would be less or equal than 3 meters, but more than 2 meters. When such contingency was detected, and the *Fear* affect program was the active program (i.e., its affective value was higher than that of all other affect programs, as described in the first stage of the action selection process), it would control the execution of the flight response as illustrated in Figure 7-4. The specific response coded by the flight behavior consisted on a startled expression and a set of commands to the robot's actuators controlling its linear velocity so that the distance between the robot and the predator would be increased until the predator was no longer in sight.

Similarly, Marvin's *Fear* affect program would select the *fright* behavior if the robot's distance to the predator, whenever detected by its releasers, would be less or equal than 2 meters, but more than 1 meter. At such range, a flight response might not be useful, as the predator is almost within reach of the robot and an attempt to

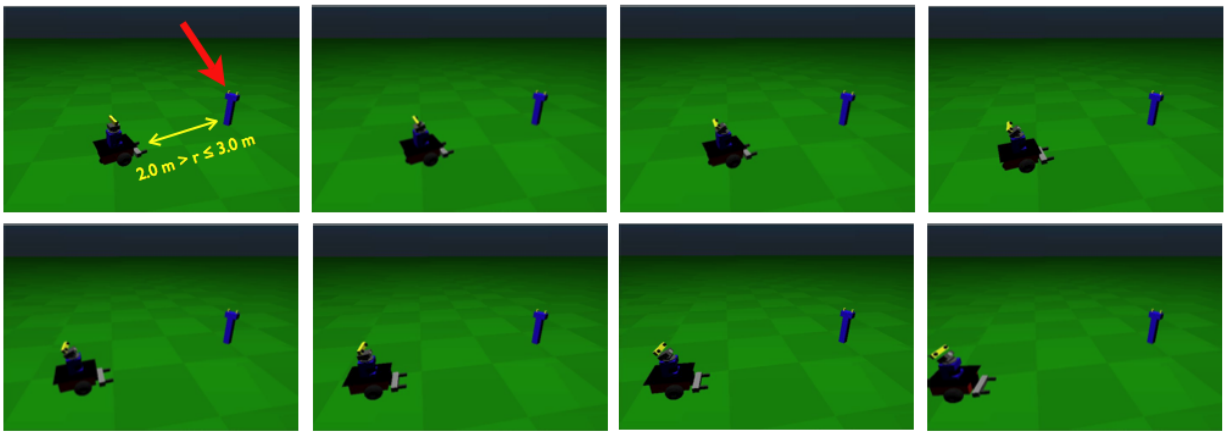

Figure 7-4: Flight Response. This sequence of frames illustrates how the Flight response becomes the selected response, as part of the second stage of action selection. Activity of releasers indicate the "predator" is present (red arrow) and the distance r between it and the robot is less or equal than 3 meters, which is the distance that would trigger such response.

escape might not be successful. In such contingencies, a "play dead" response might be of better use as it does not draw the attention of the predator until a possibility for escaping arises. When such a contingency was detected, and the *Fear* affect program was the active program, it would control the execution of the fright response as illustrated in Figure 7-5. The specific response coded by the fright behavior consisted on a set of commands that would stop all of the robot's movements and the robots head would be tilted down, "playing dead".

Finally, Marvin's *Fear* affect program would select the *fight* behavior if the predator would be detected withing 1 meter of range of the robot. At such distance, a flight response is impossible, and the fright response would not work either given that the distance is so short that the predator's attention would most likely be drawn by the robot. In such event, the only other option is to fight the predator. When this kind of event was detected, and the *Fear* affect program was the active program, it would control the execution of the fight response as illustrated in Figure 7-6. The specific

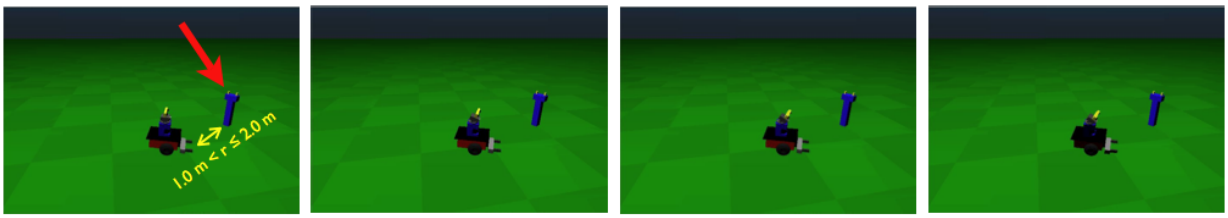

Figure 7-5: Fright Response. This sequence of frames illustrates how the Fright response becomes the selected response, as part of the second stage of action selection. Activity of releasers indicate the "predator" is present (red arrow) and the distance r between it and the robot is less or equal than 2 meters, but greater than 1 meter which is the distance that would trigger such response.

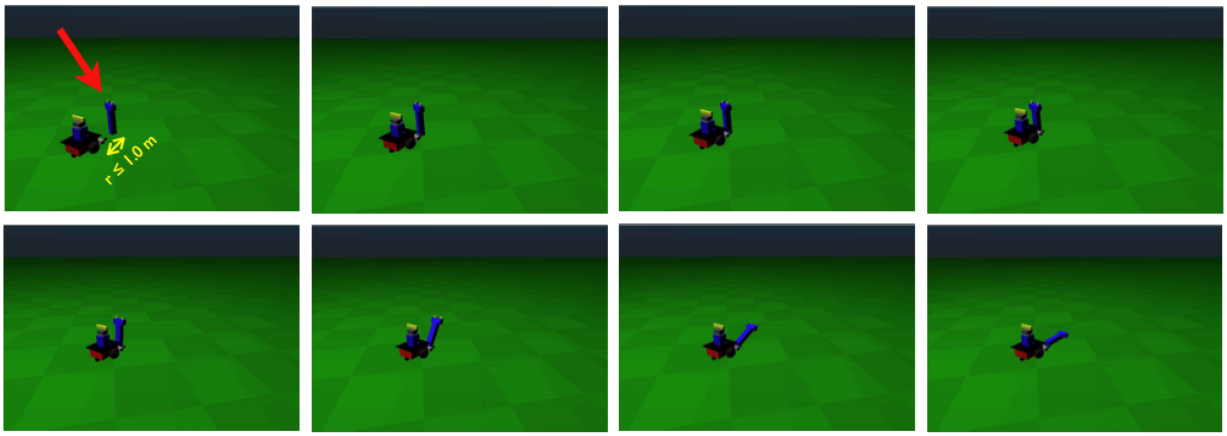

Figure 7-6: Fight Response. This sequence of frames illustrates how the Fight response becomes the selected response, as part of the second stage of action selection. Activity of releasers indicate the "predator" is present (red arrow) and the distance r between it and the robot is less or equal than 1 meter, which is the distance that would trigger such response.

response coded by the fight behavior consisted on a set of commands that would control the robots linear velocity so that it would attempt to run over the predator, as illustrated in the figure.

7.3 Anticipatory and Consummatory Behaviors

Our evaluation of the affective behavior demonstrated by the robot Marvin, described in the previous section, suggested a more fundamental organizational principle for action. Considering the notion of distal and proximal stimuli, and the different routes for behavior activation, we delved deeper into these issues and found inspiration from earlier observations in psychology and neuroscience as to how behavior might be organized in a more effective and modular manner.

Sir Charles Sherrington (1906), a pioneer in research on neural mechanisms of spinal reflexes, provided at the beginning of last century a general conceptual framework for how adaptive behavior is generated. Specifically, he introduced an influential distinction that categorized behavior in terms of *preparatory* (also referred to as *anticipatory* or *appetitive*) and *consummatory* responses.

Preparatory behavior is flexible and thus it can vary considerably. It usually implies the exploration and appraisal of the environment, which can later be combined with previous experience. According to Sherrington, these responses were triggered by visual, auditory, and olfactory perceptual systems–what he called "distance-receptors"–which provide information at a spatiotemporal distance that gives organisms the opportunity to prepare for environmental contingencies. An essential feature of preparatory behavior is that it corresponds to the first part of a sequence of multiple behaviors that together promote survival by bringing an organism into proximal contact with a goal or incentive.

Consummatory behaviors, on the other hand, correspond to more rigid, fixed action patterns that serve to fulfill a particular goal. A consummatory response is essentially the end of the sequence of motivated behavior enunciated above and is usually triggered by tactile and taste perceptual systems–what Sherrington called "proximal-receptors"–that inform organisms about what should be the appropriate immediate

Table 7.1: Comparison of *Preparatory* and *Consummatory* Sensorimotor Pathways.

Sensorimotor Pathways	
Preparatory	***Consummatory***
Composed of flexible-responses that are non-specific to a particular stimulus (e.g., approach)	Composed of fixed responses that are specific to the different stimuli encountered in the environment (e.g., chewing, gnawing)
Triggered by distance-receptors (e.g., visual and auditory systems)	Triggered by proximal-receptors (e.g., gustatory, olfactory and tactile systems)
Provide information at spatiotemporal distance, which is used to determine the general preparatory response	Provide sensory-specific information about the kind of stimulus present, and thus the kind of response that is most appropriate for such stimulus
Allow organisms to prepare for soon-to-be proximal contingencies	Allow organisms to fulfill goals (e.g., enhance welfare)

responses in order to enhance their welfare. Usually, these behaviors result in consummatory reactions, such as the ingestion of foodstuff (nutrients) or withdrawal from life-threatening situations, as depicted in our previous scenario with Marvin and its main predator.

Table 7.1 relates both preparatory and consummatory behavior, according to their main features.

7.4 Approach and Avoidance

Related to Sherrington's consummatory and anticipatory ideas, the concepts of approach and withdrawal responses also acquired significance at the time. Schneirla (1959), in his biphasic "A-W Theory", was among the first to argue for the existence of two fundamental behavioral processes: one for approach and another for withdrawal that might be the end result of completely different systems. He stated:

> *"Much evidence shows that in all animals, the species-typical pattern of behavior is based upon biphasic, functionally opposed mechanisms insuring approach or withdrawal reactions according to whether stimuli of low or high intensity, respectively, are in effect. This is an oversimplified statement; however, in general, what we shall term the A-type of mechanisms, underlying approach, favors adjustments such as food-getting, shelter-getting, and mating; the W-type, underlying withdrawal, favors adjustments such as defense, huddling, flight, and other protective reactions. Also, through evolution, higher psychological levels have arisen in which through ontogeny such mechanisms can produce new and qualitatively advanced types of adjustment to environmental conditions"* – Schneirla (1959, p. 4)

It seems reasonable then to believe, that there are in fact anticipatory and consummatory components for both approach and withdrawal systems. Konorski (1967) further classified the basic activities of organisms into four different categories: "preservative preparatory", "preservative consummatory", "protective preparatory", and "protective consummatory".

The term "preservative" corresponds to approach, while the term "protective" corresponds to withdrawal.

Based upon these ideas, we modified our affect program abstraction so that it would incorporate this basic action organizational principle as part of its overall architecture, as illustrated in the Figure 7-7 and Figure 7-8.

7.5 Evaluation of Sensorimotor Pathways

To test these ideas, we implemented a variety of affect programs in the robot Coco, which served as evaluation testbeds for the notions of preparatory and consumma-

Figure 7-7: Pathway for Consummatory Behaviors. This figure illustrates the main organizational principle for action in the proposed framework as it relates to the notion of a separate sensorimotor pathway for consummatory behaviors that help bring the organism to be in proximal contact with goal objects, thus promoting survival.

Figure 7-8: Pathway for Preparatory Behaviors. This figure illustrates the main organizational principle for action in the proposed framework as it relates to the notion of a separate sensorimotor pathway for preparatory or anticipatory behaviors that help the organism prepare to deal with environmental contingencies.

tory pathways. In particular, we implemented the *Seeking* affect program, as the main mechanism to explore the world, approaching objects of interest, and more importantly promoting and coordinating responses that would correspond to solutions to impending internal needs (e.g., recharging the battery when its levels are low). The complete description of the *Seeking* affect program will be deferred until the next chapter. For now, let us focus on how different sensorimotor pathways were implemented.

7.5.1 Coco's Distal and Proximal Releasers

In Section 6.6, we presented the idea of releasers as the main computational processes that filtered information resulting from perceptual systems. Releasers thus act as a mechanism that provides an assessment of perceptual information in order to determine whether a particular stimulus is of affective significance and therefore should motivate behavior.

We extended this mechanism to include the notion of perceptual distance, which, although related to the classic perception issue of distal versus proximal stimuli (Heider, [1930]/1959; Koffka, 1935; Brunswik, 1956; Gibson, 1979), it is not as concerned with the differentiation between the "actual" object (distal stimulus) and the images it produces on the receptors (proximal stimulus), as it is with determining the relative distance of a stimulus from the robot's perspective. In other words, proximal and distal releasers are implemented as the means to identify percepts that are either close to (proximal) or at certain reachable distance (distal) from the robot.

There are many different kinds of releasers defined for Coco that were built as part of our implementation of the *Seeking* affect program. By exploiting the regularities of the environment in which Coco was situated (e.g., the circular shape, color, and actual size of balls) we incorporated a straightforward measure of distance in the robot's visual perceptual system. This gave Coco the ability to detect whether balls that were conceived as appetitive stimuli, and thus were affectively significant, were within reachable distance.

The following scenario, as depicted in Figure 7-9 illustrates these concepts further. In this scenario, Coco was situated in our lab environment and the object of interest (i.e., the purple ball) was placed within the robot's visual perceptual field. As soon as an appetitive stimulus is detected by the robot's distal releasers, the information related to the goal's spatiotemporal distance is sent via the preparatory pathway, where

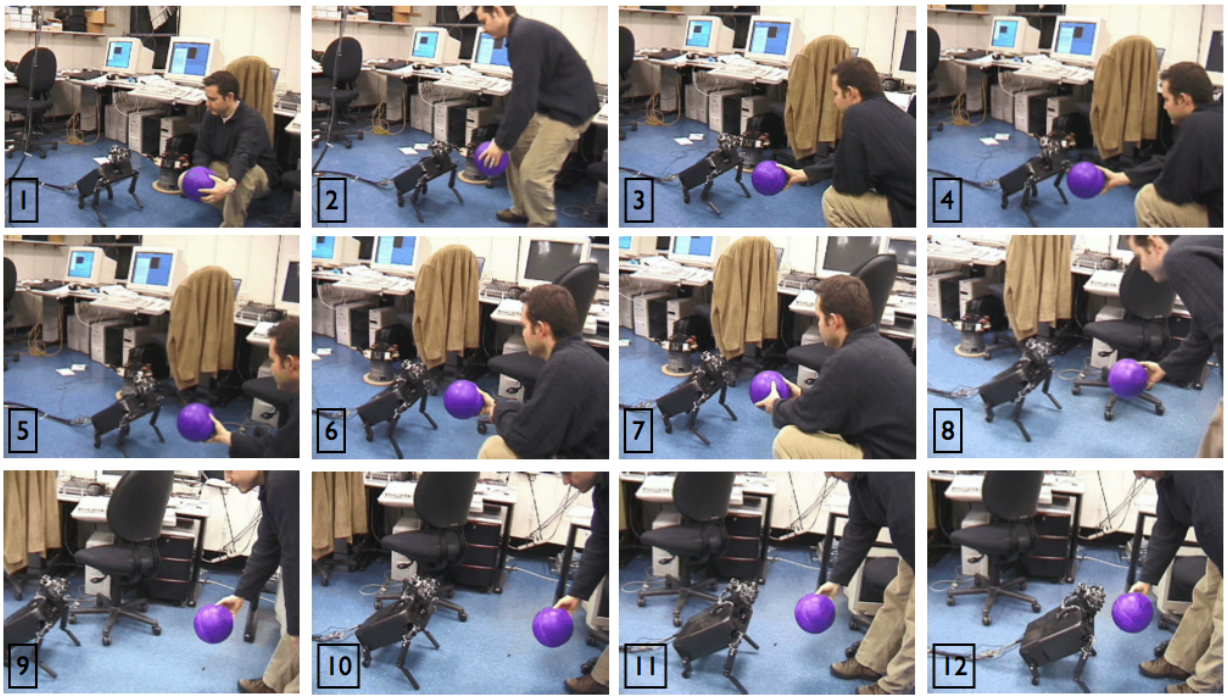

Figure 7-9: Distal and Proximal Releasers. This sequence of frames illustrates how the distal and proximal releasers that are used in Seeking Affect Program work. Distal releasers detect affectively significant stimuli at a spatiotemporal distance and send that information through the preparatory pathway, as described in the text, triggering an approach behavior until the significant stimulus is within certain pre-determined range, in which case the approach behavior is stopped.

ultimately an approach behavior, mediated by the *Seeking* system ensues. Frames 1 thru 4 illustrate this behavior. As soon as the object of interest is within certain specified range, the distal releasers stop responding to their preferred contingency, and thus the approach behavior also ends, as illustrated in frames 5 thru 7. Finally, separating the stimulus from the robot, illustrated in frames 8 thru 11, triggers the same responses all over again, until the stimulus is within reach (frame 12), when the scenario ends.

7.5.2 Coco's Sensorimotor Pathways

Distal and proximal releasers trigger very different affective responses mediated by the *Seeking* affect program. Through the preparatory pathway, distal releasers signal the presence of an "interesting" stimulus in the robot's world. This stimulus might be of interest to the robot because it has an affective significance, whether this has been pre-programmed (detected through a *Natural Releaser*), or learned as we will describe later on (and thus detected by a *Learned Releaser*), or because it is a novel stimulus and thus is worthy of investigation. In any case, distal releasers detect this kind of stimuli and trigger flexible approach responses, in the case of an appetitive stimulus, or avoidance responses in the case of an aversive one.

Conversely, through the consummatory pathway, proximal releasers signal the detection of stimuli that are in close contact with the robot, and thus can be be potentially manipulated or consumed, which would activate a consummatory behavior that is specific to the kind of stimulus being detected.

To complete the scenario described above, a combination of both preparatory and consummatory behaviors needed to be demonstrated. We thus implemented an *Interest* behavior as part of the *Seeking* affect program, which would correspond to a Consummatory behavior for Coco. Figure 7-10 illustrates this scenario. As with the previous case, the robot was placed in our lab environment and the purple ball was used as an appetitive stimulus. Frames 1 thru 8 illustrate how the preparatory behavior (e.g., approach to target) ensues, once an affectively significant stimulus is detected by the robot's distal releasers. Frames 9 thru 12 illustrate how the consummatory behavior of *Interest*, which simply executed side-to-side movements of the robot's head, ensued once proximal releasers detected the object of interest within close reach. Notice how as soon as the proximal releasers are active, the flow of information through the preparatory pathway changes and the preparatory behaviors

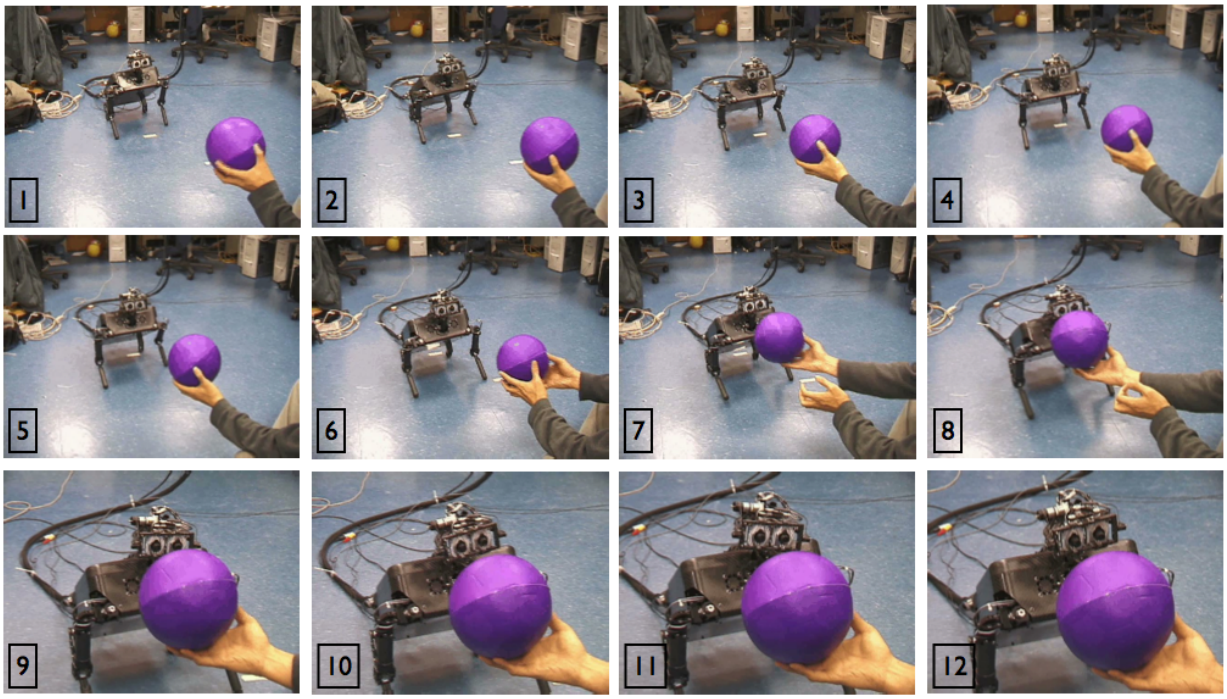

Figure 7-10: Preparatory and Consummatory Responses. This sequence of frames illustrates how the different sensorimotor pathways work, by activating first a preparatory behavior (approach) when a distal releaser has detected an affectively significant stimulus, and second, activating a consummatory behavior (interest—moving the head side to side) once the stimulus is proximal to the robot, as detected by a proximal releaser.

are inhibited.

The separation of anticipatory and consummatory pathways into two different sensorimotor routes, as demonstrated in these scenarios, is advantageous for several reasons, as it:

- Modularizes behavior
- Promotes learning
- Facilitates chaining responses, from preparatory ones to consummatory ones that are essential for the organism in its efforts to attain goals

- Promotes the reuse of preparatory responses (i.e., approach can be targeted at a recharging station or at a person)

- Relates directly to the notions of *appetitive* and *aversive* events (Craig, 1918), which would connect to approach and avoidance responses.

7.6 Affect Programs in the Robot Yuppy

To end this chapter, let us briefly review some of the full instantiations created for the robot Yuppy, which corresponded to our first attempts at synthesizing affective behavior. It should be noted that these instantiations do not necessarily follow all the principles enunciated above in an exact manner, but rather served as inspiration for their design.

Yuppy was used as the first platform to study affective processing on a real-time system that used real sensory and motor systems. This posed some interesting questions regarding the extension of a model of affect to be used in a robotic system, as most previous models of affect had been primarily used to drive synthetic characters.

As part of this instantiation of the framework, we implemented a set of basic Affect Programs for Yuppy, which included the following: *Anger, Fear, Distress/Sorrow, Joy/Happiness, Disgust, and Surprise.* This set of affect programs roughly corresponded to those proposed by several theorists as the basic emotions (Ekman, 1992; Johnson-Laird & Oatley, 1992; Panksepp, 1998), only in our case these were extended to the notion of basic affect programs.

The Surprise Affect Program

The circuits implemented by the Surprise Affect Program dealt with novelty, anticipatory expectancy, and other issues that have been considered essential components of

Figure 7-11: Illustration of an instance of the Surprise Affect Program in the robot Yuppy. Three main pathways were coordinated by Yuppy's Surprise Affective Evaluation mechanism. The first pathway mediates startle responses due to the occurrence of high intensity sounds. The second pathway mediates Orienting Responses (OR) of the robot's head to attend to and "foveate" the object of interest. The third pathway mediates full ORs that align the robot's body and head so that it faces the object of interest, perhaps to initiate an approach response toward it.

a general attentional system, including orienting to sensory stimuli, executive functions such as the detection of target events, and maintenance of a general "alert" state (Posner & Badgaiyan, 1998).

Figure 7-11 illustrates the Surprise Affect Program's set of Releasers (left side of figure) and associated Fixed Responses (right side of figure). There are three main pathways that are executed and coordinated by the Surprise Affective Evaluation Unit whenever its Releasers are present. The first pathway mediates Startle Responses—also referred to as the *Acoustic Startle Reflex*—due to the occurrence of high intensity sounds. The second pathway mediates Orienting Responses (OR) of the robot's head to attend to and "foveate" the object of interest—usually a pink bone as described in Section 5.1. The third pathway mediates full ORs that align the robot's body and head so that it fully faces the object of interest, allowing its sensory system to collect more data regarding the stimulus and if appropriate, perhaps initiate an approach response toward it.

Figure 7-12: Illustration of an instance of the Fear Affect Program in the robot Yuppy. Three pathways were coordinated by Yuppy's Fear Affective Evaluation mechanism. The first pathway mediates cowering responses that were evoked when the presence of Yuppy's predator (the blue pool-pony) was detected. The second pathway involves the regulation of expressive behavior as it pertains to fear. Any threatening stimulus (i.e., blue pool-pony or darkness) would elicit a fearful expression to different extents. The final pathway regulates navigation for the robot according to the level of brightness perceived by the robot's vision systems. If the robot is headed toward a dark region, navigation is controlled so that the robot veers away from that region.

The Fear Affect Program

The Fear Affect Program implemented as part of Yuppy's emotional repertoire dealt mainly with "dangerous" contingencies for the robot, which ranged from situations in which its sensory systems would not work properly (e.g., dark environments for its vision system), to the detection of "predators" (e.g., blue pool-ponies as described in Section 5.1).

Figure 7-12 illustrates the three main pathways coordinated by Yuppy's Fear Evaluation Unit. The first pathway regulates a Cowering Response that is elicited when the blue pool-pony is detected. The second pathway involves the regulation of expressive behavior as it pertains to fear. Any threatening stimulus (i.e., blue pool-pony or a dark environment) would elicit a fearful expression to different extents. Finally, the third pathway regulates navigation for the robot according to the level of brightness perceived by the its vision systems. If the robot is headed toward a dark region, nav-

Figure 7-13: Illustration of an instance of the Joy Affect Program in the robot Yuppy. Three different pathways are coordinated by the Joy Affective Evaluation Unit. The first pathway regulates interactions with people through a Greet-Person Response. The second pathway regulates joyful expressions given the presence of any or all of the aforementioned stimuli. Finally, the third pathway mediates approach responses toward stimuli that were deemed "interesting" (see text).

igation is controlled so that the robot veers away from that region. This makes sense from an adaptive perspective as the robot's primary sensors (i.e., its visual sensors) do not work very well under such environments.

The Joy/Happiness Affect Program

The Joy Affect Program orchestrates affective processing of appetitive stimuli, which includes the detection of "friends" (i.e., people and pink pool-ponies), "food" (i.e., pink bones), and toys (i.e., green cylinders or big yellow balls).

Figure 7-13 shows the three different pathways that were coordinated by the Joy Affective Evaluation Unit. The first pathway regulates interactions with people. Once the presence of a person has been detected, a greeting response is initiated. The second pathway regulates joyful expressions given the presence of any or all of the aforementioned stimuli. The third pathway mediates approach responses toward stimuli that have been deemed "interesting". The measure of how interesting each stimulus is depends on the confidence—indicated by the level of activation of its Releaser—that the stimulus corresponds to any of the objects the robot has been pre-programmed

to consider as appetitive stimuli[1].

The Distress/Sorrow Affect Program

The Distress Affect Program is occupied with responding to certain kinds of stressful stimuli in the robot's environment. For instance, all regulatory mechanisms release distress when their error signal (drive) goes above certain threshold. Responses to this kind of stimuli typically lead the robot to pursue the needed resources in its environment (e.g., seeking people to interact with, or pink bones and yellow balls to play with).

Yuppy's regulatory mechanisms included:

- Recharging Regulation: Monitors the robot's battery level (composed of two different batteries, one on the synchrodrive base and one on the body proper).
- Imitation: A particular instance of a social drive that "urged" Yuppy to promote interaction with people by imitating sounds.
- Play: A different instance of a general social drive that urged Yuppy to interact with people by triggering Fixed Responses that appeared to be dancing moves.
- Fatigue: Unlike the previous two mechanisms, Fatigue regulated Yuppy's activities by shutting-out the external world instead of promoting interactions with it.

Figure 7-14 shows the three different pathways that make up the Distress Affect Program. The first pathway regulates seeking behaviors. Having no stimuli in sight will increase the activity of the Distress Affective Evaluation and Control Unit, which will promote the execution of a Look-Around Response in which the robot scans its

[1]Pre-programming the robot to "like" or "dislike" certain stimuli is akin to our innate predispositions to different stimuli in the world.

Figure 7-14: Illustration of an instance of the Distress Affect Program in the robot Yuppy. Three different pathways are coordinated by the Distress Affective Evaluation and Control Unit. The first pathway regulates seeking behaviors in search of objects of interest. The second pathway regulates sorrowful expressions according to the activity level of the Distress Affect Program. Finally, the third pathway mediates control responses that help the robot maintain balance of its regulatory mechanisms.

environment seeking out for any object of interest. The second pathway regulates sorrowful expressions according to the activity level of this affect program. Finally, the third pathway mediates control responses that help the robot maintain balance of its regulatory mechanisms. Some of these responses include overt behaviors, such as the Imitate and Dance responses, or internal actions such as the simulated "resting" that balanced its fatigue drive.

7.7 Affective Phenomena

The robotic instantiations of the framework provided us with the opportunity to assess the kinds of affective phenomena that could be synthesized and thus the feasibility of the framework as a model for affect.

7.7.1 Fast Primary Emotions DP 1.2

Section 6.1.1 discussed one kind of affective phenomena that has been shown to be pan-cultural and for which homologues exist in other species. This type of affect

has been defined by many as the *primary emotions*. They are primary in the sense that these are innate, pre-organized, and more primitive mechanisms that process affective information and from which other affective phenomena might be derived. This is precisely the approach we followed in this work, by implementing the affect program abstraction. We believe that a set of more primitive mechanisms exists, such as those reviewed in Chapter 4, which deal with innately prepared stimuli in very specific, stereotyped ways. The affective processing of these stimuli and the subsequent control and coordination of ensuing responses is thus the primary function of Affect Programs, as described throughout this thesis.

Given the instantiation of the framework described in the scenarios above, fast primary affective processing is possible, as implemented with the instances of affect programs described above.

7.7.2 Emergent Emotions and Emotional Behavior DP 1.1

The following anecdotal account illustrates a scenario in which emergent emotions and emotional behaviors were produced with the Cathexis framework. In this scenario, two different people interact with the robot Yuppy, each of them holding a pink bone, which, as it was mentioned above, elicits activity in the primary Joy Affect Program. Inadvertently, one of the participants involved in this scenario was wearing a a red sweater. Unbeknown to any of the participants, the color of the sweater was detected by the Pink Bone Releaser indicating the presence of the bone, albeit with a lower confidence[2]. This resulted in the activation of the Look-At-Stimulus Response targeted toward one pink bone (the one being held by the participant with the red sweater) more than the other. Thus, the robot would focus on the pink

[2]In normal people, vision systems are robust enough to distinguish among the color spectrum. However, the artificial vision system used with Yuppy was not always able to distinguish between certain shades of red and pink, thus generating false positives in the detection of objects that would trigger affective responses.

bone that this person held, and much like the behaviors observed with Braitenberg's vehicles, from an observer's point of view, this appeared to be an emergent emotion of the robot, "liking" or "preferring" one pink bone, better than the other[3]. After analyzing the situation, to us, designers of the system, it was readily apparent what was really happening: The combination of the red sweater and the pink bone acted as a super stimulus for the robot, much like those used in ethological experiments (Tinbergen, 1951), thus resulting in the observed behavior. Notwithstanding this realization, it is interesting to observe that such emergent emotional reactions can take place in our robotic systems, considering there is plenty of evidence showing that similar kind of behaviors can be observed in real animals, as it is the case with fixed action patterns and their activation through super stimuli and "fake" sign stimuli (Tinbergen, 1951; Lorenz, 1973).

7.7.3 Emotion Blends

As it has been suggested before, all affective phenomena are believed to be derived from the primary emotions. One common suggestion made by many adherents to the affect programs theory is that emotions which do not fit the model of primary emotions are simply blends of activity in one or more affect programs (Izard, 1977; Plutchik, 1994). For instance, according to Plutchik (1994), fear and surprise would generate alarm, whereas joy and fear would produce guilt.

It seems clear from the work of Izard, Ekman, Plutchik, and others that there can be blends of basic affects. In fact, Cathexis does support the synthesis of such blends when two or more of the basic affect programs are activated simultaneously at levels below their activation thresholds. When activity surpasses the activation

[3]When naive observers took part of this experimental scenario it was common to hear expressions such as "It likes you", referring to the fact that the robot would seem to prefer one person (rather than the bone) instead of the other.

thresholds in any of these systems, the winning Affect Program will tend to inhibit other active Affect Programs.Thus, although it is possible to synthesize affective blends in the model, it is less apparent, and rather unlikely, however, that these mixtures can account for the whole domain of affective phenomena, which includes moods, temperament, and other kinds of affective processing that are important to consider in a model.

7.7.4 Other Affective Phenomena

Many researchers distinguish emotions from moods. The differences among these affective phenomena are not completely clear, however, nor the specific function of moods with respect to emotions or affect in general. A possible interpretation for moods is that they correspond to low level activity of the same affective processing systems responsible for coordinating emotional responses, but which, given their lower level of activity, do not issue specific stereotypical responses (i.e., overt motor behaviors that deal with a specific contingency), but rather modulate all cognition, including further appraisal of affective stimuli. Thus, moods may serve the purpose of assessing the propitiousness of the environment for specific action. We follow this approach in Cathexis, and model moods as low levels of arousal of the same Affect Programs. In other words, while high arousal of Affect Programs will surpass activation thresholds and thus will tend to inhibit other Affect Programs, mild arousal may very well allow several Affect Programs to be concurrently active, leading to the chance of a modulation of multiple systems without issuing a specific response. This representation is consistent with the enormous subtleties of human moods, since the possible combinatorial states of the primary Affect Programs (taking into account their overall intensities, time courses of activity, and the interactions within their Releasers) are enormous (Panksepp, 1994). It is also consistent with the common observation that

moods seem to lower the threshold for arousing certain emotions because Affect Programs that are aroused at low levels, as it happens in the representation of moods, are already providing some potential for their full-blown activation. 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 Affect Programs will decay more slowly.

7.8 Summary

This chapter presented several experimental scenarios that illustrate the notion of affect programs as instantiated in our framework. We have shown how behavior can be generated and controlled by the activity of the affect programs and introduced a scheme for the organization of behaviors that is based upon earlier ideas and observations from psychology which divide action into two different sensorimotor pathways: a *preparatory* or anticipatory pathway and a *consummatory* one.

We further showed how these principles prove to be effective in the organization and control of behavior, as mediated through affect programs.

Chapter 8

Affective Learning

> *The sway that the response-reinforcement framework (Spencer, Thorndike, Hull, Skinner) has held on the behavioral sciences for nearly a hundred years is finally ending. The strength of this framework lays in providing concepts and methods for studying the effects of hedonic (reinforcing) stimuli on the repetition of specified responses acquired in instrumental training situations of various kinds. Its weakness lays in the invalidity of its central assumptions, stimulus-response association and response-reinforcement, which could not deal with motor equivalence and flexibility (or "intelligence") in behavior* (Bindra, 1978, p. 41)

All of us can relate to the great power of our emotions. Not only do they govern our impulses and actions, but also direct our attention to the events in the world that could be significant to us, as embodied organisms situated in the different environments we inhabit.

A very important function of affect programs, as repeatedly mentioned throughout this work, consists of determining what objects, events or contingencies are of importance to the survival of the organism. As such, affect programs are great resources for the individual, as they allow it to detect contingencies of importance in the world and prepare the appropriate responses to deal with them. These responses, as we have discussed earlier, might be the stereotypical responses that have proved

useful in our evolutionary past, or they can be whatever other responses we have learned as relevant throughout our interactions with the world, and thus have been incorporated to our behavior repertoires.

8.1 Multiple Stages for Affective Learning

Based on evidence stemming from multiple disciplines that have studied affective phenomena from different standpoints, we propose that affective learning occurs in a sequential set of events that take place when an organism is exposed to signals or cues that predict affectively significant events. The proposed model is depicted in Figure 8-1. This model indicates: (a) the hypothetical psychological constructs that occur at each stage; (b) possible behavioral correlates; and (c) the computational components that are associated with, and support the events in each stage. Although the model suggests a sequential order, the fact that the behavioral components follow this order is not meant to imply that the brain mechanisms underlying these different behaviors and stages also function in a sequential order. In fact, quite the contrary occurs and depending on what is being studied, parallel processing (and even competition) occurs in the underlying brain mechanisms. The main issue that we want to point out with this model is that the formation of associations based on affectively significant events, produces predictable behavioral changes that are associated with multiple learning systems involving a variety of processes.

The model can be described as follows. The first stage shown in Figure 8-1 represents an "attention" (also often called "arousal") stage that involves the response to novel stimuli, most usually associated with "orienting responses" (ORs). In addition to the relational behavior associated with the ORs, novel stimuli also elicit complex autonomic changes (e.g., changes in heart rate and blood pressure, hormonal release into the bloodstream). If the eliciting stimulus is not of affective significance (either

directly or because it has been associated with a stimulus that is), the OR habituates until it is no longer generated. This first stage of affective learning will be described in Chapter 9.

In contrast, the co-ocurrence of otherwise *neutral* stimuli with affectively significant ones, would elicit specific behaviors associated with the next stage. This second stage corresponds to the first step in the development of associative affective learning. In this stage, neutral stimuli acquire affective significance and thus become important to the organism, as they become reliable predictors of events of biological importance. The development of affective significance associations is related to the appearance of learned specific and non-specific responses, as will be discussed in more detail later. Non-specific responses have been referred to as *preparatory*, since they occur regardless of the nature of the learning contingencies and presumably in order to prepare the organism for the specific events that will follow. For instance, predicting the presence of a predator through signals in the environment, might trigger a set of preparatory responses that include accelerating the heart rate, releasing specific hormones such as adreno-cortisol, and sending blood to the limbs, all in preparation for escape. Specific responses, on the other hand, have been referred to as *consummatory*, as they end the preparatory phase of behavior and consist of actions that are specific to the affective event (e.g., escaping once the predator is actually detected). Thus this second stage is perhaps the most important stage in affective learning, as it is in here that stimuli are "coded" with affective value and meaning is ascribed, at its simplest level, to the events occurring in the world. This stage is described in this chapter.

In the third stage, this same kind of meaning is ascribed, but this time to actions. In this stage, flexible responses are learned based on the association of the outcome of an action (in affective terms), and the action itself. This stage comprises a set of highly complex events for which there is yet no complete understanding. However, we do know that the learning of more flexible responses starts to occur and when

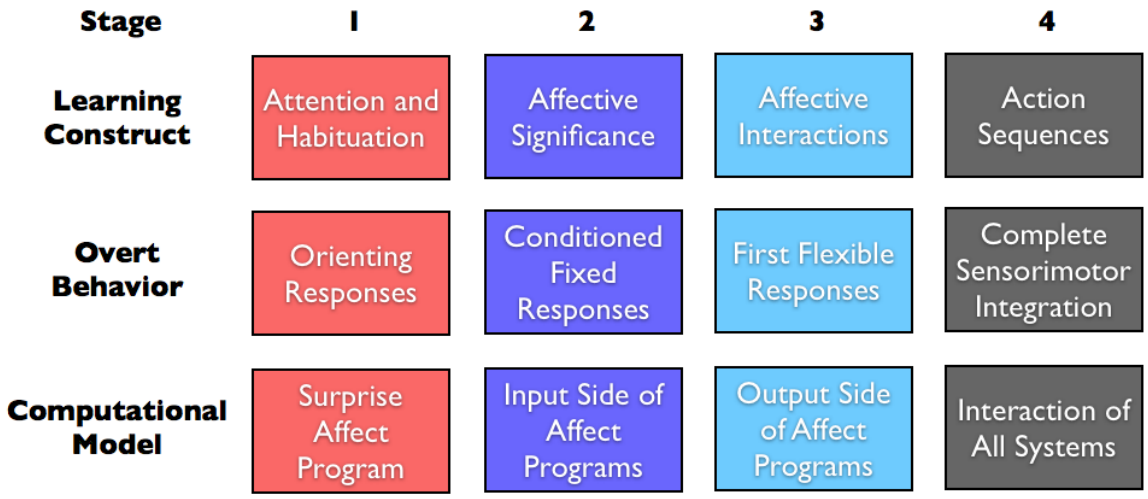

Figure 8-1: A multi-stage model of affective learning.

the events that led to this learning are repeated in a predictable manner, learning reaches asymptotic levels and the production of these responses become habitual. Some examples of these kinds of interactions are presented in Chapter 9.

Finally, in the fourth stage, these habitual responses are organized into behavioral "chunks" composed of sequences of behaviors that represent the highest form of sensorimotor integration.

8.2 Attributing Affective Significance

As mentioned earlier, in their most basic form, affect programs are closed and appear to offer limited response flexibility. Their releasers are pre-wired and their responses are often short and stereotyped. Like many other traits, however, affect programs are deeply developmentally ingrained.

A major aspect of this thesis involves the extension of the affect program abstraction so that fixed affective responses can be gradually refined and eventually

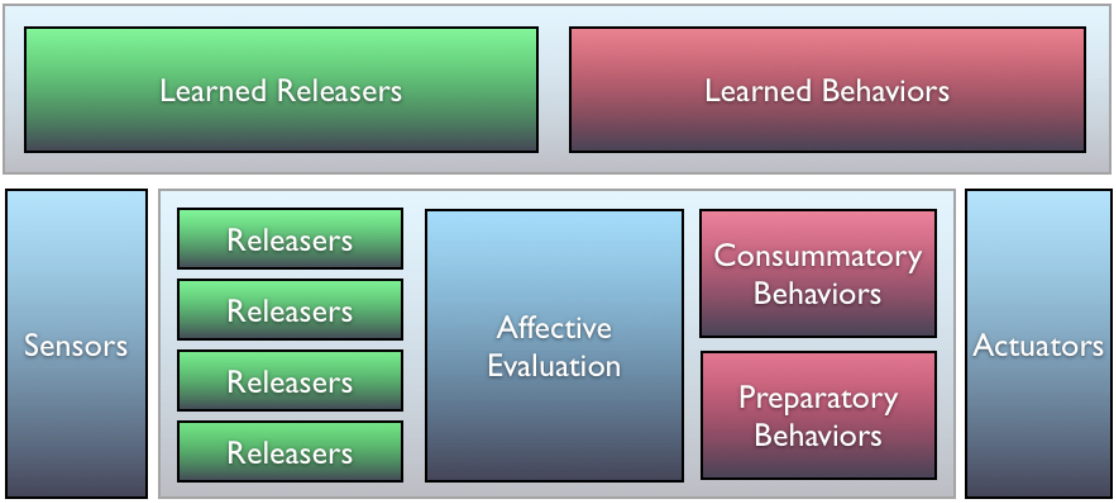

Figure 8-2: Development and subsumption of affect programs

subsumed into more complex and flexible responses that reflect culture and individual development.

In this view, the central role of affect programs is maintained, but it is now possible to introduce new higher-processing elements into each emotional response. This is in many ways similar to what Izard refers to as *Affective-Cognitive structures* (Izard, 1993), and what Damasio (1994) refers to as *secondary emotions*.

This refinement or learning process can occur both on the input, as well as the output sides of affect programs. On the input side, while basic or primary affect programs are elicited by natural releasers, secondary affect programs will be elicited by learned releasers, which correspond to stimuli to which the organism has become sensitized through experiencing its environment. On the output side, primary affect programs have fixed, stereotypical responses. Secondary affect programs, on the other hand, will implement more flexible responses that can exert control over the fixed ones, either inhibiting, enhancing, or simply bypassing them. The resulting abstraction is illustrated in Figure 8-2

Several mechanisms that demonstrate the feasibility of affective learning have been implemented as part of this work. A preliminary implementation consists of an associative network comparable to Minsky's K-lines (Minsky, 1986), in which salient stimuli (e.g., features and percepts representing objects and agents) are connected to primary affect programs when these have become active throughout the robot'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 new releaser and the active emotional system.

We have used various forms of Hebbian learning including simple Hebb, postsynaptic rules, and covariance rules. Figure 8 below, shows an example of the results obtained using a postsynaptic Hebbian rule similar to that described in (Floreano & Mondana, 1998) and shown in Equation 8.1

$$\Delta w = w(-1 + x)y + (1 - w)xy \tag{8.1}$$

Where w is the weight between both units, x is the activity level of the presynaptic unit (learned releaser), and y is the activity level of the postsynaptic unit (emotional system).

8.3 Emotion-Based Learning Systems

Based on the affective significance of events, emotion-based learning allows the robot to take into account internal needs and external stimuli in deciding what should or should not be learned in a particular situation.

Using this model, we have implemented several different emotion-based learning systems with Yuppy. These include examples of both nonassociative and associative

learning.

8.3.1 Affective Conditioning

Using the model for affective learning described above, we have developed several systems of affective or emotional conditioning across different sensory modalities (e.g., pairing tactile and auditory stimuli, or visual and auditory stimuli). Some of these systems include examples of alpha conditioning, as well as classical fear and appetitive conditioning. Figure 8-4 illustrates one example of affective conditioning in which the *Fear* affect program acquires new releasers for specific sound frequencies.

Figure 8-3 illustrates this classical scenario of *fear conditioning*. A natural releaser (presence of the blue pool-pony) generates a fearful response corresponding to the *Flight* behavior (frames 4 and 5). Different tone frequencies played with a flute, however, do not produce any activation in the *Fear* affect program, thus the *Flight* behavior does not become active either (frames 1 thru 3). If both stimuli are presented simultaneously (frames 6 thru 8), the *Fear* affect program forms a new learned releaser for the sound stimulus (see top chart of Figure 8-4) After only one trial, the newly formed releaser for the specific sound frequency that was paired with the blue pool-pony is capable of producing some activation of the *Fear* emotional system. After several more trials, the connection between the sound frequency releaser and the *Fear* affect program is strong enough to produce activation of the *Flight* behavior, and thus an emotional memory is formed (frames 11 and 12).

These results suggest that emotional conditioning is possible under the proposed model using a robot that operates in the real world, with real sensing and real action. Furthermore, it demonstrates how extending the affect program abstraction with learning mechanisms may be used to mediate and bias the robot's action-selection process.

Figure 8-3: Fear Conditioning. This sequence of frames illustrates our first attempt toward affective learning. In this fear conditioning scenario, the robot is trained to predict the occurrence of the predator via a particular sound frequency.

8.3.2 What is Learned in Affective Conditioning?

The left hand side of Figure 8-4 illustrates the associations made in this affective conditioning scenario. Essentially, what has been learned here corresponds to an association of the neutral stimulus (i.e., the sound frequency), and the affective value of the affectively significant stimulus (i.e., the evil blue pool-pony), as determined by the activity of the *Fear* affect program. This means that the neutral stimulus has acquired the properties (affective value) of the affective stimulus and now can also mediate the same kind of responses that this stimulus generates.

Besides learning associations to the specific affective value of events, could we create a model that would also learn about the general motivational implications of

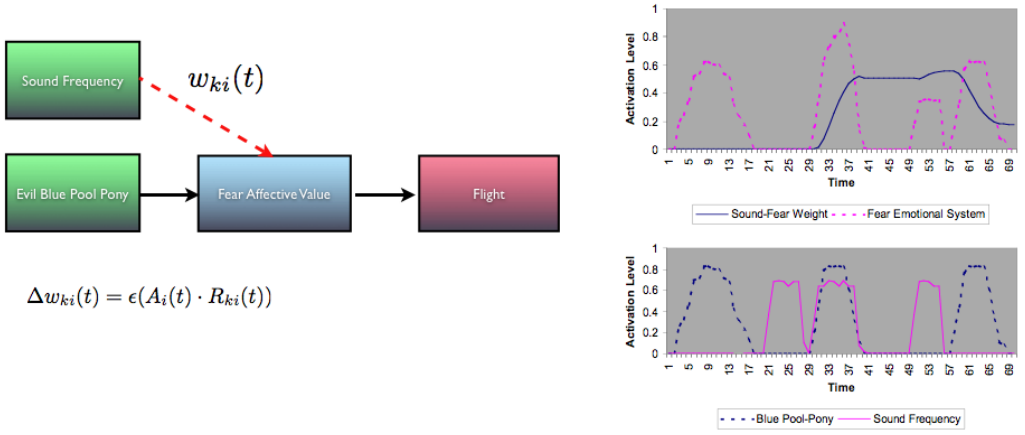

Figure 8-4: Results from Yuppy's affective conditioning to a sound frequency after pairing the sound with the blue pool-pony. Activity of the *Fear* affect program and the weight of its new connection to a sound releaser is illustrated on the top chart, whereas the occurrences of both stimuli are plotted in the bottom (right hand side of figure).

these events, and which would be influenced by regulatory mechanisms, as described in Section 3.2?

In order to do this, we added a new construct to the affect program abstraction. This construct would represent the appraisal of events, as related to their general motivational significance and how they might be influenced from regulatory mechanisms. We called this construct *Incentive Value*, and it is formalized in Equation 8.2 and depicted in Figure 8-5.

$$I_i(t) = A_i(t) \cdot M_i(t))) \tag{8.2}$$

Where $I_i(t)$ is the incentive value for affect program i at time t; $A_i(t)$ is its affective value, and $M_i(t)$ is the summed motivational influence exerted by regulatory mechanisms.

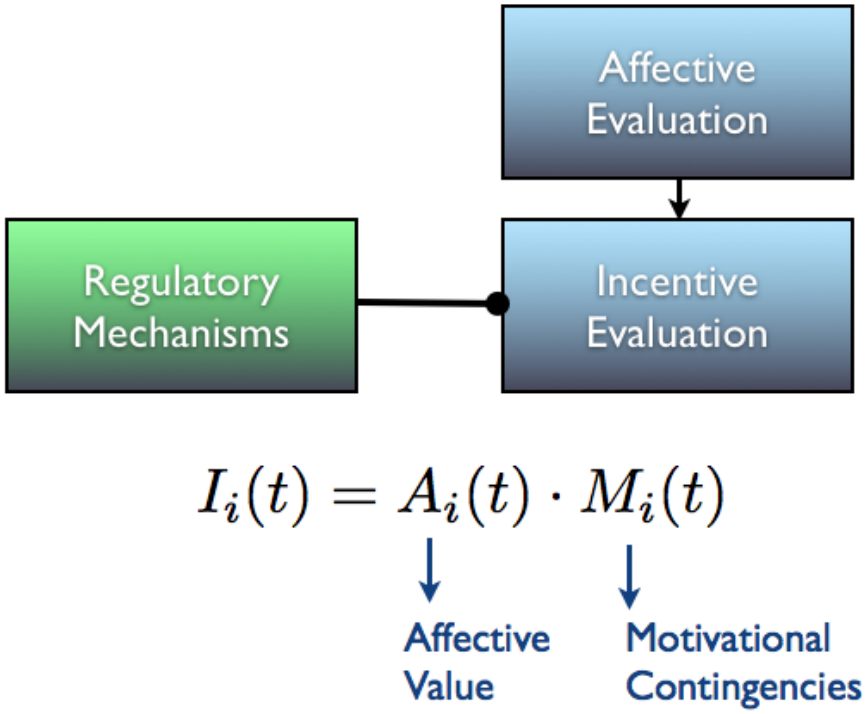

Figure 8-5: Computing the Incentive Value of Events. This figure illustrates how the Incentive Evaluation Unit computes the *incentive value* of the different contingencies detected by the affect program's releasers. This incentive value refers to the general motivational value that those contingencies represent and which can be influenced by regulatory mechanisms such as those described in Section 6.7. The specific activity pattern follows the description of Equation 8.2 in the text.

An interesting consequence of this addition is that now affect programs have two main evaluative processes: a ''liking' pathway, as mediated by the affective value of events, and a 'wanting' pathway mediated by the general incentive value of the same events. This corresponds to the same notions described in Section 3.3.1, and illustrated in Figure 8-6.

Interestingly enough, these same process have direct relationship with the separation of sensorimotor pathways into preparatory and consummatory behaviors de-

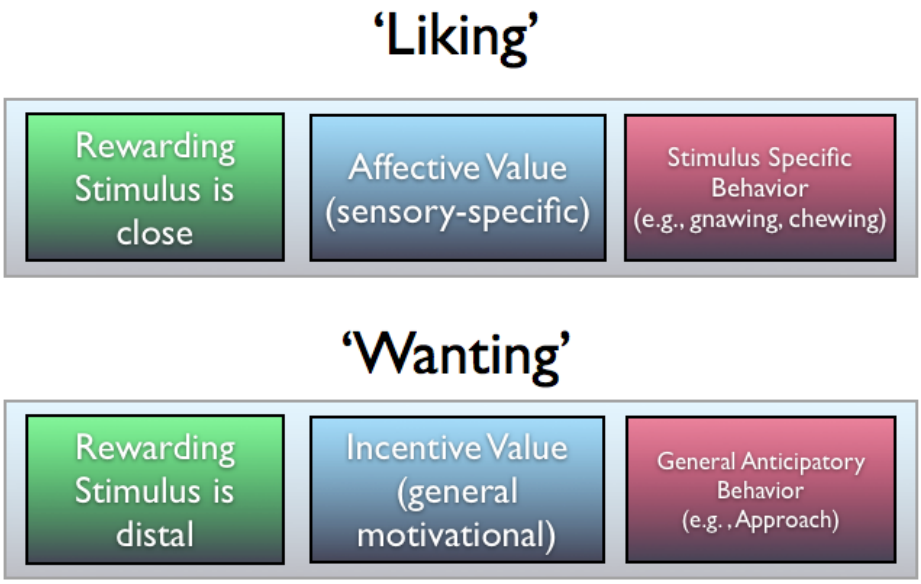

Figure 8-6: 'Liking' and 'wanting' pathways in affect programs. This figure illustrates the two pathways that mediate evaluative processes in the affect program abstraction. A 'liking' pathway mediated by the affective value of events, and a 'wanting' pathway mediated by the general motivational value of these same events.

scribed earlier, and as such are useful in the evaluation and control of such kinds of responses.

8.4 Incentive Salience

The process through which these events, or their perceptual and sensory features to be precise, acquire affective significance and hence the ability to promote and elicit actions and responses is referred as *Incentive Salience* (Panksepp, 1998; Balleine, 2004; Berridge, 2006).

The concept of incentive salience or incentive learning arose from studies that followed traditional paradigms of learning. These studies posed questions regarding the nature of rewards (i.e., unconditioned stimuli), as changes in their quality, quantity,

and time of presentation (delay) produced systematic and striking effects on behavior. Given predominant behaviorism thought, these effects were initially attributed to intrinsic properties of material objects, but as more emphasis was placed on underlying neural mechanisms, incentive motivation came to be viewed as part of those brain processes that mediate adaptive behavior. (Bindra, 1968; Bolles, 1972; Bindra, 1978; Toates, 1986; Dickinson, 1994).

It is important to distinguish this associative process of incentive learning from other types of learning, such as those discussed in Chapter 4. Incentive learning is different from learning relationships among environmental stimuli, which provides a propositional-type of knowledge base regarding the structure of the world, and which organisms can do without the intervention of unconditioned, rewarding stimuli. This type of learning is usually referred to as *declarative learning* or *stimulus-stimulus learning*, whose major neural substrates are the hippocampus and anatomically closely related neural systems (Squire, 1992; Squire & Zola, 1996; White, 1996; White & McDonald, 2001).

Incentive learning is also different from *Habit Learning*, which corresponds to the traditional stimulus-response learning whose major neural substrates have been attributed to the neostriatum and related neural systems (Knowlton et al., 1996; White, 1997; Graybiel, 1998; Hikosaka, 1998).

Summarizing, incentive learning or motivation refers to processes involving environmental stimuli predicting the perception of unconditioned stimuli. This type of learning is crucial in organisms and it has been implemented as part of this thesis allowing the robots to seek out and anticipate various rewards in their environments.

8.5 Neural substrates of Incentive Learning

The system that mediates incentive learning, together with flexible approach/avoidance responses has been named differently by many: Gray (1990) called it a *Behavioral Activation System*, Depue & Iacono (1989) calls it a *Behavioral Facilitation System*, and more recently, Panksepp (1998) has called it a *Expectancy/Seeking System*. Regardless of the name, most researchers now agree that this is a general incentive or appetitive motivational system that regulates "wanting" as opposed to just "liking".

This system is part of the mammalian brain and regulates what could be called an exploration, interest, curiosity or investigative mechanism that leads organisms to eagerly pursue available resources and extract meaning from the various situations in their environments. As part of the latter function, this same system helps the organism signify the importance of novel stimuli because of their association to opportunities for consummatory behavior.

As with all other emotions, this system might be initially without intrinsic cognitive content, it gradually helps define the perception of causal connections in the world. In essence, it drives mental phenomena that, in humans, would be associated to the experience of persistent feelings of interest, curiosity, sensation seeking, and perhaps even the search for "higher" meaning.

The main circuits for this emotional system are concentrated in the extended Lateral Hypothalamus (LH) continuum, running from the Ventral Tegmental Area (VTA) to the Nucleus Accumbens (NAS), as depicted in Figure 8-7. This system responds unconditionally to homeostatic imbalances and environmental incentives. It spontaneously learns about environmental events that predict resources, although the detailed mechanisms of such learning are still not well understood.

Electrical Brain Stimulation (EBS) of this circuit will elicit energized exploratory and search behaviors in animals (Ikemoto & Panksepp, 1999).

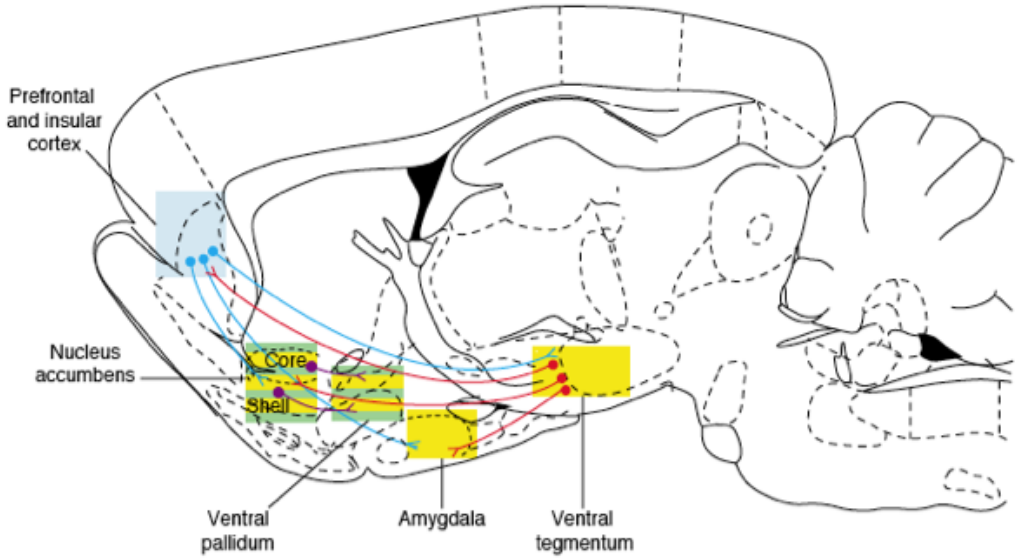

Figure 8-7: This figure depicts the main circuitry involved in incentive learning. This system is evoked by (1) regulatory imbalances; (2) external unconditioned stimuli and (3) conditioned stimuli that predict the occurrence of unconditioned stimuli and thus the opportunity for consummatory behavior. Reprinted with permission from Berridge, K. C. and Robinson, T. E. (2003), Parsing reward, *Trends in neurosciences*, 26(9), 507513.

Similarly, when NAS DA transmission is artificially enhanced using such methods as local tissue microinjections of various substances (e.g., amphetamine), it activates a state of incentive motivation and exploratory arousal, thereby generating flexible approach-seeking behaviors toward salient environmental stimuli. Conversely, the disruption of NAS DA transmission should blunt the ability of organisms to approach salient stimuli. It is not that animals lose the ability to recognize salient stimuli. Their perception and memory-retrieval processes remain intact, even though it is likely that some aspects of their attentional resources are compromised. Such animals simply are not aroused into sustained attentional-investigatory patterns by novel stimuli. It is also not the case that animals lose the physical capacity to perform instrumental tasks

or consummatory responses. Clearly, animals lacking NAS DA are not behaviorally incompetent. Rather, their deficits may arise from NAS DA no longer being able to amplify behaviorally energized states of expectancy. In other words, the output of the declarative-perception system to the approach motor system is compromised. Also, there may be deficits of positive feedback between the incentive modulator and the declarative-perception system. On the other hand, well-established conditioned responses in familiar environment contexts (i.e., habits do not appear to be disrupted by decreased NAS DA transmission. In sum, the NAS DA system is more involved in addressing unfamiliar situations or stimuli which deserve to be investigated and determining whether novel stimuli predict rewards rather than in dealing with familiar situations where organisms already have stable behavioral priorities (i.e., they have habits).

In humans, the affective state produced by stimulation of this system doe not resemble the pleasurable feelings we normally experience when performing several consummatory behaviors, but rather it resembles the "energization" felt when anticipating rewards.

Traditionally, all behaviors have been divided into *appetitive* and *consummatory* behaviors. The difference being that the former are oriented toward the seeking and approaching the various stimuli (resources needed for survival) found in the world, whereas the latter are the specific interactions and responses generated once these resources are found. The system mediating incentive learning thus appears to control appetitive activation—the search and exploratory behaviors—that all organisms must exhibit before they engage in consummatory behaviors.

8.6 An Approach to Incentive Learning in Robots

As mentioned above, incentive learning or motivation are processes that allow organisms to effectively seek out and anticipate various rewards in their environments. While research in robot control—and more generally, in agent architectures—has involved approach and avoidance behaviors in one way or another(Blumberg, 1996; Breazeal, 2000), not much has been done in terms of implementing an integrated system that regulates a variety of processes that orient attentional resources to novel events, in order to extract "meaning" from these sensory experiences, and to energize a unique class of investigatory and appetitive approach responses to a variety of affectively significant stimuli.

Our approach to building such system as part of an integrated robot architecture follows the same line of modeling described in Chapter 6. That is, we have implemented a simple mechanism for incentive learning or motivation as an instance of the *Affect Program* kind. As such, this system has all the properties of affect programs, including a set of releasers, activation and saturation thresholds, an associated decay function, and a set of associated responses (behaviors). In fact, this system is perhaps one of the best examples of affect programs as it involves not only the more fixed responses that account for rapid emotional responses, but also the idea of learning and subsumption of these systems.

It may strike as odd to some that an approach and incentive system, such as the one being described here, might be considered a candidate for an affect program as these implement emotional processes. However, this system actually has an important affective component which is in fact the primary reason for motivating approach behaviors in the first place. The affective component corresponds to what some would refer to as interest, curiosity, expectation, vigilance, or anticipation, which, in fact, have been considered by many to be part of the so-called set of basic emotions

(Tomkins, 1962; Izard, 1971; Frijda, 1986; Panksepp, 1998; Plutchik, 2001).

An important feature of this system is that its organization is broadly diffused through much of the architecture, including integration with perceptual, attentional, learning, and motor processes that have to be coordinated in order to generate successful approach responses. In other words, for successful approach, the robot must be able to recognize environmental stimuli, in addition to the ability to change behavior as a function of experience in order to maximize the probability of obtaining material resources needed for its "well being". Thus, the overall approach system presumably consists of motor, attentional, motivational, emotional, mnemonic and other cognitive sub-processes. Although each one of these sub-processes can be studied separately, we argue that in normally functioning organisms, a global system helps coordinate all of the sub-processes needed for adaptive approach toward goals. More recently, from a more psychodynamic perspective, this system has been called *the Seeking system*. Here we shall continue to argue that the meso-accumbens DA system is a part of such preparatory approach-seeking system.

Drawing upon these ideas, the Cathexis model was extended to incorporate a specific affect program that mediates all flexible approach and avoidance responses for the robot and also regulates incentive learning. This affect program was named the *Seeking* affect program as it has been described by Panksepp (1998) and it is depicted in Figure 8-8.

8.7 The Seeking Affect Program

Figure 8-8 summarizes a conceptual model that highlights the role of the Seeking affect program in regulating behavior and learning. A crucial aspect of this affect program is that it has distinct sensorimotor pathways for approach and consummatory responses as it was described in Section 7.5.2. Essential components underlying the

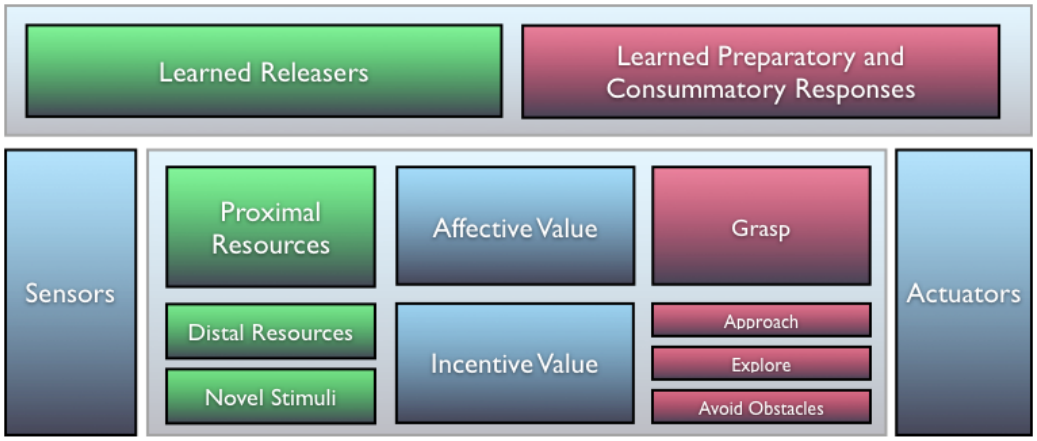

Figure 8-8: The Seeking Affect Program

approach responses include preparatory and consummatory appraisals (assessing the affective significance of the stimuli), a system involved in the formation of habitual approach/avoidance responses, and incentive-cue formation systems (see Section 8.8 below) that essentially implement the concept of incentive learning as described above.

It can also detect salience of environmental stimuli i.e., novel stimuli, conditioned stimuli, and innately salient stimuli by contrasting present input with previous memories. Thus, an important feature of the current view is the recognition of two distinct types of approach response systems: a habit response system which operates in well-trained animals and a flexible response system which operates preferentially when animals are learning about incentive contingencies in their environments.

In other words, the Seeking affect program can facilitate flexible approach responses in the presence of various salient stimuli (e.g., incentive stimuli and novel stimuli). In summary, the primary role of the Seeking affect program is to facilitate flexible approach responses by modulating incentive motivation processes. Let us develop this idea further by describing two stages of instrumental approach performance.

8.7.1 Motivating Exploratory Behavior

The Seeking affect program plays an essential role in invigorating flexible approach responses when the robot encounters salient stimuli (e.g., incentive and novel stimuli). Distal releasers and anticipatory appraisal processes detect various salient stimuli (e.g., arising from all novel events) and energizes the Seeking affect program, which in turn will spread its activation to flexible approach-seeking behaviors. Figure 8-8 highlights the routes of control involved in such processes.

8.8 Results for An Incentive-Cue Formation System

It appears that the brain is organized in such a way that the detection of unconditioned stimuli can automatically promote learning. The brain is tuned to the appearance of novel stimuli, and the Seeking affect program is activated especially by those that are associated with affectively significant events. Although these associations are not essential to consummatory reactions per se, they do establish an implicit knowledge of situations within which consummatory behaviors can be optimally expressed. The Seeking affect program is critically involved in such incentive learning processes. More specifically, changes in the Seeking affect program may first be involved in investigatory activities and more gradually in signifying the importance of environmental stimuli because of their association with opportunities for consummatory behaviors. This linking of external events with opportunities to stimulate various proximal sensory receptors is here conceptualized as the "incentive learning effect". Thus, heightened levels of the Seeking affect program add incentive properties to declarative knowledge so environmental stimuli that are predictive of unconditioned stimuli come to facilitate and energize approach (or avoidance) responses. Normal

activation of the Seeking affect program, on the other hand, maintains such incentive motivation, thus, a decrease in activity of the Seeking affect program reorganizes internal mechanisms involved in incentive representations of declarative knowledge so that environmental stimuli that are not predictive of the perception of unconditioned stimuli will no longer activate approach responses.

The following scenario illustrates the results obtained when placing our simulated robot in a situation in which neutral stimuli are contingent on incentive stimuli, and thus, come to signal their occurrence once *incentive salience* has been attributed to them, as described herein. In this scenario, Marvin is situated in an environment in which the recharging station (an incentive stimulus) is detected by the robot's distal releasers. The robot's battery level is low and thus an impendingneed of recharging is mediated by the robot's *Seeking* system. Figure 8-9 illustrates this scenario. In frames 2 thru 5, it is shown that the detection of the recharging station triggers an approach response as mediated by the *Seeking* system. In a similar manner, in frames 6 thru 8, the detection of the recharging station by proximal releasers elicits the consummatory grasping response also controlled by the *Seeking* program, which increases the battery level to appropriate values and ends the scenario. Throughout this time, however, the neutral stimulus represented by the purple block has been contingent in all of the activity of the *Seeking* system. By means of the same associative learning rules described earlier, two new releasers are associated to the *Seeking* system. The first releaser is a distal releaser for the purple block that is associated to the preparatory pathway of the *Seeking* system, whereas the second releaser is a proximal releaser for the same purple block, but which is associated to the consummatory pathway and thus it would be used by this affect program to trigger grasping responses targeted at the purple block.

The second part of this scenario illustrates the results obtained when placing our simulated robot after the processes of incentive salience, described above, have taken

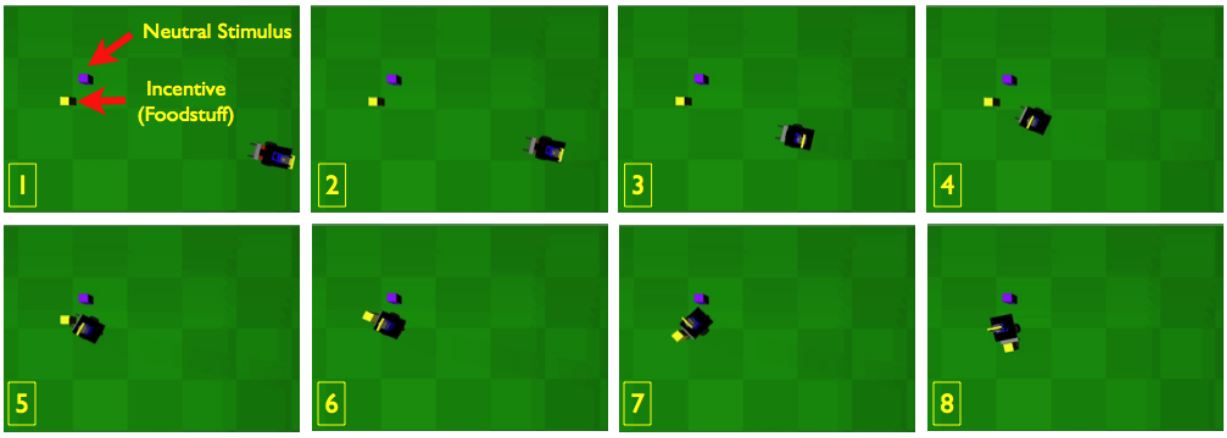

Figure 8-9: Incentive Salience. This sequence of frames illustrates how the Seeking Affect Program works, triggering approach responses to a recharging station or foodstuff (lower red arrow) in response to a regulatory goal (low battery level). In the process, the neutral stimulus (upper red arrow) will be detected as well and incentive or motivational properties will be attributed to it as it is a cue that predicts the occurrence of the goal, and thus it facilitates consummatory behavior.

place. The setup is the same as before and is illustrated in Figure 8-10. In frame 1, the robot is shown with the two same stimuli, only the stimulus marked as neutral has already been attributed the motivational properties of the incentive (the yellow block). In frames 2 and 3, it is shown that the detection of the previously neutral stimulus now also triggers an approach response as mediated by the *Seeking* system. This happens because the purple block now has general motivational value, based upon the incentive learning that occurred in the last scenario. In frames 4 and 5, the robot detects the presence of the recharging station, and given its higher incentive value, an approach response targeting this incentive ensues. Finally, in frames 6 thru 8, the detection of the recharging station by proximal releasers elicits the consummatory grasping response in the same manner as before, which increases the battery level to appropriate values and ends the scenario.

Thus, the Seeking affect program plays an important role in integrating reward-

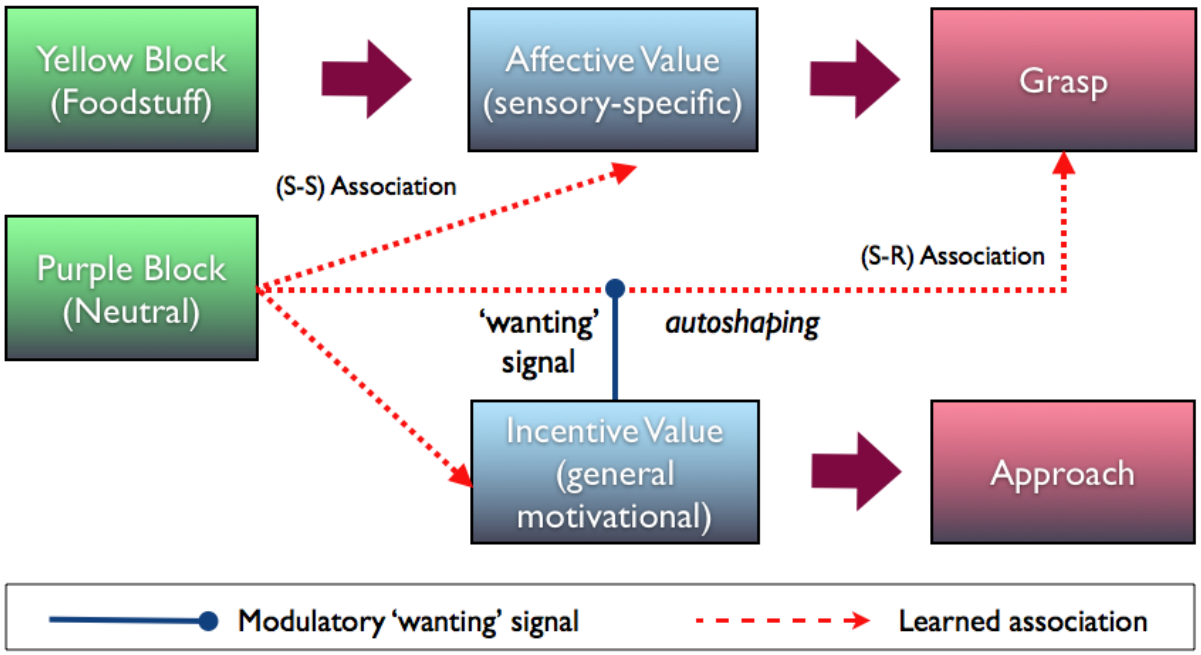

Figure 8-10: Incentive Learning Associations. This figure depicts the kinds of associations made when hebbian learning takes place under the incentive learning paradigm. Incentive salience is attributed to the purple block (neutral stimulus) by means of two simple S-S associations: (1) between the stimulus and the general incentive ('wanting') pathway of the *Seeking* system, which mediates approach responses; and (2) between the same stimulus and the affective-specific ('liking') pathway of the same system.

related information on specific aspects of the environment into conditioned approach-seeking reactions. Once such conditioning has been established in a specific context, however, heightened the activity of this affect program is no longer necessary for its expression, unless the robot experiences new opportunities for consummatory responses in those contexts. The following figure highlights key routes involved in such incentive modulation processes. As described above, the Seeking affect program is only involved in learning in the sense that it modulates the initial behavioral responses to potential incentives, and the development of conditional incentive stimuli (i.e., the automatic valuation of neutral environmental events). It is not involved in

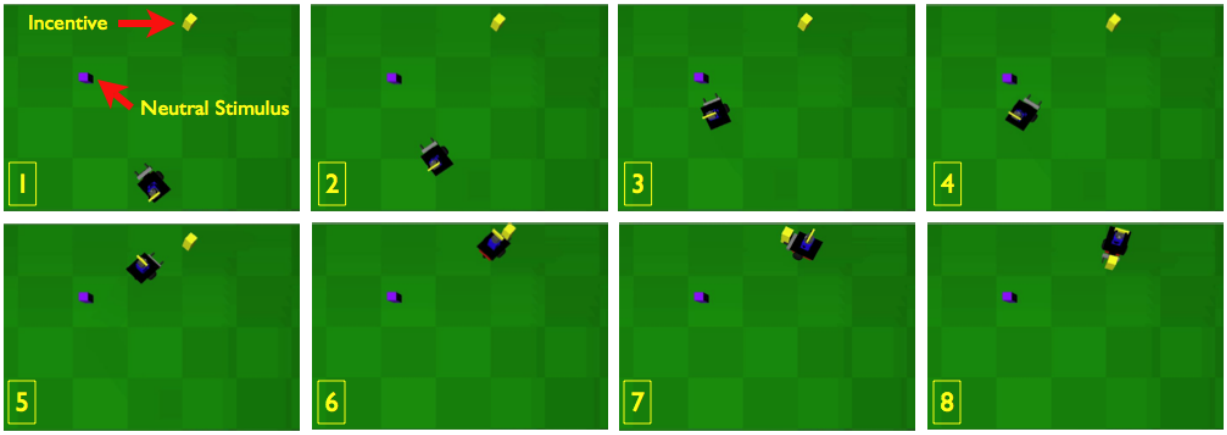

Figure 8-11: Incentive Attribution. This sequence of frames illustrates how the Incentive Salience Model works by attributing incentive or motivational properties to an otherwise neutral stimulus. In the frames, the purple block acquired incentive properties and now the robot approaches it whenever detected, by doing so, the robot has come closer to the goal (yellow block), which when detected, is approached and a final consummatory behavior corresponding to a grasp response is executed.

declarative memory formation or retrieval nor in procedural learning, which are types of learning out of scope from the present work.

It should be noted that all approach and navigation-related behaviors include the implementation of the Nearness Diagram (ND) obstacle avoidance algorithm (Minguez & Montano, 2004).

8.9 Habit Learning

Many acts that we perform regularly become so routine that we carry them out almost without conscious effort. We depend on these habits to free us to think and to react to new events in the environment. For instance, as I write this chapter, I repeatedly press the combination of keys that allow me to save the document. I do this in a preventive effort to maintain a saved version of the document with the latest changes.

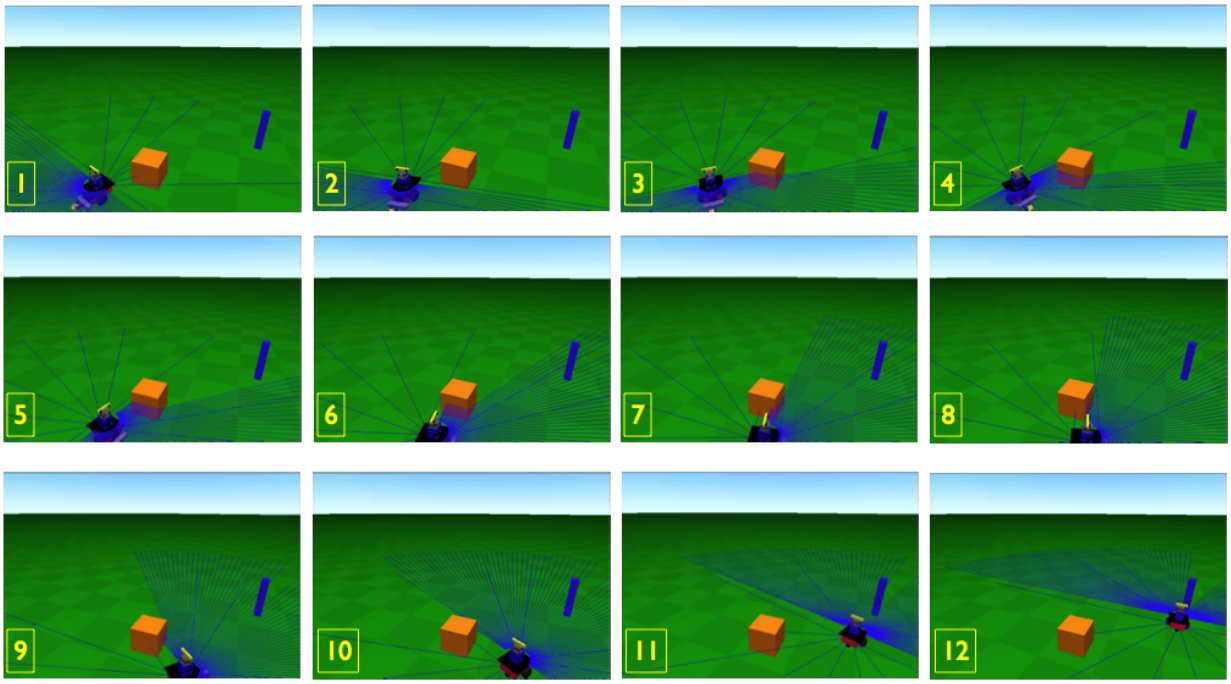

Figure 8-12: Obstacle Avoidance. This sequence of frames illustrates how the approach behavior subsumes an obstacle avoidance behavior, which implements the Nearness Diagram (ND) Navigation by Minguez & Montano (2004).

This trivial task seems to be happening somewhere outside the realm of consciousness as I do not even notice when it happens, and only now, that I stop to think of an example of a habitual response do I realize that it has been happening all along.

The formation of habits is an essential aspect of learning, and as it was described in Chapter 4.3, its neural underpinnings are dissociated from all other types of learning that we have described and reviewed so far, and which seem to be related to the hippocampus system and anatomically close neural systems (Squire & Zola, 1996; White & McDonald, 2001).

Previous sections presented work toward the development of refined affect programs that subsume the activity and functionality of primitive ones on their input side (learning mechanisms for the acquisition of new releasers). This work addressed

the issue of incentive learning or affective conditioning, which corresponded to stage 2 of affective learning. This chapter describes the development of a learning mechanism for the output side of affect programs that implements simple habit learning, and which corresponds to some of he interactions involved in the third stage of the model for affective learning described at the beginning of the chapter.

8.9.1 Definitions

Before discussing the different mechanisms involved, some definitions are due. Let us start with describing the type of learning addressed in this chapter. Habit learning refers to the traditional type of stimulus-response (S-R) learning, also called reinforcement-response learning. In this type of learning (see Figure **??**) organisms may accidentally or unintentionally make any response (R) in the presence of an environmental stimulus (S). If a reinforcer (S*) is encountered at around the same time as these events, an association between the stimulus and the response is strengthened or enhanced, thus increasing the probability that the stimulus will elicit the same response in the future. In the S-R model, any response that is performed can, in theory, become associated with any stimulus that happens to be present if the two are temporally contiguous with a reinforcer. This function of reinforcers was originally described by Thorndike (1911), who referred to it as the "stamping-in" of stimulus-response (S-R) bonds. The basic mechanism was adopted by Hull (1943) for his theory of learning, and later elaborated by others (Estes, 1959; Schacter, Chiu & Ochsner, 1993). Because this type of learning is thought to proceed in an automatic, unconscious manner, it has been called "habit" learning by Mishkin (1984).

8.9.2 Correlational Learning

Most biologically plausible models of learning assume that modifications of synaptic efficacy (or more generally, the strength of the connection among computational processing units) can account for a variety of forms of learning.

All learning mechanisms contributed as part of this thesis follow three main principles:

1. **Simplicity**: These mechanisms rely on simple correlational learning rules that have a limited set of features and functionality that makes them accessible for analysis.

2. **Specificity**: These mechanisms are not intended to be general purpose learning systems. Their activity and resulting mapping is constrained by the affect programs to which they belong. Thus, all learned relations are tied to that affect program's domain.

3. **Biological feasibility**: These mechanisms have interesting relations to biological data that supports their main assumptions and constructs.

We have implemented a habit learning mechanism that, as with our approach to incentive learning, follows these principles and focuses on learning rules that can account for not only the temporal coincidence of events, but also their temporal order.

Most accounts of synaptic plasticity—which is considered as the main mechanisms to support learning and memory in the brain—rely on neural activity to change synaptic function. Usually, the notion of some correlational learning rule is used to model this type of plasticity. A correlational learning rule, often called a Hebbian learning rule[1], uses the correlation between presynaptic activity and postsynaptic

[1]Donald Hebb (1949) proposed a learning rule for the modification of synaptic strengths that can be summarized as follows: If neuron A repeatedly participates in the activation of neuron B, the synapse from A to B is strengthened.

response to drive changes in synaptic efficacy (Churchland & Sejnowski, 1992).

8.9.3 Beyond Temporal Coincidence

Hebbian learning rules, such as those described as part of the incentive learning mechanisms, are correlational because the changes in the efficacy of the connection between computational units is tied to the associations between inputs. However, there are several cases in which just using temporal coincidence is not sufficient in order to learn certain relationships about events. Consider for instance the case of an organism learning a sequence of events (or actions) that lead to a specific reward. In such cases, learning the specific order of events (or actions) is important and thus capturing information relative to the temporal order of such events becomes necessary.

Hebbian learning rules are sensitive to the temporal coincidence of inputs, but not to their temporal order. In other words, a simple Hebbian learning rule is not sensitive to whether input A follows input B, or viceversa. Therefore, a Hebbian rule would not be sufficient to develop the predictive relationships that occur between stimuli during an instrumental conditioning task as defined above in Section 8.9.1.

Predictions are very useful for organisms to prepare themselves for the future, given their current state, the state of the environment, and the possible set actions that could be performed. With a robot the case is not different. For a robot to fulfill its many different and often conflicting goals, it has to decide what action is most appropriate given a specific situation. In such case, predictions become useful in many ways. For instance, they can be compared to actual events and thus compute possible errors to improve their prediction in the future, which would increase the probability of achieving better performance (such a mechanism is believed to be in place in the *Effective-Reinforcement Hypothesis* described in Section 4.4.3). Similarly, predictions can be used in order to prepare a set of actions in anticipation of a particular event,

which would also yield better performance.

Given the importance of predictions, it is conceivable to expect that there exist several predictive systems in the brain that act at a variety of spatiotemporal scales.

8.9.4 Predictive Learning

Predictive learning is a requisite of the type of learning involved in habit learning. In other words, we would argue that habit learning requires the existence of anticipatory mechanisms which predict or signal the expectancy of a reinforcer, and in which the temporal order of actions is accounted for so that the appropriate relationships between those actions and their outcomes can be learned.

Previous work, both theoretical and modeling, has focused on the need of such mechanisms to explain animal learning (Rescorla & Wagner, 1972; Sutton & Barto, 1981). The work presented here extends these ideas and ties them to an affective perspective. The following sections present some modeling approaches and arguments that address predictive mechanisms as a principal component in the implementation of habit learning.

A Global Predictive Learning System

Information about reward is passed to many different brain structures in part through diffuse ascending systems of axons that originate in small nuclei in the midbrain and basal forebrain (e.g., Cooper 1970). This work extends those ideas and implements a system that in a similar way, sends information to other systems regarding the expectancy of reward and thus modulates how stimulus-response associations are formed.

A predictive system, labeled P, receives convergent input from perceptual systems, as well as inputs carrying information regarding the occurrence of rewards or any

other salient stimuli. This predictive system computes its change in activity over time, which can be described as:

$$\delta(t) = V(t) - \bar{V}(t) \tag{8.3}$$

$$V(t) = \sum_{j=1}^{\eta} w_j x_j + r(t) \tag{8.4}$$

Where $V(t)$ is the net input to the predictive system at time t, including the reward stimulus $r(t)$, and $\bar{V}(t)$ is a running average that can be represented by Equation 8.5

$$\bar{V}(t) = \lambda V(t) + (1 - \lambda)\bar{V}(t - 1) \tag{8.5}$$

where $0 < \lambda < 1$ is a constant that defines how much into the past, the activity is averaged. As λ approaches 0, the average reaches farther into the past. As λ approaches 1, the averaging interval becomes short, and the output of P thus closely approximates the net input $V(t)$ at time t.

From Equations 8.3 and 8.5, we obtain:

$$\delta(t) = (1 - \lambda)[V(t) - \bar{V}(t - 1)] \tag{8.6}$$

The output of P reflects a scaled temporal difference between the current net input and the previous running average of the net input.

Any change in the strength of the connections between the perceptual system units and rewarding stimulus and P follows a simple correlational rule, such as that generalized in Equation 8.7

$$\Delta w(t) = \eta x(t)\delta(t) \tag{8.7}$$

Thus, the output $\delta(t)$ is not simply the magnitude of the reinforcer, but rather a comparison of the net input throughout time.

When a rewarding stimulus is first met, it increases the ouput of P because at that time, the output $\delta(t)$ is proportional to $r(t) + \bar{V}(t-1)$, where $V^*(t)$ is the total net input to P not contributed by $r(t)$. As the actual delivery of information about the reward rises and falls to baseline, the running average $\bar{V}(t)$ will follow slowly. During learning, the weights are changed according to Equation 8.7 until the running average of the input $\bar{V}(t)$ from the perceptual systems correctly predicts delivery of reinforcement, so that $\delta(t) = 0$. This kind of predictive mechanisms has been used in a variety of engineering contexts (Sutton & Barto, 1981; Sutton, 1988). Here, these mechanisms are used to drive prediction so that it can be used as a global signal that drives learning of habits.

8.9.5 Implementing Habit Learning

This section describes the extension the affect program abstraction with learning mechanisms on its output side. Once again, the idea is that the fixed responses that are triggered by various stimuli, as they activate any given affect program, will be subsumed into more flexible responses that reflect the learning of the robot. As it was illustrated in Figure 6-2, these flexible responses, which may be composed of multiple action sequences, will be able to exert control (inhibition or enhancement) over the more fixed ones, and even bypass them through their direct coupling with actuators.

1. To construct and reinforce contexts: The predictive signal is a general response to affective stimuli and does not differentiate among the different kinds of such stimuli. The combination of specific releaser input, the predictive signal, and input from specific affect programs (e.g., fear), can specify a more detailed context for the processing of behaviors.

2. To selectively enhance certain behaviors, and focus attention: As releasers present repeated coherent input patterns, loosely termed here as contexts, the dopamine-like signal can increase signal-to-noise ratio in the selection of behaviors, exerting a focusing effect whereby only the strongest inputs are selected and weaker activity is lost.

   Mechanisms analogous to this one have been suggested and believed to be mediated by the interactions of D1 and D2 families of dopamine receptors in striatal neurons (Nicola, Surmeier & Malenka, 2000). In addition, such mechanism would correspond to some of the ideas described in the "switching" hypothesis reviewed in Section 4.4.3.

3. To use the signal as a reinforcer of behaviors that are active during appropriately rewarding contexts: In conjunction with predictive Hebbian learning (see Section 8.9.4 above), we have used this signal for the acquisition of behaviors that lead to certain outcomes (i.e., habit learning). This strategy corresponds to the effective-reinforcement hypothesis, suggested by Schultz (1998), that was reviewed in Section 4.4.3. The main idea is illustrated in Figure 8-13 and can be described as a two-stage learning process. In the first stage, the predictive system acquires responses to reward-predicting stimuli. In the next step, the resulting predictive signal would specifically strengthen those connections that are active at the time of the reward-predicting stimulus, whereas the inactive ones are left unchanged. This essentially constitutes a three factor learning rule where the changes in connection strengths depends on the pre-synaptic activity (input from releasers), the post-synaptic activity (activity level of behavior unit), and the predictive signal.

   Thus, a simple learning rule to modify the weights in those connections is described in Equation 8.8

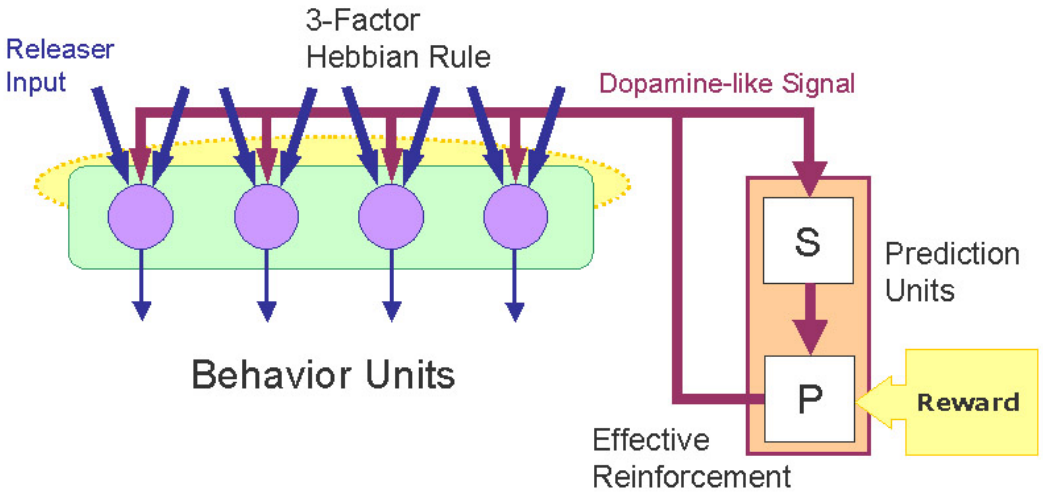

Figure 8-13: Predictive hebbian learning in the formation of habits

$$\Delta w = \epsilon \cdot p^{\cdot} \cdot i \cdot b \tag{8.8}$$

where p is the 'wanting' signal, i is input activity from releasers, b is the activity from a behavior unit, and ϵ is some learning rate.

There are many other alternatives for Hebbian rules, and this one is only suggested as part of this implementation. The important issue here is the incorporation of predictive models in the form of three-factor Hebbian learning rules.

4. To reinforce behaviors that were active some time steps before the predictive signal: This is the same strategy as in 3. The only difference is that this time the 3-factor rule considers traces of pre- and post-synaptic activity as well as the existence of the predictive signal. This requires some kind of memory, or as Schultz calls it, an eligibility tracing mechanism (yellow ellipse in Figure 8-13; (Schultz, 1998). Although the existence of such tracing mechanism is not

known, mechanisms mediated by NMDA receptors and nitric oxide (NO) have been suggested as possible candidates (Houk, Adams & Barto, 1995).

8.9.6 Competition Between Systems

The mechanisms for habit learning just described run in parallel with those existing for incentive learning. They, however, are mechanisms that are both located at different places (output and input sides of affect programs, respectively) and designed for very different purposes in mind. However, there are certain tasks in which both mechanisms might be doing the same exact type of learning. Consider for instance a scenario in which, through incentive learning, an association is learned between a previously neutral stimulus S and a reinforcer or unconditioned stimulus S*. Through this association, an unconditioned response R usually triggered through S* will also be triggered now by S. So far no problems. Consider now, that, through habit learning mechanisms, an association will be made between the stimulus S and the response R as long as there is a contingent presentation of the reinforcer S* together with the stimulus S.

This shows how two different learning systems actually may compete with one another in learning the same type of information. The interesting thing here, however, is that there is actually evidence that this is the case in the mammalian brain as well. The type of learning mediated by the amygdala (stimulus-unconditioned stimulus associations) competes with that of the striatum (stimulus-response associations) (White & McDonald, 2001).

8.10 Limitations and Extensions

The declarative-perception system that mediates learning of stimulus-stimulus relationships, such as as those mediated by the hippocampus is a complex assembly of

heterogeneous cognitive processes, outside the scope of the present work, that represent knowledge about the environment allowing organisms to acquire and recall the relationships among environmental objects and events without the essential intervention of unconditioned stimuli (i.e., stimulusstimulus or declarative learning). A future extension of this affect program might involve the construction of such system.

8.11 Summary

Incentive learning or motivation in the present thesis refers to processes involving environmental stimuli detected by distance releasers predicting the perception of unconditioned stimuli, usually detected by proximal releasers, which allow agents (in this case our robot), on future occasions, to effectively seek out and anticipate various rewards in their environments both (material objects as well as immaterial ones, like safety).

Chapter 9

Affective Interactions

The solutions implemented by affect programs, in order to deal with the contingencies that organisms face in their environments, must involve the activity of a variety of mechanisms that operate at various levels. We have suggested that the coordination and synchronization of these mechanisms is, at an operational level, the main reason for emotion. Yet, emotion, as we have defined it here, must interact with a number of processes in order to effectively implement these solutions.

This chapter describes some of the interactions between affect programs and other mechanisms we have deemed important in producing coherent and intelligent behavior. To illustrate these principles and ideas, we focus on the interactions between affect and attention.

9.1 Modulation of Attention

Information flow through the affect programs has been arranged in such a manner that it promotes taking actions specifically to obtain more information from the world. This is commonly referred to as goal-directed perception (Maes, 1995).

Drawing its inspiration from current knowledge on information-processing in the

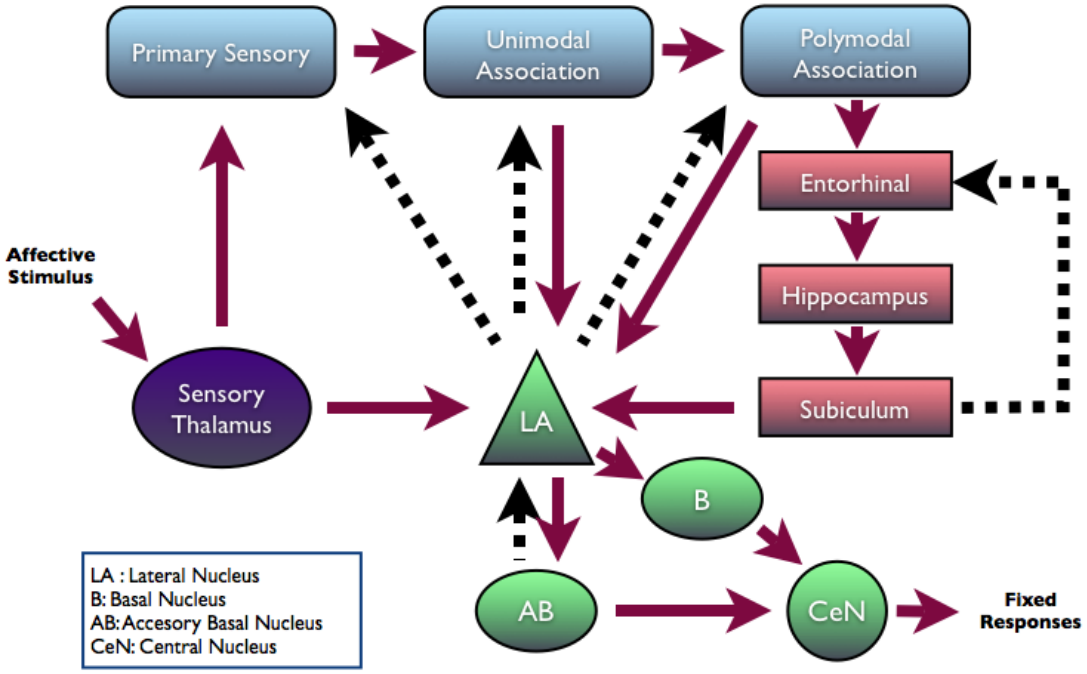

Figure 9-1: Information about external stimuli reaches the amygdala by way of direct pathways from the thalamus as well as by way of indirect pathways from the thalamus to primary and associative cortices and then to the amygdala (pathways shown correspond to auditory cortex). Bypassing the cortex in the direct pathway allows for a fast response from the amygdala but does not benefit from cortical processing and thus only low-level features of the stimulus may be detected. LeDoux (1996) has speculated that this direct pathway may be responsible for those affective responses we do not fully understand. (Adapted from (LeDoux, 1993) and (LeDoux, 1996).)

amygdala (see Figure 9-1) in which information regarding possibly significant stimuli reaches the evaluation centers of the amygdala through different pathways (LeDoux, 1996), the perceptual systems in Cathexis initiate two different pathways from which information about external and internal stimuli reaches Affect Programs. Figure 9-2 shows an example of affective processing through these pathways. The low activation pathway is a direct connection from simple perceptual systems to an Affect Program. Because of its simplicity in processing information, this is a faster transmission route compared to the high pathway. However, it only provides the Affect Program with a

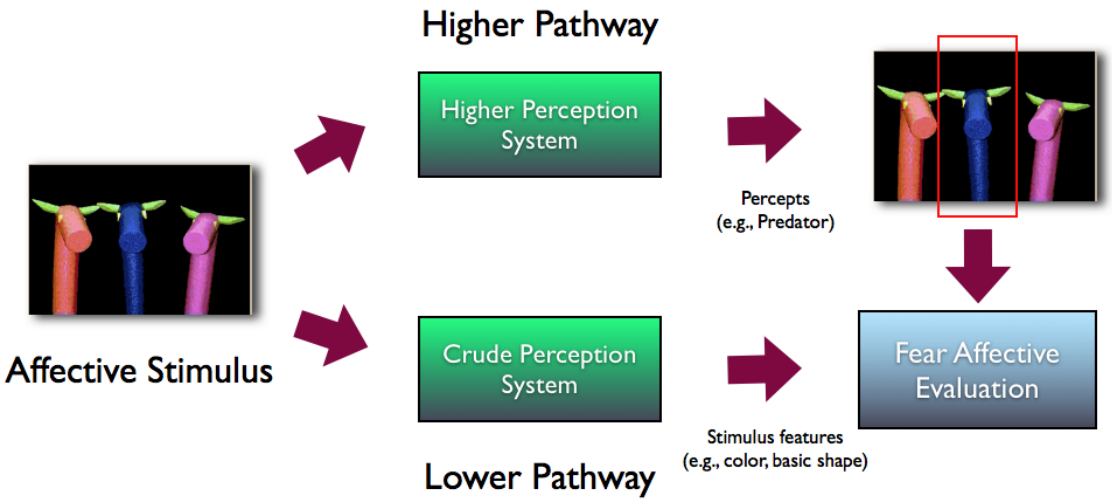

Figure 9-2: Affective-processing pathways in the Cathexis framework. Perceptual systems are connected to affective evaluation systems via two separate pathways: a direct or low pathway that provides emotional systems with fast, but simple representations of a stimulus, and an indirect or high pathway that involves more complex, time-consuming processing, but provides a richer representation of the stimulus

crude representation of the stimulus. The high pathway, on the other hand, involves more complex, time-consuming processing that provides a richer representation of the stimulus.

This mechanism of multi-scale activation pathways allows for flexible affective processing and facilitation of attention as the robot can initiate appropriate responses before fully identifying a particular stimulus. In addition, the multiple time scales are useful to integrate and deal with perceptual processes that are computationally expensive (e.g., auditory processing and vision). To illustrate this, consider the following experimental scenario in which a ball is passed in front of the robot. In such situation, simple vision systems might detect features of the stimulus (e.g., motion and color), and send this information, via the low pathway, to systems that can initiate appropriate responses (e.g., orienting towards the stimulus and perhaps looking at its approximate location). At the same time, more complex processing systems may

be at work determining the nature of the stimulus. Once this information is available, it is sent, via the high pathway, to the appropriate systems in order to enhance or suppress ongoing responses (e.g., allow an approach response if it is a ball, suppress it otherwise).

Arranged in these two pathways, we implemented different perceptual systems for our robots based on visual and auditory processing that were capable of providing both stimuli features (e.g., color, shape, motion, intensity of sounds), and objects (e.g., people, specific sound frequencies, styrofoam bones and horses).

9.2 Orienting Responses and Habituation

While attention has usually been related to almost every aspect of cognitive processing (e.g., from the modulation of perceptual processes, to the control of goal-directed behavior and even further to the mechanisms that may mediate consciousness), rarely has attention been related to emotion.

Given our functional perspective on emotion and affective processing, we argue that attention is inherently related to the notions of affect programs, as the many functions that comprise what has been referred to as attention are also essential functions of affect programs in their role of coordinators of responses that deal with an organism's biologically significant contingencies.

As defined in Chapter 3, from an operational point of view, attention corresponds to the set of mechanisms that serve to provide coherent control of behavior. When organisms are confronted with a variety of stimuli, they must be able to selectively use some of the features (including both sensory-specific and general affective features) of certain stimuli in the attentional space, and ignore others, while performing some specific behavior in response to environmental events.

As part of our solution to generate coherent behavior from an affect programs

perspective, we have implemented an instance of the *Surprise* affect program which deals with issues of novelty and behavioral functions that may be described as aspects of attention, including most notably the modulation of Orienting Responses (ORs).

An illustration of this affect program is depicted in Figure 9-3. As this figure indicates, the preparatory pathway for this system deals with the detection of distal stimuli that are usually novel to the agent and thus promote head ORs. The consummatory pathway, on the other hand, deals with events that are proximal and may be of attentional importance. Thus, when such events have been assessed as affectively significant, full body ORs are produced, which have two main effects: (1) by virtue of the perception-processing pathways described earlier (see Section **??**), a full OR toward an affectively significant stimulus guarantees that the agent directly faces such stimulus and thus it can process it further with higher perception systems (i.e., it guarantees the apportionment of further processing of the stimulus)[1]; and (2) it facilitates the activity of other affect programs such as the *Seeking* or the *Fear* systems which might implement a more appropriate solution to the specific contingency, once it has been fully evaluated.

Interestingly enough, recent research suggests that the central nucleus of the amygdala is involved in these same attentional functions (Holland & Gallagher, 1999). Thus, our implementation of a *Surprise* affect program, while at a much higher level of abstraction, is consistent with current knowledge of some of the neural substrates behind the interactions between affect and attention.

[1]This is an example of taking an action specifically to obtain more information about the world. A process usually referred to as goal-directed perception.

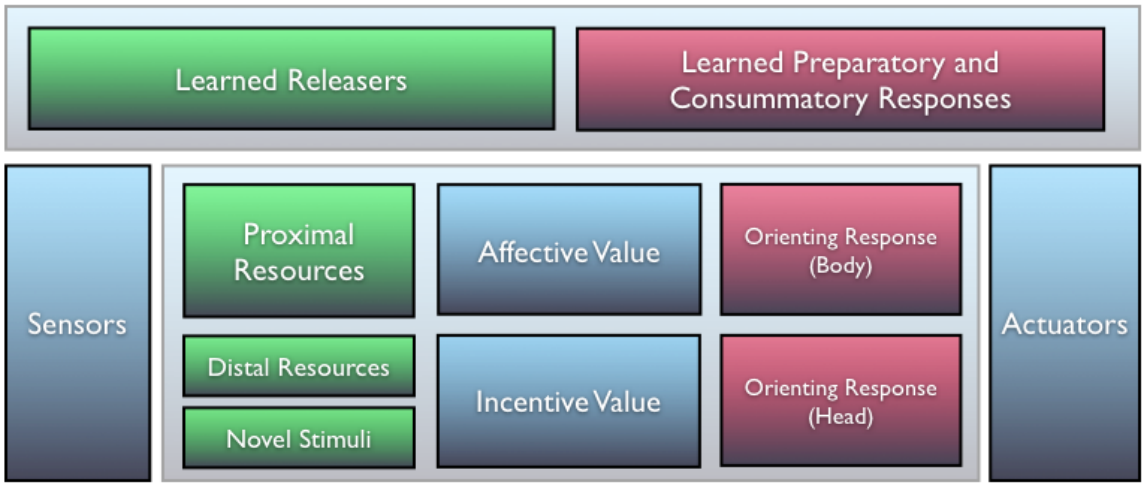

Figure 9-3: The *Surprise* Affect Program. This figure illustrates a pictorial description of the Surprise affect program which deals with behaviorally separable functions that have been described as aspects of attention, including the detection of novel stimuli and the mediation of Orienting Responses (ORs). As indicated in the figure, the preparatory pathway for this system detects distal stimuli that are novel to the agent and promote head ORs. The consummatory pathway deals with events that are proximal. When such events have been assessed as affectively significant, full body ORs will be produced.

9.3 Mediation of Orienting Responses

Affectively significant events must be capable of interrupting an organism's activities be they the execution of a particular behavior or ongoing stimuli processing. In our simulated robot, and through the activity of the *Surprise* affect program, novel stimuli as detected usually by distal releasers (e.g. sudden sounds, or the appearance of an object in the visual field) raise the affective value of this system, which, when selected through the action selection process described before, triggers and controls head Orienting Responses (ORs).

With repeated presentation of these same stimuli, however, and with no other significant consequence (i.e., the stimulus is not assessed to be of affective significance),

these ORs are diminished and habituate over time. We achieve this behavior through the habituation mechanisms described earlier in Section 6.6.2.

9.3.1 Results of ORs and Habituation

The following scenario illustrates the results obtained when placing our simulated robot in a situation in which novel stimuli are detected by its distal releasers. In this scenario, Marvin is located in an environment in which a prominent object (i.e. a big red ball) suddenly appears in the robot's perceptual field, as detected by its laser range finder systems, which feed directly into the robot's releasers. Figure 9-4 illustrates this scenario. In the first frame, the red ball (marked by a yellow arrow) appears in the robot's perceptual field. In frames 2 and 3, it can be seen how this event triggers a head OR on the robot, which is mediated by the *Surprise* affect program. The extent of the OR can be seen both by the movement of the robot's head, as well as through the robot's visual system (i.e., images captured from the robot's camera) as illustrated in the lower right segment of each frame. The red ball has no specific significance to the robot. In other words, it was not included as part of the affective stimuli that the robot would be interested in, and in this particular case, it had not acquired any learned significance either. In frames 4 and 5, the response is retracted and the robot's head faces the same direction as its body. The continued presence of the stimulus, triggers the same activity through the *Surprise* affect program, which can be seen in frames 6 to 9. After a couple more occurrences of this kind of activity, the releasers habituate and hence so do the ORs, as seen in frames 10 and 11 where the OR was much smaller, and finally in frame 12, where ORs are no longer generated.

These results are also consistent with research on amygdala function, which have implicated this set of nuclei in the control of orienting behavior toward salient and

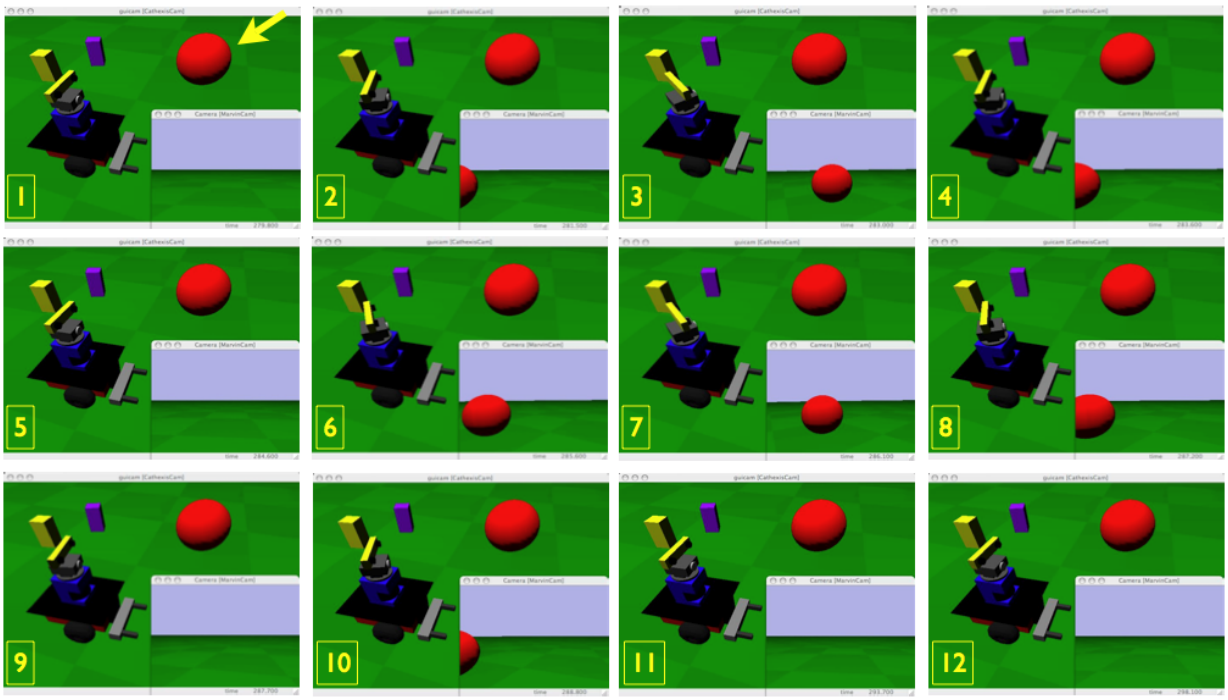

Figure 9-4: Habituation of Orienting Responses (ORs). Through the mechanisms of habituation described in Section 6.6.2, attention to stimuli can be modulated and the onset of ORs will habituate over time, as the stimulus that triggered the response (marked by the yellow arrow) is not deemed to be of affective significance for the robot.

novel stimuli. Even as early as 1951, researchers noted that electrical stimulation of the amygdala complex produced both "attentional" and "affective" responses, which habituated over time when the stimulus was presented repeatedly (Kaada, 1951). These findings have been confirmed more recently and they have showed that once the stimulus is paired with food delivery or with other affectively significant stimulus, the ORs reemerge, often attaining a level considerably higher than that which occurred when the stimulus was first presented. This kind of potentiation is suggested to be a product of the same sort of associative and affective learning processes we described in this work.

These ideas were also tested to some extent, by instantiating the *Surprise* affect program on the physical robot Coco. In the following scenario, we demonstrate how the same affective-attention interactions prove useful in integrating different sensory modalities, as well as promoting goal-directed attentional processes in a physical robot.

This scenario illustrates the results obtained when placing Coco (our baby gorilla robot) in a situation in which novel novel sounds with abrupt onsets (e.g. the sound of the experimenter calling the name of the robot) generate head ORs and facilitates the apportionment of further processing of the present stimuli, which ultimately results in complete body ORs. In this scenario, Coco is situated in our lab and the experimenter calls the robot's name while positioning himself on either side of the robot. These abrupt sounds are detected by Coco's auditory processing systems, which feed directly into the robot's distal releasers. Figure 9-5 illustrates this scenario. In the first frame, the experimenter calls the robot's name. In frames 2 and 3, it can be seen how this event triggers a head OR on the robot, which now facilitates the processing of stimuli via other sensory modalities, in this case the robot's visual system. This simple behavior, which mediates goal-directed perception, All of these responses are mediated and coordinated by the *Surprise* affect program. Given that the detection of people is an event deemed to be of affective significance to the robot (i.e., it was hard-wired into its affect program's releasers), the appropriate response can be coordinated and executed. In this instance, the *Surprise* affect program's proximal releasers detect the presence of the experimenter and a full body OR ensues as seen in frames 3 and 4 (notice the position of the robot with respect to previous frames). Had the presence of a person not been of affective significance, the head ORs would have habituated much as it was described in the results of the previous scenario with our simulated robot. In frame 5, the experimenter repeats the procedure (i.e., calls the robot again), this time from the other side. In frames 6 thru 8, the robot exhibits once again head ORs

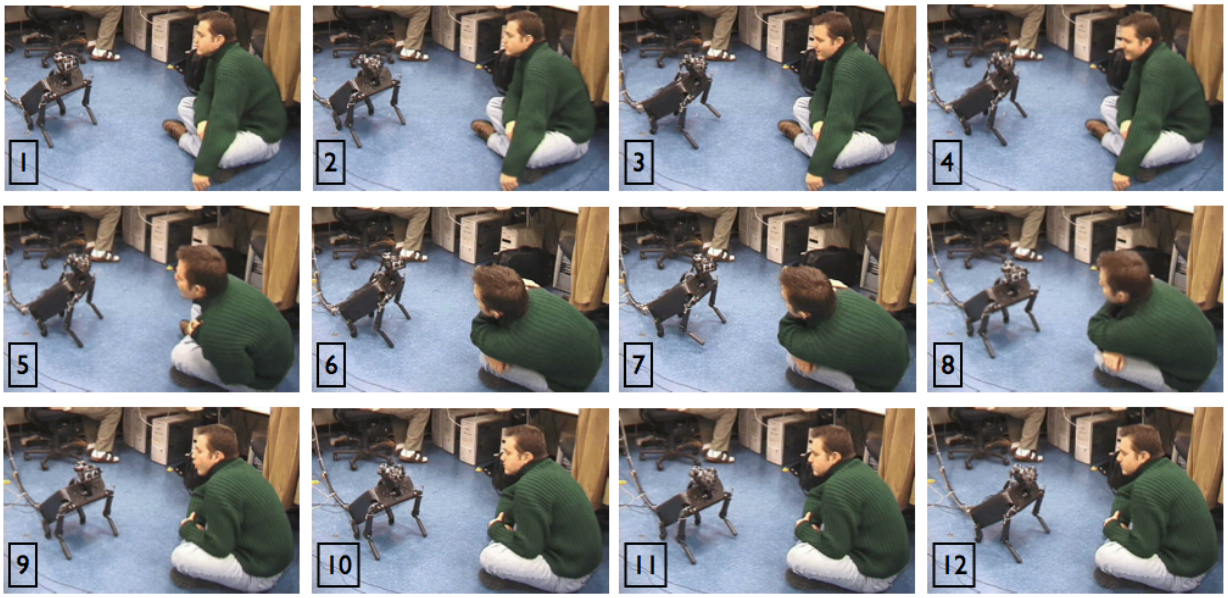

Figure 9-5: Full Orienting Responses to Affective Stimuli. If a stimulus is evaluated as being of affective significance, as it is the case with detecting a person for the robot Coco, the Head Orienting Responses do not habituate, but rather a full Body Orienting Response is produced. It is worth to note that the Head Orienting Responses occur in response to an auditory stimulus, when the experimenter calls the robot's name.

toward the approximate location of the detected sound, which facilitate the detection of the person and hence promote full body ORs. The same procedure is repeated in frames 9 thru 12. An important consequence of these results is the idea mentioned in Section 9.2, which suggested that the activity of an attentional system such as this implementation of the *Surprise* affect program, facilitates the activity of other affect programs such as the *Seeking* or the *Fear* systems which might implement a more appropriate solution to specific contingencies. In this scenario, this would mean that after the full body ORs are generated, the activity of the *Surprise* affect program transiently dissipates, and an affect program that deals with contingencies related with the detection of people might ensue.

9.4 Incentive Salience and Attention

Consider the notion of incentive salience we reviewed in Chapter 3 and which was ultimately implemented as part of the *Seeking* affect program. Based on these ideas, what does it mean to *really* 'want' something? If that "something" corresponded to a particular goal, such as obtaining a Ph.D., would this mean that one would work "harder" in order to obtain that goal? Would one also work more focused and less distracted in order to attain the object of desire? It seems reasonable to suggest this, at least from a colloquial perspective. Like the issues described in the previous section, these are separable functions that can be associated both with attentional and affective processing.

How might we implement such kinds of interactions? The first answer to our problem came in the form of the *Seeking* affect program. We had already implemented 'wanting' signals as part of this system (described in Chapter 8), now we only needed a way for these signals to be able to modulate and influence attentional processes. Along came the *Surprise* affect program, which as we reviewed earlier, deals with novel stimuli and attentional responses including ORs. If you recall, in Section 6.3.3 we described how affect programs could influence each other by sending excitatory or inhibitory input which was taken into account when computing the affective value of the affect program. Now everything was in place, and we could explore these novel ideas on the modulation of attention by incentive motivational processes, simply by connecting the *Seeking* and *Surprise* systems via their incentive ('wanting') and affective ('liking') value signals, respectively.

The end result of such ideas is illustrated in Figure 9-6. The incentive value of the *Seeking* system, which depends on the multiplicative effects of regulatory mechanisms (i.e., drive signals) was thus connected as an inhibitory input to the *Surprise* system. In vernacular terms, this means that when impending motivational needs arise (e.g.,

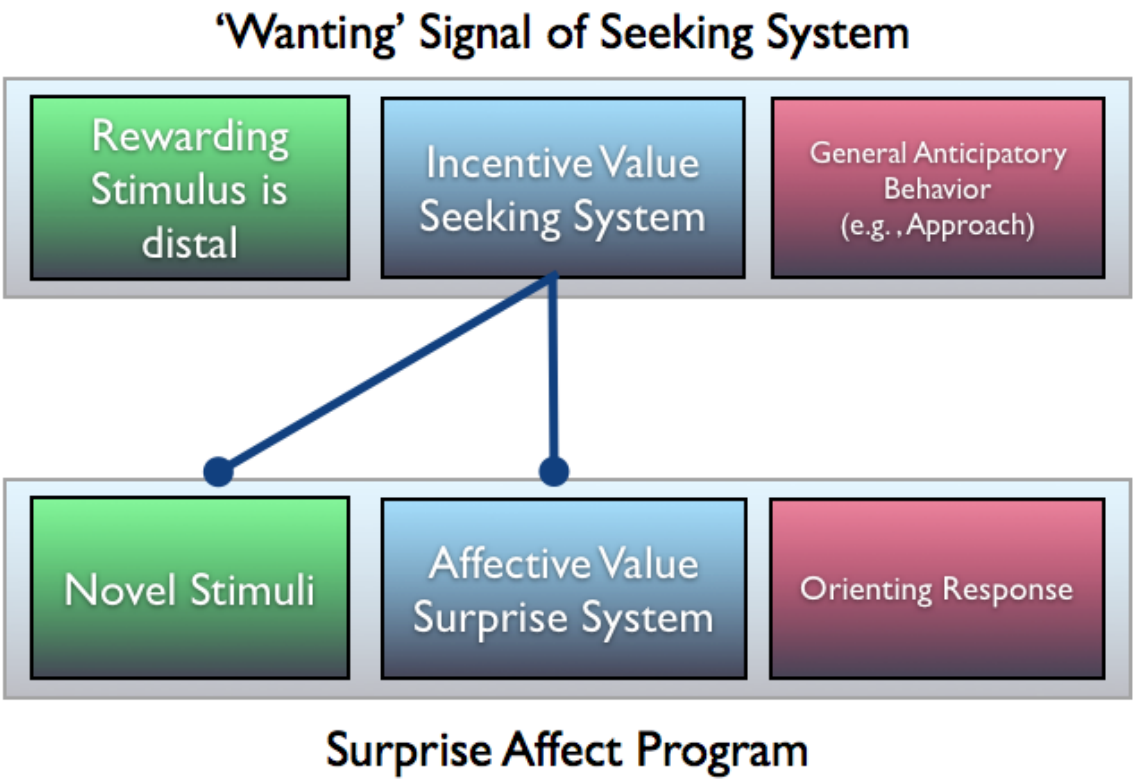

Figure 9-6: 'Wanting' Modulation of Attention. This figure illustrates how we implemented a simple yet powerful mechanism for the modulation of attentional processes via the 'wanting' signal of the *Seeking* affect program. With such mechanism, goals that have a high incentive value can be pursued by the robot with little, if any distractions, whereas goals that have lower motivational values will be pursued in a less coherent manner and external stimuli will cause distractions, triggering ORs via the *Surprise* affect program.

the need to recharge the battery in the case of our simulated robot), as long as an incentive is detected (e.g. the recharging station) which has sufficient incentive value, the *Seeking* system should promote approach responses while at the same time, it will also inhibit the activity of the *Surprise* system, hence reducing possible distractions which would correspond to the detection of novel stimuli in the robot's path toward the incentive goal.

This is precisely what happened! To test these ideas, we set up two different

scenarios in which the robot's battery level was set to be low and a recharging station was placed nearby, so that it would be easily detected by the robot's distal releasers. In the first scenario, we implemented the interactions between affective and attentional processes as described above, and watched the results of the behaviors this produced. In the second scenario, we artificially amplified these interactions by doubling the output value of the *Seeking* system's incentive value signal, which was also connected as an inhibitory input to the *Surprise* system, and also watched the behavioral results.

Figure 9-7 illustrates the first scenario. In the first frame, the incentive (i.e., recharging station) is marked by the top most red arrow, and two other neutral stimuli (a red ball and a purple block) also marked by the other two arrows, were placed in the same environment as distracting stimuli for the robot. The robot's battery level was set to a low value, which would generate positive multiplicative effects in the incentive value of the recharging station, and hence of the *Seeking* affect program, once its releasers detected this incentive. In frames 2 and 3, it can be seen how the detection of the recharging station triggers the approach response, which we now commonly associate to a preparatory response controlled by the *Seeking* system. In frame 4, the red ball is also detected, as is the purple block, but the saliency of the red ball triggers a head OR on the robot, mediated by the *Surprise* affect program. In frame 5, these head ORs where repeated and later followed in frame 6 by full body ORs. In frame 7, the position of the robot was such that the purple block was salient and also triggered a head OR. Finally, in frames 8 and 9, the robot resumes its approach response targeted at the incentive goal (i.e., the recharging station—yellow block). In frame 10, proximal releasers detect the yellow block and a consummatory response corresponding to a grasping behavior, as mediated by the *Seeking* program ensued, which ultimately increased the battery level to appropriate values and thus no impending motivational needs were present, which ended the scenario. It should be noted that this experimental scenario took place for several minutes and only the

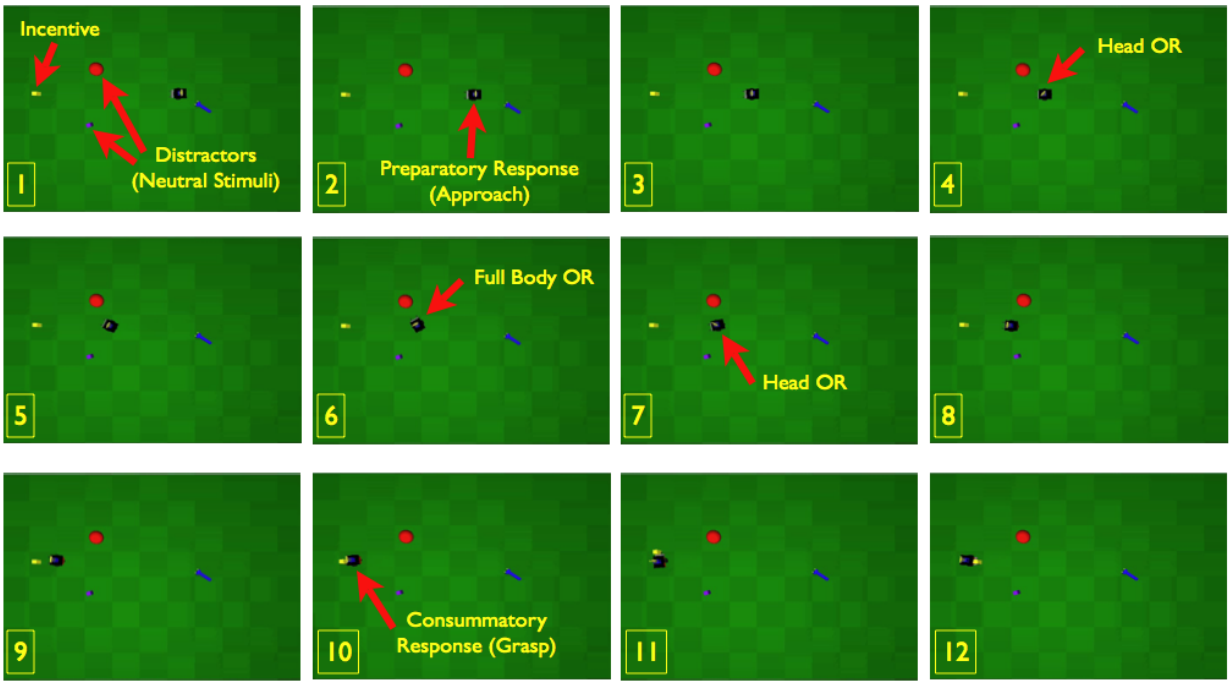

Figure 9-7: Modulation of Attention — Distraction. This sequence of frames illustrates the robot's attempt to pursue a goal, when multiple objects exist in the path of the goal (yellow block). In the case of a low incentive value for the goal, all other neutral objects trigger Orienting Responses (ORs) through the Surprise Affect Program, causing distractions which result in the robot's active pursue of the goal to be interrupted. Thus, the robot's behavior is not coherent and it dithers between approaching the goal or exploring the novel objects.

key events are represented in the illustration.

These results suggest that at normal levels of incentive value, the 'wanting' signal is not sufficient to inhibit the attentional distractions produced by novel stimuli in the path of the agent when pursuing incentive goals. As we described earlier, in order to test whether this idea was possible at all, we implemented another scenario in which the 'wanting' signal was amplified.

Figure 9-8 illustrates this second scenario. All conditions are the same as in the first scenario, with the only difference that the 'wanting' inhibitory signal coming

from the *Seeking* affect program and into the *Surprise* affect program was amplified by doubling its actual value. The first frame shows the robot in the same environment and in the same position with respect to all objects. In frame 2 it can be seen that the detection of the recharging station also triggers the approach response mediated by the *Seeking* system. In contrast to the first scenario, however, when the red ball is detected in frame 3, the robot comes to a full stop (a surprising event to us), and as if it hesitated, slowly reinitiated its path toward the incentive goal in frames 4 and 5. No head ORs were triggered by the detection of either distractor and in frame 6, the detection of the recharging station by proximal releasers elicits the consummatory grasping response controlled by the *Seeking* program, which finally increased the battery level to appropriate values and ended the scenario. In contrast to the first scenario, the amplification of the 'wanting' signal did have the anticipated modulatory effects in the *Surprise* affect program, thus demonstrating the possibility of modulation of attention by means of incentive salience processes. Furthermore, from a behavioral perspective, it was clearly shown how the robot appeared more "focused" in its pursuit for the goal, something that could also be assessed by the amount of time it took to achieve its goal, which was less than one minute.

Interestingly, recent evidence stemming from research on hyperdopaminergic mutant mice showed that these animals, which have a dopamine transporter (DAT) knockdown mutation that preserves only 10% of normal DAT, and therefore causes mutant mice to have up to 70% elevated levels of synaptic dopamine in comparison to normal mice, exhibited quite similar responses when pursuing goals in their environments. In a runway task, these mice demonstrated enhanced acquisition and greater incentive performance for a sweet reward. As described by the researchers, *"Hyperdopaminergic mutant mice leave the start box more quickly than wild-type mice, require fewer trials to learn, pause less often in the runway, resist distractions better, and proceed more directly to the goal"* (Peciña et al., 2003). As with our simulated

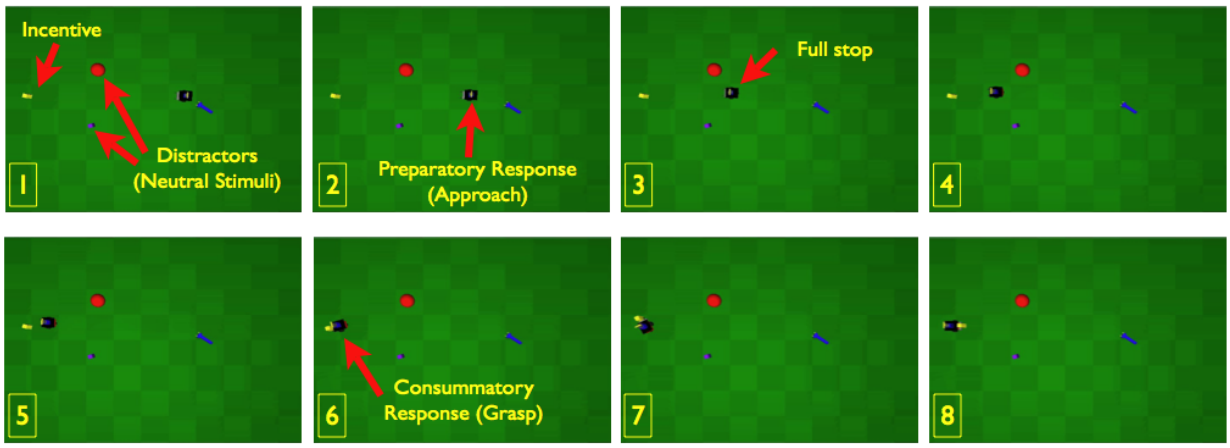

Figure 9-8: Modulation of Attention — No Distraction. This sequence of frames illustrates the robot's attempt to pursue an incentive, when multiple objects exist in the path to the goal (yellow block) and the *Surprise* Affect Program, which mediates Orienting Responses (ORs) to novel stimuli, is modulated via the incentive value of the *Seeking* Affect Program (The 'Wanting' Pathway). This 'wanting' signal was amplified 2x and connected through an inhibitory interaction to the *Surprise* Affect Program as illustrated in Figure 9-6. In this case, having a high incentive value did focus the robot's attempts and produced coherent approach behavior targeted at the incentive.

robot and its amplified 'wanting' signal, these observations seem to suggest that hyperdopaminergic mutant mice attribute greater incentive salience ('wanting') to rewards.

Chapter 10

Toward a Computational Theory of Affect

The computational model described in previous chapters ties together theories and evidence stemming from various disciplines that have a long-standing tradition in the study of emotion. This computational account introduced the notion of an affect program as the primary theoretical construct for investigating the function and the mechanisms of emotion, as they are instantiated in a variety of embodied agents.

As many researchers in Artificial Intelligence will attest, building systems that exhibit autonomous and intelligent behavior is a highly complex research endeavor for which many questions remain open. We believe that when we commit to such a task, with the main purpose of instantiating affect programs into these systems to understand the set of computational problems they face in their specific environments, we are in fact providing a very rigorous test of the theory of emotion upon which the computational account is based.

We would thus argue that in the same manner that we have drawn inspiration from other fields, these disciplines would also benefit from the lessons learned through an approach like the one described in this thesis, especially as they refer to the many

seemingly unimportant, but in fact quite relevant, issues raised by attempting to instantiate a mechanistic account of affects into an embodied system. To capture several essential features of emotional systems, an approach like this forces us to think about them at multiple levels of abstraction, as well as at several spatial-temporal scales. Furthermore, it requires that we address details that would otherwise be considered unimportant at the level in which most psychological constructs are thought of within these disciplines.

This chapter summarizes some of the lessons learned and freely speculates about interesting ideas that might be related in some way or another to the notions put forward in this thesis, and that as such would correspond to possibly interesting future research undertakings.

10.1 Summary of Contributions

The following are the main contributions made by this thesis:

- From a more theoretical perspective, one of the main contributions of this work consists on a reconceptualization of the notion of emotion into one that departs from traditional views that focus on the *experience* of emotion, and instead views emotions as functionally distinct processes–rather than states[1]–which implement specialized solutions to prototypical situations that organisms (or robots) regularly face in their environments.

- We presented a unified computational framework for the study of emotion that accounts for different affective phenomena, including a variety of emotions

[1] I thank Roz Picard for rightly suggesting that "snapshot views" of these processes could represent instantaneous affective states as well, and could be treated as such when these descriptive levels are useful.

and mood-like activity, and integrates these with several notions traditionally deemed to be integral components of intelligent behavior.

- As the main component of this framework, we introduced a novel computational construct for an *affect program* as a biologically plausible abstraction for emotion. The primary function of affect programs is to mediate, control and synchronize the activities and interactions of several subprograms, including those that govern perception, attention, physiological regulation, goal selection, motor control, expressive social communication processes, action selection and learning, and so forth. Each of these affect programs establishes a mode of operation for the robot, which involves the coordinated adjustment and entrainment of these subprograms (responses) so that the whole system exhibits coherent behavior as a response to the confrontation with specific eliciting situations.

- We presented a model for incentive salience, which attributes motivational properties to stimuli and actions that signal the occurrence of events of emotional (biological) significance. An incentive salience approach contrasts with other views that propose that reward or incentive learning, as mediated by the brain's mesolimbic dopamine systems, is based upon global teaching signals that code for the errors in the prediction of reward. These views have found further acceptance in the neurosciences given that elegant computational counterparts, such as the reinforcement learning models, seem to work in a similar manner. However, an incentive salience approach, such as that proposed in this thesis, provides an alternative explanation for the activity of these same brain systems and through simple and localized learning rules, together with the organizational principle that separates action into preparatory and consummatory behaviors, can account for some of the evidence seen in experimental paradigms. Something that reinforcement learning models, at least in their original form, cannot

account for.

- We proposed an agent architecture that follows an affective-based decomposition and which provides a novel alternative to the control of intelligent robotic systems. In this approach we suggest that a different organization of action is pursued, one based not upon the desired external behaviors of the robots, but rather on the set of prototypical fundamental situations that the robots will encounter, and view these as affective situations for which a set of coordinated responses can provide solutions, much in the spirit of these biological schemas we have referred to herein as the *affect programs*.

- Finally, we proposed a multi-stage model of affective learning that relates evidence from psychology and neuroscience regarding classical paradigms of associative learning and bridges these notions with a possible computational substrate, in the form of affect programs.

10.2 Some Lessons Learned

In AI, we have long strived toward creating systems that can exhibit autonomous, and intelligent behavior. Depending on the approach taken, these efforts vary from symbolic computational approaches, to those that deal with neural networks, genetic algorithms and robotics. As philosophers have long suggested, autonomy is, and should be, a central tenet of AI, just as it has been with the philosophy of the mind (Dennett, 1987; Griffiths, 1997).

Traditionally, we have not taken advantage of the many fruits that a research agenda into computational emotion would provide. Partly due to previous cognitive and symbolic efforts that looked deceptively simple to achieve and thus gave the impression that these were uninteresting challenges to pursue, and partly because

our own notions and uses of these psychological constructs led us to work at different levels of abstraction that not always produced fruitful results or which seemed to be resolved questions altogether.

10.2.1 Mechanistic Precision and Psychological Constructs

Throughout this work, we have suggested that affect, like any other construct, can be studied from many different standpoints and at many different levels of abstraction. Given our design principles (described in Section 6.1), we follow a computational perspective that attempts to elucidate these issues at the mechanistic level. We are interested not only in accounting for high-level constructs, but also for determining how these could be implemented in robots that are situated in real environments.

In following this approach, we have faced an interesting predicament related to how certain theoretical constructs, such as the notion of reinforcement and reward, or the general conception of how Pavlovian conditioning works, which are widely used in our high level discussions and theories, might mean completely different things or cease to exist altogether when considered at a much lower level of abstraction.

Consider for instance the notion of *reward*. Rewards, like emotions, motivations, and many other "big words" have come to elicit many different meanings and thus act like the "suitcase" terms that Minsky (2006) and others have talked about before. Rewards are a central tenet in all of our discussions about motivation and they are, no doubt, an essential component of many of the psychological explanations we use for intelligent behavior and learning. Without the notion of reward, contemporary learning theories would be meaningless, as would their computational counterparts. Although the notion exists since earlier times, Thorndike's and Pavlov's experiments certainly "stamped-in" (forgive the pun) the idea of rewards in our current discussions about learning. In Pavlov's classical conditioning paradigm, for instance, we have

come to use the term "reward" when referring to the Unconditioned Stimulus (US) that evokes Unconditioned Responses (URs). From these constructs many have taken to believe several things: for one, that USs are innately detected, which Pavlov never intended to mean, and in fact he clearly stated otherwise; and second, that we can use these high-level constructs without paying much attention to some of the details they entail. Is food a reward? We believe so. Is food a US? It must be, if we consider the experimental preparations that are widely used in the behavioral neurosciences. But is this really the case when we think of these constructs at a mechanistic level? Food cannot be reward, at least not in its natural form, unless it has been signified as such (through associative processes that link the sensory properties of the foodstuff with nutritional features that are innately rewarding) by the organism that is consumming it. Food pellets like those given to rats in behavioral experiments, or a slice of pepperoni pizza that we might have for lunch, are not natural rewards. They act like ones now, but they came to do so by a necessary association via the very basic learning and affective processes that we attempted to elucidate in this work.

Thus, we would argue that only by addressing these kinds of constructs at specific levels of abstraction will such issues arise. These situations are not typically faced by those who attempt to model the same phenomena at higher levels of abstraction. These other models abstract much of these issues, and thus decision-making processes such as action selection arbitration become not a matter of conflicting actuators and choosing, executing and controlling the selected response, but rather selecting one of several possibilities represented by an input vector, for instance.

These issues are certainly not novel, as we are clearly not the first ones attempting to build robotic creatures that can be endowed with mechanisms and processes like the ones described in this work. However, we shall make an effort to call attention to them as we believe they are of high importance to a general understanding of the complex processes that are involved in intelligent behavior.

More than forty years ago, for instance, W. Grey Walter faced similar predicaments while attempting to recreate these phenomena in his mechanical "tortoises". While attempting to have his *Machina speculatrix* replicate conditioned reflexes, the high-level abstraction of conditioning, as used by Pavlov and his disciples of the time became quite problematic for his low-level needs, as he intended to recreate such constructs in a physical machine made out of tubes, relays and simple circuits[2]. Walter attempted to seek out guidance by constructing a working model of how Pavlovian conditioning could work, as explained by the high level psychological constructs used at the time. He already had a model of a machine that could move via a reflex which was responsive to light. As Walter described, *"it was a simple addition sum to provide a second reflex circuit to be made responsive to sound."* The further addition of a conditioning association between the two reflexes turned out to be no simple operation (as perhaps it could be in a higher level model) but in fact would require a higher level of precision. His attempts fruitfully ended in the creation of a Conditioned Reflex Analogue (CORA) which could be demonstrated on the bench in his *M. docilis*, but which would require even much more precision if it were to be included in *M. speculatrix*, which was a mobile machine (Walter, 1961).

Like Walter's "tortoises", our computational models for controlling robots are intended to include parsimony, goal-seeking, and incorporate positive and negative tropism. It seems clear to us that even though our current technology is advanced enough to build fairly complex robots with humanoid form, that can perform tasks in dynamic environments, the nature of these tasks make them seem much more complex than the simple overt behaviors of "approach", "avoid", "like", and "dislike", that are possible with our models. However, we would argue that we are still far from elucidating the very basic mechanisms responsible for such kind of behaviors and their

[2]I am indebted to Rod Brooks for pointing out to me Grey Walter's similar predicaments and directing me toward his fascinating work on mechanical tortoises.

organizational principles. However, it should also be clear that if our goal is to mimic behavior, then we have already achieved this, thus the engineering goal with respect to these simpler tasks has already been attained. Our contention here is more related to obtaining an understanding of the general organization of such constructs from scientific goal of obtaining an understanding of the mechanisms underlying affective processes, and as such these simple behaviors are of high interest to us.

Let us end this argument with a much more succinct example. By forcing precision onto psychological constructs such as Pavlovian conditioning, other processes have been elucidated, including that of Evaluative Conditioning (Houwer, Thomas & Baeyens, 2001), which corresponds to an associative process that even though might be directly related to the Pavlovian paradigm, exhibits marked differences and might ultimately be the process by which a stimulus such as a food pellet becomes to mean a "reward" later on, once it has been associated with affective and motivational processes like those described in Chapter 8.

10.2.2 Meaning Machines

Based upon this work, we have also come to think of emotions as valuation engines that effectively tag stimuli in the world by deciding which events are significant and which are not, and in what sense they are so (e.g., in the joyous, the infuriating, or the fearful sense), thus providing meaning of the world to the organism. As such, emotions can be seen as significance or meaning machines, that bias action, modulate perception and attention and which lay the foundation for other more cognitive processes by giving affective meaning to the world contingencies.

10.2.3 Action Comes Before Abstraction

Action is fundamental in the notion of autonomy, which as we alluded to before, has been considered paramount to the problem of intelligence. Our primary means for dealing with the world are largely motivational, and from our viewpoint, mainly regulated by our affect programs. From our psychoevolutionary perspective, emotions came about as a set of responses, both preparatory and consummatory, that facilitated solutions to life's fundamental situations. Our basic approach to engineering affect, and hence to controlling robots, suggests that instead of building general purpose learning systems we focus on how these affect programs can reduce the learning space via local learning systems that become active once there are contingencies of importance to the organism (or robot). Likewise, we propose to build control systems that are not based on broad organizations of behavior, but rather which are organized into different pathways, preparatory and consummatory, both for perception and for action. Finally, we propose to create effective action programs that can be released by commonly faced situations and tasks. These action programs are nothing but our affect programs, which can later be used as schemes for higher more abstract processes.

Could the same programs (i.e., the same mechanisms for assigning affective and incentive value, as well as those for synchronizing and coordinating responses) be combined with more abstract releasers and responses? If releasers are not of the material kind, based primarily on our perceptual systems, but rather based on abstract thoughts and the learned relationships between stimuli, could we not start to elucidate the meaning of secondary or more cognitively produced emotions? The *Seeking* system, as it currently stands, is integrated with motor behavior which results in approach and exploratory responses. If this same control programs were to be integrated with abstract thought, might they *seek out* solutions to imagined situations based on

imagined or abstract actions? Might a mechanism such as this be the basis for cognitive constructs such as planning or even creative processes? Highly speculative, of course, but definitely worth thinking about.

Consistent with these ideas, though, is the evidence that stems from subjects who exhibit pathological cases in which these kinds of systems become overactive and the "imagination" of such individuals seems to run amok, determining causality where only simple correlations might exist and ascribing reality to fantasy (Ikemoto & Panksepp, 1999).

10.3 On Engineering Affect: Related Work

To explore both what we can learn from our understanding of affect programs and to suggest how this understanding might be improved by our work in computational modeling, we distinguish between different kinds of computational models of emotion. Depending on the goals and the nature of the model, we divide these efforts into *shallow* and *deep* models of affect.

10.3.1 Shallow Computational Models

One of the earliest models that included a notion of emotion, from a symbolic classical Artificial Intelligence perspective, was the work of Colby in his PARRY program (Colby, 1974). PARRY was a program with which one interacted via a computer terminal. PARRY was designed to respond to the users with exaggerated affective terms, and thus the idea was to model a simplistic account of human paranoid schizophrenia.

Like PARRY, other text-based work related to emotion, or at least to the understanding of narrative accounts that included emotional terms was Dyer's BORIS program (Dyer, 1982). Also based on symbolic approaches, BORIS was a narrative analyzer that could make simple inferences regarding the affective terms and states

that made part of these stories.

Most of these models were essentially reasoners that could analyze emotional terms and thus make inferences regarding related emotional states. These systems treated emotions simply as labels that had no meaningful connections to many of the issues related to affective phenomena, and in no way accounted for how these emotional labels were generated in the first place, what would be their eliciting conditions, their components, or how would they interact with other constructs and processes.

From these efforts, multiple models came about, including Clark Elliott's Affective Reasoner (Elliott, 1992), that attempted to alleviate some of these shortcomings. The Affective Reasoner was based on the model by Ortony et al. (1988), which was (and still is) a predominant cognitive theory of emotion. Elliott's model allowed users to interact with simple characters enhanced with multimedia capabilities (e.g., music), who appeared to have different emotions and could *reason* about them, hence the name of the model.

Other earlier accounts, which were embedded in more complex agent architectures that included functional systems for perception, planning and action, included the work by Reilly (1996) and the Oz agents. Part of a wider project to create emotional agents at CMU, Reilly's work would focus primarily on the exploitation of emotional expression to create believable agents. That is, on the use the expressive signals of affective processes in order to suspend disbelief in humans that would interact with these agents in short-lived experiences.

Most of these models were based upon the cognitive appraisal theories reviewed in Section 2.2.1. The main reason why these were the predominant theories upon computational models were based was that the problem could easily be reduced to a fairly direct mapping between the cognitive appraisal taxonomies and production rules that implemented them in a symbolic fashion. One has to be careful in ascertaining what is it exactly that these models account for. For instance, in the case of Elliott's

Affective Reasoner, the implemented agents seem to be able to *reason* about the kind of emotion they would be in, given a particular situation. The important thing to note here is that the goals and application for which these models are built are an important indicator of the kind of model that is provided. In the case of models that are built for entertainment purposes, such as Reilly's (1996) and even Elliott's (1992), the main purpose is not to understand affect from a computational perspective, and thus these cannot be evaluated as such. If that were to be the case (i.e., use them to understand affect), we would argue that these models would not provide much information regarding what emotions are and what are the kinds of computational problems they involve, but they would provide information regarding how people interpret and reason about emotional situations.

In a similar fashion, other models have been developed as part of robotic systems (or their control architectures to be more precise) that do have to sort complex computational tasks in their environments, which would require the set of solutions for which we believe emotional systems were adapted. At a glance, these models seem more complex and are in fact linked to the robot's architecture in interesting manners. However, with the exception of a few, most of these models have restricted their use of emotional processing to that of directing expressive behaviors that would "trick" humans into suspending disbelief (albeit in the short-term) and thus facilitating human-social interactions. Indeed, a very interesting aspect in which emotions are involved (and contribute to its regulation) corresponds to that related to the signaling of internal states that are essential to regulate social communication and interaction. This is certainly an exciting area that has recently drawn attention in robotics research and many important questions in this respect remain open. However, as we have argued throughout this thesis, we do not believe emotions evolved as superordinate programs simply to modulate social communication. Rather, we see their main function to be that of directing the activities and interactions of subprograms that

regulate perception, attention, motivation and goal choice, action selection, learning, motor control, and so on, as part of the implementation of solutions that have provided adaptive in recurrent life and survival-related fundamental situations.

This type of models in which systems can mimic the appearance of having emotion, or which can recognize and respond to emotion in simple ways, but otherwise do not involve the set of computational issues involved in affective processing, are what we refer to as shallow models of affect. Let us further clarify that by "shallow" we do not mean for any negative connotation to be ascribed to these models. We simply mean that their goals and applications should be considered when discussing about their explanatory power, as they are not developed with the purpose of understanding affect from a computational perspective.

10.3.2 Affect-Related Models: Reward Learning

The hypothesis that dopamine signals between neurons are an important component in the neural substrates that causes reward learning has gained great prominence in recent years. This view puts forward the idea that the activity of mesolimbic dopamine neurons acts as a global, teaching signal that modulates learning processes in order to *'stamp in'* and associatively reinforce new associations between *stimulus-stimulus* or *stimulus-response* contingencies.

A major appeal of this learning hypothesis for dopamine function results from the realization that elegant computational models based on reinforcement learning (Sutton & Barto, 1981), which were not originally developed with such explanatory purposes in mind, do fit the data stemming from electrophysiological studies on the phasic activity of dopamine neurons as they may code for errors in the prediction of rewarding events (Schultz et al., 1997; Montague, Hyman & Cohen, 2004; Schultz, 2006).

Although these models are *not* models of affect in the sense posited in this thesis, they certainly are related to specific affective processes. These models attempt to account for some of the constructs related to incentive or reward learning, at least with respect to the 'liking' and learning components of reward, described in earlier chapters. As such, a brief review of their main features and their relation to that relevant aspect of the framework proposed herein (i.e. that which relates to the *Seeking* affect program) is in order.

Reinforcement learning models based on temporal difference (TD) learning, such as those proposed initially by Sutton (1988) and Sutton & Barto (1998), and which are based on earlier prediction error models that suggested a plausible progression of associative learning (Rescorla & Wagner, 1972), are the most widely adopted computational accounts for both Pavlovian and instrumental conditioning, and now for dopamine activity as well, as suggested above. These models, albeit elegant and practical in their computational nature, do not account for several key findings in the rich literature of such phenomena.

Thus, at least with respect to explaining affective phenomena such as those described in this thesis, reinforcement learning approaches do not account for motivational influences on reward learning, nor do they account for influences in attentional processes as evidenced by large bodies of work (for a review see (Dickinson & Balleine, 2002)).

In particular, TD-based models do not take into consideration how motivational shifts (like those described with physiological mechanisms like hunger or thirst) exert immediate effects on behavior and learning. A recent and notable extension to such models has been proposed by Dayan & Balleine (2002) in an attempt to compensate for such explanatory deficiencies. This model includes extensions that relate the pursuit of rewards to two different predictive systems: one for Pavlovian conditioning and one for instrumental conditioning, and which take into account some motivational

influences in incentive learning. The model, however, as an extension to TD-based algorithms, still suffers from the rigidity in the selection of actions as they relate to the pursuit of rewards, a notable feature of this type of model.

In terms of the organizational principles suggested in our framework (i.e. consummatory and preparatory pathways for behavior), we should mention that a signature finding of Pavlovian conditioning is the notion of conditioned responses, which implies more than just the predictions of rewards. The extent of the Pavlovian conditioning paradigm reaches the behavioral level in order to consider the consequences of such predictions, as evidenced by learned or conditioned responses, such as the approach response to the purple block that signals the occurrence of the yellow recharging station in the case of our simulated robot. The Pavlovian consequence of such signals is a preparatory behavior such as the approach response (undeniably observed throughout experimental studies). Thus the organism (or robot) will approach the signal (be it a light, the approximate location for a sound or another object) regardless of whether or not doing so is optimal (or even functional) for obtaining the reward[3]. In TD-based models, the performance of such responses is not accounted for, and much less so when they could imply the reduction of reward.

Furthermore, traditional models do not account for the important attentional processes involved in classical conditioning paradigms. Our framework suggested a possible interaction effect between the 'wanting' signal particular of incentive salience processes, and the modulation of attentional processes in the form of inhibition of selective Orienting Responses, which resulted in fewer "distractions" when our robot attempted to pursue a goal. At a much lower level of attentional processes, however, Dayan and colleagues have shown how Kalman filters could be used to extend attentional processes beyond the issue of managing limited attentional resources and

[3]See the work on training chicks to access food by Hershberger (1986) for an interesting arrangement that demonstrates this issue.

suggest how such mechanisms might work to determine what it means to be an important and relevant stimulus from an affectively significance perspective (Dayan, Kakade & Montague, 2000).

We should mention that most of the models reviewed here are still based on the same principles of temporal difference learning, and while these extensions account for specific phenomena (e.g., motivational influences, attentional influences) none of these models provide a comprehensive account (at the architectural-level) that integrates all of these features as we have proposed in the current framework. In all fairness, however, this is not the goal of such models, as they are not models that attempt to account for the synthesis of affect, but rather focus mainly on the idea of reward learning in classical paradigms. As such, they would only be comparable to the *Seeking* affect program described herein, which in contrast to these reinforcement learning approaches, is based on the notion of incentive salience and can offer an alternative explanation for the system-level function of dopaminergic systems in the brain, as proposed by others (Panksepp, 1998; Ikemoto & Panksepp, 1999; Berridge & Robinson, 2003; Berridge, 2006; Robbins & Everitt, 2006; Salamone & Correa, 2002; Salamone, Correa, Mingote & Weber, 2005).

A very interesting and recent work by Ahn & Picard (2006), also attempts to model the activity of such *Seeking* system, albeit at a higher-level of abstraction. This work provides a probabilistic account for affective-cognitive interactions, and uses affective signals in order to influence decision-making processes, in a way similar to that proposed by Damasio's somatic marker hypothesis (Damasio, 1994; Damasio, 1999). Although their affective models are limited to binary states of "feeling good" or "feeling bad", this work is one of the few attempts to integrate both extrinsic and intrinsic motivational inputs in the processes of learning and decision making. It would be interesting to see how a model such as this would scale if it were to consider a wider span of affective processes or states, such as those described in this framework.

We believe these accounts are complementary perspectives on a similar process. While our work proposes a biologically plausible mechanistic account for affect programs, including the *Seeking* system, their account would account for higher-level cognitive interactions, but both approaches follow similar assumptions with respect to the activity of dopaminergic systems and the existence of 'wanting' pathways.

Finally, McClure, Daw & Montague (2003) have suggested a model for incentive salience that relies on the same associative learning mechanisms of temporal difference learning methods, but which does not take into consideration any of the motivational factors that are at the essence of the incentive salience ideas. It essentially equates incentive salience to associative predictions of reward. In that sense, it differs significantly from accounts of incentive salience as an integrative motivational process in which physiological states have multiplicative interactions that contribute to determining the incentive value of stable learned signals such as it its proposed by Toates (1986) and Berridge (2004), and which is precisely modeled in our framework.

To close these ideas on the learning hypothesis of dopamine activity, it should be mentioned that recent evidence indicates that dopamine is neither necessary, nor sufficient, to mediate changes in hedonic 'liking' of rewards. Likewise, other recent evidence suggests that dopamine is not needed for new learning either, and also not sufficient to directly regulate learning processes through teaching signals based on errors in the prediction of rewards. Instead, growing evidence indicates that dopamine does contribute causally to incentive salience processes, as used in a framework like the one proposed here. Dopamine seems to be necessary for normal 'wanting', and in fact sufficient to promote cue-triggered incentive salience (Berridge, 2006).

As a general overview, table 10.1 summarizes the main approaches to synthesize affect from a computational perspective, as well as those models that in spite of not being proposed for the synthesis of affect, are related to a particular component of affective processes (i.e. reward learning) and as such provide interesting contributions

to the field.

10.4 Future Work

Affective processes involve such a wide variety of mechanisms and phenomena, that clearly one single framework, even if it is composed of many subprograms for attention, behavior, learning and so forth, leaves behind much to be explained, and many interesting possibilities for extensions and future work. This section reviews some of the most interesting ones, from our perspective, and which seem as natural extensions of the work presented here.

10.4.1 Social Emotions

While there is considerable evidence to support a psychoevolutionary approach to emotion like the one presented here in the form of affect programs, a different account might be needed to explain social, more cognitive emotions, such as *Guilt*, *Jealousy*, *Shame*, and other complex psychological constructs such as *love*. The approach taken in this work has been that through learning and developmental processes, some of these emotions could be accounted for, but clearly other affect programs exist in the mammalian brain that were not studied in this framework and which might be essential to build such developmental strategies. We are referring to affect programs such as those for *maternal care*, which might be the basis of attachment processes, and *separation distress*, which might be intimately related to affiliation processes and social phenomena, among others. In any case, other social accounts for emotion (without recurring to extreme views that discount biological determination) might certainly be appropriate and useful approaches to explore from a computational standpoint (Frank, 1988).

Table 10.1: Comparison of Models of Affect or Affect-Related Phenomena.

	Cognitive Models	**Dimensional Models**	**Reinforcement Learning Models**	**Affect Programs – Cathexis**
Affective Phenomena	- Emotions as states - Labels of primary emotions - Labels of secondary emotion - Non-emotional states	- Primary emotions - Secondary emotions - Mood-like states - Non-emotional states (e.g. sleepy)	- Not models of affect, but include affect-related constructs - Rewards modeled as values	- Emotions as processes - Primary emotions - Secondary emotions suggested as learning processes - Emergent emotions - Moods as tonic-level activation of affect programs - Parameter-based simple account for temperament
Affective Behavior	- Mainly directed at signaling of emotional state (i.e., Emotional expression) - No organization for behavior suggested	- Primarily for signaling emotional state (i.e., Emotional expression) - No organization for behavior suggested	- Many possible behaviors - Not "motivated" by affect, but influenced by "reward"	- Action is at center stage, determined and regulated by affect - Separation of sensorimotor pathways into preparatory and consummatory behaviors - Emotional expression - Responses other than relational behavior possible
Affective Learning	- No specific learning interactions addressed - Some include "affective tags" (i.e. labels of emotion assigned to objects)	- No specific learning interactions addressed by these models	- Temporal difference learning - Reinforcement based on errors in prediction of reward General conditioning	- Nonassociative learning (e.g., habituation) - Evaluative conditioning - Incentive learning
Affective Interactions	- Appraisals theoretically involve complex cognitive processes that interact to produce affect - Computational models do not account for other interactions, but see (Thagard, 2006)	- Theoretical models may include interaction with bodily processes and attention (through an arousal dimension) - Computational models do not account for these interactions (but see (Ahn & Picard, 2006) for an example that uses the valence and arousal dimensions to account for motivational and cognitive interactions)	- Do not account for motivational, attentional, or any other interactions (but see (Dayan & Balleine, 2002; Dayan et al., 2000; McClure et al., 2003) for extensions to this type of model that attempt to account for some motivational and cognitive interactions)	- Motivations directly influence affective processes and learning - Modulation of attention by incentive value - Stimulus vs. goal-directed action

10.4.2 Misbehavior

Under certain conditions, organisms exhibit a variety of behaviors that defy explanation in traditional drive reduction theories or response reinforcement models such as the ones reviewed above. Such behaviors are often called "misbehaviors" as they appear to be unrelated to the contexts in which they occur, and inappropriate with respect to the stimuli that are present in such events.

These phenomena include responses such as *autoshaping*, *misbehaviors*, and *displacement* behaviors, reviewed by Bolles (1972) as part of his motivational theories, but initially identified in the late 1960's by others (Breland & Breland, 1961; Brown & Jenkins, 1968; Williams & Williams, 1969). In a classic example, racoons being trained to put "money" (coins) into a metal box would exhibit non-reinforced behaviors in which the racoons started rubbing the coin against the inside of the box, taking it back out and clutching it firmly for several seconds, before finally letting it go. This misbehavior occurred as if the racoons were "washing" the coins, much as they wash food before eating it. Interestingly, these misbehaviors became worse as time went on, in spite of non-reinforcement by the experimenters. Similarly, pigeons that were presented with light signals followed by freely available food, came to start pecking robustly at the light signal whenever it came on, despite the fact that they had never been reinforced to do so—rats exhibit similar behaviors, gnawing and biting the lights when presented under similar conditions. This kind of behavior is known as *autoshaping*, due to the similarities of this phenomenon to instrumental or operant conditioning (also referred to as "shaping"), in which an animal's response such as pulling a chain or pressing a lever is increased by the reinforcement given by the experimenter. As suggested by Berridge (2000), however, autoshaping is purely an incentive process that requires no reinforcement whatsoever. Thus, there is no rational reason for the animal to work for a reward, less so to peck at or bite the sig-

nal for it, at least when attempting to explain such phenomena from a reinforcement perspective. Finally, displacement behaviors are also commonly reported, including *schedule-induced polydipsia*, in which animals drink exorbitantly large amounts of water for no apparent reason while under certain eating (not drinking) training conditions, or self-grooming and scratching, displayed when an animal apparently has a conflict between different motivations, such as the desire to pursue an incentive, while at the same time being fearful of that incentive.

As early as 1972, Bolles had suggested that these misbehaviors could not operate by the response reinforcement and drive reduction theories that were the darling theories of the moment. Still now, our models for action selection and reward learning, and the generally accepted models for classical conditioning, such as reinforcement learning models, cannot account for these phenomena and do not even bother considering explanations for it[4]. There is a motivational explanation for some of these misbehaviors, however, and it relies precisely on the notion of *Incentive Salience* proposed by Berridge & Robinson (1998), and which provides the theoretical basis for our computational approach to incentive learning, as described in Chapter 8.

In our framework, misbehaviors such as those exhibited by the racoons, and autoshaping can be accounted for as they would be mediated by the 'wanting' signal related to the incentive value processes discussed in Chapter 7 and Chapter 8. The S-R association that is learned between the neutral stimulus and a consummatory behavior via a three-factor hebbian rule, is modulated by this 'wanting' signal. In some cases, depending on the contingencies and the learning rates associated to these hebbian rules, the 'wanting' signal can amplify this S-R association (which is usually learned at a slower rate than other associations) and make the neutral cue (CS)

[4]I am indebted to Patrick Winston for drawing my attention to misbehaviors, including autoshaping and displacement behaviors which led me to think about possible ways a framework like the one described here could account for these phenomena.

to be highly 'wanted', in such way that a consummatory behavior (e.g., chewing or gnawing), which would otherwise be associated to a specific goal stimulus (e.g., food), would now become associated to the neutral cue and thus activated by proximal releasers that provide sensory-specific information about such stimulus. In this sense, our robots would "chew" the lights (neutral stimuli) that signal the occurrence of the rewarding stimuli. In the case of our simulated robot, the "grasping" behavior would also apply to the purple blocks which are neutral and usually only trigger preparatory responses (i.e., approach behavior).

In other words, the neutral stimuli (CSs such as a light or the purple block in our simulated world) that predict food to a hungry pigeon or a robot that needs to recharge its battery, become attractive, potentially "consumable", and possibly even 'liked' food-like or rewarding objects (Berridge, 2006). They elicit approach responses like those demonstrated in Chapter 8 and even consummatory behavior that would ordinarily be directed to the incentive stimulus itself. All this activity is modulated by motivational processes (e.g., regulatory mechanisms such as hunger or recharging need) in such a way that they have multiplicative effects on these object's attributed incentive value, but when these motivations are no longer present (i.e., the animal is no longer hungry or the robot does not need to recharge its battery), the same signal is simply a predictive signal, lacking any motivational properties.

Displacement behaviors are more elusive, however. The specifics of how adjunct or displacement behaviors are produced remain an open question, but what we do know is that they are under the control of the same dopaminergic systems that mediate the neural circuits involved in the seeking system, and thus are directly related to the interactions between these same 'liking' and 'wanting' pathways that are at the essence of our model for incentive salience. Robbins & Koob (1980) showed a dissociation between water-deprived drinking and polydipsia in rats. Lesions to the dopaminergic systems projecting to the nucleus accumbens eliminated the adjunct drinking behavior

in the lesioned animals, yet they could still exhibit drinking behaviors when water-deprived. This suggests a possible motivational origin for such behaviors, but the answer remains at large and it is not clear that a model such as the one presented here would account for displacement behaviors.

Another possibility for the occurrence of such behaviors might be at the action arbitration level. It has been theorized that displacement behaviors occur as the possibly conflicting motivations neutralize each other and in the process, this neutralizing effect disinhibits the displacement behavior which thus becomes active. The details of such activity remain a matter of speculation, at best. Notwithstanding, from a mechanistic perspective, for our model to account for such possibility the computation of behavior value and the arbitration scheme would have to be modified from an excitatory winner-take-all strategy to an inhibitory looser-take-all one. This would be consistent, nonetheless, with current thinking on basal ganglia mediated action selection, which is thought to involve a double inhibition (i.e., disinhibition) of tonically-active behaviors.

In any case, the implementation and further accounting for misbehaviors is certainly an interesting question which remains open and quite amenable for future work as an extension of the ideas presented in this thesis.

10.4.3 Incentive Value and Vigor

Our notion of incentive value is a simple one, as it only considers the multiplicative effects that physiological mechanisms such as those regulating hunger or thirst, would have over the incentive value of rewarding events or of those neutral stimuli, once incentive salience has been attributed to them. An idea related to the 'liking' and 'wanting' systems is the notion of *vigor* which describes the strength or rate of responding that organisms exhibit when working for rewards. In other words, this

notion refers to how "hard" organisms would work for rewards, which, from our perspective would be directly related to how much these rewards are 'wanted'. Thus, incorporating the notion of vigor to account for the simple, and well-established observation that hungrier animals would work harder (e.g., circumvent more obstacles) for food than satiated animals is a natural extension of this work. Incorporating a notion that includes the costs of executing certain actions could certainly make the computation of incentive value a much complete one. To this end, one could look at the quite recent work by Niv, Daw, Joel & Dayan (2007) which attempts to account for such ideas and relate them to the tonic activation of dopamine neurons.

10.4.4 Other Interactions

Investigating other affective interactions is certainly a matter worth pursuing:

- It would be a natural extension to this work to consider how affective processes interact with other constructs and processes that were theoretically included in our framework, but not actually implemented, such as the influence of affect in low-level attention processes like the ones described in earlier sections (Dayan et al., 2000), or how a process like that of incentive salience could account for superstitious behaviors and attentional regulation when affective significance is attributed to neutral stimuli through processes that were coincidental but not really contingent.

- Likewise, there are well-established relations between affect and memory storage and retrieval. The proposed framework only hints as to how associative representations, mediated by affect, might trigger memory-congruent events, but no real implementation or model is in place, beyond the associative learning rules described earlier.

Like these, there are multiple kinds of interactions, some established, some hypothetical, that would provide interesting avenues of research as an extension of this work. Some examples would include affective-immunological interactions, affective-cognitive expectancies in the control of action (Dickinson & Balleine, 1993; Dickinson & Balleine, 2002), and the highly complex and controversial implications of emotion and consciousness (Panksepp, 2005; Ellis & Newton, 2005)

10.5 Afterthought

One of the most interesting issues that we believe rises from this work is the reconceptualization of affect and how this might change our view of the mind.

From an evolutionary perspective, and after reviewing the practical issues involved in modeling affect programs and instantiating them into embodied systems, we would argue that the mind is essentially a host of affective capabilities that can act as valuation engines that effectively tag the stimuli in the world by deciding which is significant or not, and in what sense (i.e., with respect to any of the basic emotions), thus providing meaning of the world to the organism. As such, emotions are meaning machines, that bias action, modulate perception and attention and lay the foundation for other more cognitive processes by giving affective meaning to the world contingencies.

In other words, from an affective perspective, we can view the mind as a set of evolved, domain specific programs, each functionally specialized for solving different adaptive problems that came about throughout our evolutionary past, including the need to avoid danger, seek out resources and "solutions" to impending internal motivational problems (e.g., obtain energy through foodstuff, hydrate, regulate temperature, choose mates, and so forth), and which become active by a different set of environmental cues that are partly hardwired, and partly learned as we showed in the

results of this work. This set of programs correspond to the *affect programs*. These affect programs schedule, control and synchronize the resources of the organism so that the right subprogram can be executed and a coherent solution implemented for any of the prototypical situations that have recurred in the organism's course of its life.

In this work we have shown how a few of such affect programs, named here as *surprise*, *seeking*, *fear*, *joy*, and *distress* can sort a variety of situations that robots face while being situated in their world attempting to achieve and maintain a set of specific goals.

Each of these affect programs establishes a mode of operation for the robot, which involves the coordinated adjustment and entrainment of several different subprograms (behaviors)—activating some, deactivating others, adjusting the functioning parameters of yet others, and in some cases even creating new instances of others through learning both on the input side or the output side of the affect program—so that the whole system acts in a coherent fashion, producing coherent behavior as a response to the confrontation with specific eliciting situations.

While most theories of emotion described in Chapter 2 attempt to reduce emotion to different dimensions, and different components, our approach suggests that emotions are not reducible to anything different than what we have described here. That is, they can only be reduced to the set of coordinated evaluation-action programs for which they have evolved instructions on how to command and process *all of them together*.

A computational approach to studying affect requires the design of a framework that delves deeper into many of these computational issues, attempting to understand the sorts of representations (at various descriptive levels) of the prototypical situations or problems that the robot will address, and the different programs of specialized solutions, that will be implemented as the synchronization of appraisal mechanisms,

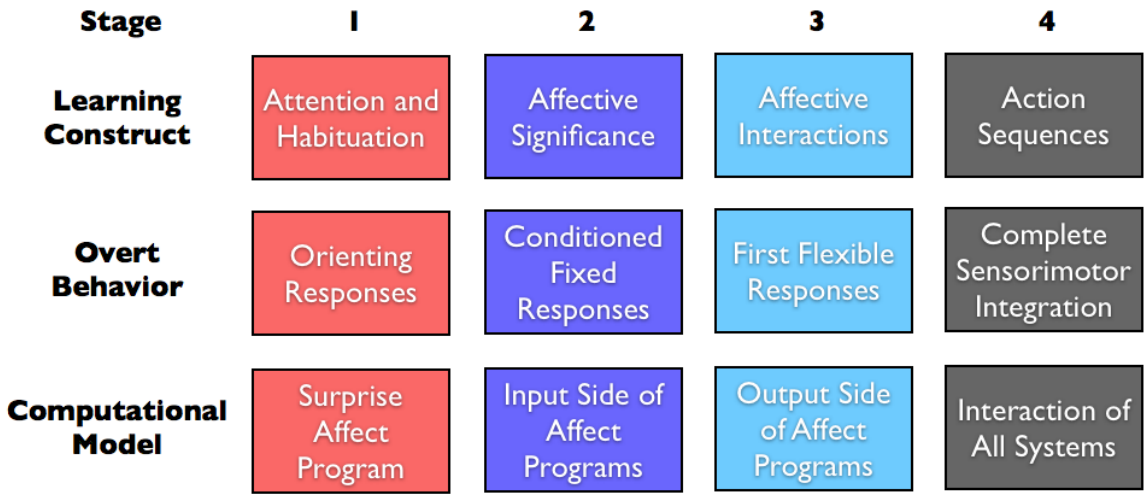

Figure 10-1: A multi-stage model of affective learning.

action arbitration strategies, attention modulation algorithms, learning models and motor control subprograms.

It is the design of such a framework, what we call a *deep model of affect*, and it is precisely this kind of model that we have built as part of this work and thus one of its main contributions. We have attempted to tie evidence stemming from multiple fields that have studied affect long since there were ever an interest in Artificial Intelligence approaches to this construct. This work has resulted in a variety of ideas, including the proposal that affective learning follows a sequence of stages such as those depicted in Figure 10-1, and which ultimately relate the many psychological constructs that are part of coherent, intelligent behavior, to the plausible computational substrates.

We hope that this framework, and the ideas contained herein, serve as a starting point and fresh perspective for the study of affect, from a computational standpoint. The space of affective programs is as vast as the minds of different species and the many environments that helped them evolve. Thus, inordinate possibilities and lessons to learn remain possible.

Bibliography

Abercrombie, E., Keefe, K., DiFrischia, D. & Zigmond, M. (1989), 'Differential effect of stress on in vivo dopamine release in striatum, nucleus accumbens, and medial frontal cortex', *Journal of Neurochemistry* **52**, 1655–1658.

Ahn, H. & Picard, R. W. (2006), Affective Cognitive Learning and Decision Making: The Role of Emotions, *in* 'The 18th European Meeting on Cybernetics and Systems Research (EMCSR 2006)', Vienna, Austria.

Arkin, R. (1990), 'Integrating behavioral, perceptual, and world knowledge in reactive navigation', *Robotics and Autonomous Systems*.

Armon-Jones, C. (1986), The thesis of constructionism, *in* R. Harré, ed., 'The Social Construction of Emotions', Basil Blackwell, New York.

Arnold, M. (1969), Human emotion and action, *in* T. Mischel, ed., 'Human action: Conceptual and Empirical Issues', Academic Press, New York.

Arnold, M. B. (1960), *Emotions and Personality*, Vol. 1 and 2, Columbia University Press, New York, NY.

Averill, J. (1980), A constructivist view of emotion, *in* R.Plutchik & H. Kellerman, eds, 'Emotion: Theory, research and experience', Academic Press, New York.

Averill, J. (1984), The acquisition of emotions during adulthood, *in* C. Malatesta & C. Izard, eds, 'Emotion in adult development', Sage, Beverly Hills.

Balleine, B. W. (2004), *Incentive Behavior*, The behavior of the laboratory rat : a handbook with tests, Oxford University Press, Oxford, pp. 436–446.

Barrett, L. F. & Russell, J. A. (1999), 'Structure of current affect', *Current Directions in Psychological Science* **8**, 10–14.

Becker, J., Breedlove, S. & Crews, D., eds (1992), *Behavioral Endocrinology*, second edn, MIT Press.

Berridge, K. & Robinson, T. (1998), 'What is the role of dopamine in reward: hedonic impact, reward learning or incentive salience?', *Brain Research Reviews* **28**, 309–369.

Berridge, K. C. (2000), Reward learning: Reinforcement, incentives, and expectations, *in* D. L. Medin, ed., 'Psychology of Learning and Motivation', Vol. Volume 40, Academic Press, pp. 223–278.

Berridge, K. C. (2004), 'Motivation concepts in behavioral neuroscience', *Physiology and Behavior* **81**(2), 179–209.

Berridge, K. C. (2006), 'The debate over dopamine's role in reward: the case for incentive salience', *Psychopharmacology*. M3: 10.1007/s00213-006-0578-x.

Berridge, K. C. & Robinson, T. E. (2003), 'Parsing reward', *Trends in neurosciences* **26**(9), 507–513.

Berridge, K. C. & Valenstein, E. S. (1991), 'What psychological process mediates feeding evoked by electrical stimulation of the lateral hypothalamus?', *Behavioral Neuroscience* **105**, 3–14.

Bindra, D. (1968), 'Neuropsychological interpretation of the effects of drive and incentive-motivation on general activity and instrumental behavior', *Psychological Review* **75**, 1–22.

Bindra, D. (1978), 'How adaptive behavior is produced: a perceptual-motivational alternative to response-reinforcement', *The Behavioral and Brain Sciences* **1**, 41–91.

Blumberg, B. (1994), Action-Selection in Hamsterdam: Lessons from Ethology, *in* 'Proceedings of Simulation of Adaptive Behavior (SAB94)'.

Blumberg, B. (1996), Old Tricks, New Dogs: Ethology and Interactive Creatures, PhD thesis, MIT.

Bolles, R. (1972), 'Reinforcement, expectancy, and learning', *Psychological Review* **79**, 394–409.

Braitenberg, V. (1984), *Vehicles: Experiments in Synthetic Psychology*, MIT Press, Cambridge.

Brauer, L. H., Goudie, A. J. & de Wit, H. (1997), 'Dopamine ligands and the stimulus effects of amphetamine: animal models versus human laboratory data', *Psychopharmacology* **130**, 2–13.

Breazeal, C. (2000), Sociable Machines: Expressive Social Exchange Between Humans and Robots, PhD thesis, Massachusetts Institute of Technology.

Breazeal, C. & Velasquez, J. (1998), Toward teaching a robot "infant" using emotive communication acts, *in* 'Socially Situated Intelligence: Papers from the 1998 Simulated Adaptive Behavior Workshop'.

Breland, K. & Breland, M. (1961), 'The Misbehavior of Organisms', *American Psychologist* **16**, 681–684.

Brooks, R. A. (1986), 'A Robust Layered Control System for a Mobile Robot', *IEEE Journal of Robotics and Automation* **RA-2**, 14–23.

Brooks, R., Breazeal, C., Irie, R., Kemp, C., Marjanović, M., Scassellati, B. & Williamson, M. (1998), Alternative Essences of Intelligence, *in* 'Proceedings of the American Association of Artificial Intelligence (AAAI-98)'.

Brown, P. L. & Jenkins, H. M. (1968), 'Auto-shaping of the pigeon's key-peck', *Journal of the Experimental Analysis of Behavior* **11**(1), 1–8.

Brunswik, E. (1956), *Perception and the representative design of psychological experiments*, The University of California Press, Berkeley.

Cañamero, L. (1997), Modeling Motivations and Emotions as a Basis for Intelligent Behavior, *in* 'Proceedings of the First International Conference on Autonomous Agents', ACM Press, New York.

Charney, D. & Redmond, D. (1983), 'Neurobiological Mechanisms in Human Anxiety', *Neuropharmacology* **22**, 1531–1536.

Chiara, G. D. (1999), 'Drug addiction as dopamine-dependent associative learning disorder', *European Journal of Pharmacology* **375**, 13–30.

Churchland, P. S. & Sejnowski, T. J. (1992), *The Computational Brain*, MIT Press, Cambridge.

Clore, G. L. (1994), Why Emotions Are Never Unconscious, *in* P. Ekman & R. J. Davidson, eds, 'The Nature of Emotion: Fundamental Questions', Oxford University Press, New York, NY, pp. 285–290.

Colby, K. M. (1974), *Artificial Paranoia*, Pergamon Press, New York, NY.

Cornelius, R. (1996), *The Science of emotion: research and tradition in the psychology of emotion*, Prentice Hall, New Jersey.

Cosmides, L. & Tooby, J. (2000), Evolutionary Psychology and the Emotions, *in* M. Lewis & J. M. Haviland-Jones, eds, 'Handbook of Emotions', The Guilford Press, New York.

Craig, W. (1918), 'Appetites and aversions as constituents of instincts', *The Biological Bulletin* **34**(2), 91–107.

Damasio, A. R. (1994), *Descartes' Error*, G.P. Putnam's Sons, New York.

Damasio, A. R. (1999), *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*, Harcourt Brace, New York.

Darwin, C. ([1859]/1998), *The Expression of the Emotions in Man and Animals*, third edn, Oxford University Press, New York. Definitive Edition.

Davidson, R. J. (1994), On Emotion, Mood, and Related Affective Constructs, *in* R. J. Davidson & P. Ekman, eds, 'The Nature of Emotion: Fundamental Questions', Oxford University Press, New York, NY, pp. 51–55.

Davis, M., Rainnie, D. & Cassell, M. (1994), 'Neurotransmission in the rat amygdala related to fear and anxiety', *Trends in Neurosciences* **17**, 208–214.

Dayan, P. & Balleine, B. W. (2002), 'Reward, Motivation, and Reinforcement Learning', *Neuron* **36**, 285–298.

Dayan, P., Kakade, S. & Montague, P. R. (2000), 'Learning and selective attention', *Nature neuroscience supplement* **3**, 1218–1223.

Dennett, D. C. (1987), *The Intentional Stance*, MIT Press.

Depue, R. & Iacono, W. (1989), 'Neurobheavioral aspects of affective disorders', *Annual Review of Psychology* **40**, 457–492.

Dickinson, A. (1994), Instrumental conditioning, *in* N. Mackintosh, ed., 'Animal Learning and Cognition', Academic Press, San Diego.

Dickinson, A. & Balleine, B. W. (1993), *Actions and responses: The dual psychology of behaviour*, Spatial representation, Blackwell, Oxford, pp. 277–293.

Dickinson, A. & Balleine, B. W. (2002), *Steven's Handbook of Experimental Psychology*, Vol. 3, 3rd edn, John Wiley & Sons, Inc., New York, NY, chapter The role of learning in motivation.

Dragoi, V. (2002), 'A feedforward model of suppressive and facilitatory habituation effects', *Biological Cybernetics* **86**, 419–426.

Dyer, M. C. (1982), *In depth understanding. A computer model of integrated processing for narrative comprehension*, MIT Press, Cambridge, MA.

Ekman, P. (1992), An Argument for Basic Emotions, *in* N. L. Stein & K. Oatley, eds, 'Basic Emotions', Lawrence Erlbaum, Hove, UK.

Ekman, P. (1994*a*), All Emotions Are Basic, *in* P. Ekman & R. J. Davidson, eds, 'The Nature of Emotion: Fundamental Questions', Oxford University Press, New York, NY, pp. 15–19.

Ekman, P. (1994*b*), Moods, Emotions, and Traits, *in* P. Ekman & R. J. Davidson, eds, 'The Nature of Emotion: Fundamental Questions', Oxford University Press, New York, NY, pp. 56–58.

Ekman, P. (1994*c*), 'Strong evidence for universals in facial expression: A reply to Rusell's mistaken critique', *Psychological Bulletin* **115**, 268–287.

Ekman, P. (1999), Basic Emotions, *in* T. Dalgleish & M. Power, eds, 'Handbook of Cognition and Emotion', John Wiley & Sons, Inc.

Ekman, P. & Friesen, W. (1986), 'Constants across cultures in the face and emotion', *Journal of Personality and Social Psychology* **17**(2), 124–129.

Ekman, P., Friesen, W. V. & Ellsworth, P. (1982), What emotion categories or dimensions can observers judge from facial behavior?, *in* P. Ekman, ed., 'Emotion in the human face', Cambridge University Press, New York, NY, pp. 39–55.

Ekman, P., Levenson, R. & Friesen, W. (1983), 'Autonomic Nervous System Activity Distinguishes Among Emotions', *Science* **221**, 1208–1210.

Elliott, C. (1992), The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System, PhD thesis, Northwestern University.

Ellis, R. D. & Newton, N. (2005), *Consciousness and Emotion: Agency, conscious choice, and selective perception*, John Benjamins Publishing.

Estes, W. (1959), The statistical approach to learning theory, *in* S. Koch, ed., 'Psychology: A study of a science', McGraw-Hill, New York.

Fanselow, M. (1994), 'Neural organization of the defensive behavior system responsible for fear', *Psychonomic Bulletin Reviews* **1**, 429–438.

Farthing, G. (1992), *The Psychology of Consciousness*, Prentice Hall, Englewood Cliffs.

Floreano, D. & Mondana, F. (1998), 'Evolutionary neurocontrollers for autonomous mobile robots', *Neural Networks* **11**, 1461–1478.

Flynn, J. (1967), The neural basis for aggression in cats, *in* C. Glass, ed., 'Neurophysiology and Emotion', Rockefeller University Press, New York.

Forum, M. (2002), 'Official MPI (Message Passing Interface) standard document, errata, and archives of the MPI Forum. http://www.mpi-forum.org'.

Fox, N. & Davidson, R. (1986), 'Taste-elicited changes in facial signs of emotion and the asymmetry of brain electrical activity in human newborns', *Neuropsychologia* **24**, 417–422.

Frank, R. H. (1988), *Passions within Reason: The Strategic Role of the Emotions*, Norton, New York, NY.

Frijda, N. H. (1986), *The Emotions*, Cambridge University Press, Cambridge, England.

Frijda, N. H. (2000), The Psychologists' Point of View, *in* M. Lewis & J. M. Haviland-Jones, eds, 'Handbook of Emotions', The Guilford Press, New York.

Gallagher, M. & Chiba, A. (1996), 'The Amygdala and Emotion', *Current Opinion in Neurobiology* **6**(2), 221–227.

Gallistel, C. R. (1978), 'Irrelevance of past pleasure', *Behavioral and Brain Sciences* **1**, 59–60.

Gazzaniga, M. S. (1998), *The Mind's Past*, University of California Press, Berkeley.

Gazzaniga, M. S. & LeDoux, J. E. (1978), *The Integrated Mind*, Plenum Press, New York.

Gergen, K. J. (1985), 'The Social Constructionist Movement in Modern Psychology', *American Psychologist* **40**, 266–275.

Gibson, J. (1979), *The ecological approach to visual perception*, Houghton Mifflin, Boston.

Gill, T. M., Sarter, M. & Givens, B. (2000), 'Sustained Visual Attention Performance-Associated Prefrontal Neuronal Activity: Evidence for Cholinergic Modulation', *Journal of Neuroscience* **20**(12), 4745–4757.

Gray, J. A. (1982), *The Neuropsychology of Anxiety*, Oxford University Press, Oxford.

Gray, J. A. (1990), 'Brain systems that mediate both emotion and cognition', *Cognition and Emotion* **4**, 269–288.

Graybiel, A. (1995), 'Building action repertoires: memory and learning functions of the basal ganglia', *Current Opinion in Neurobiology* **5**, 733–741.

Graybiel, A. (1998), 'The basal ganglia and chunking of action repertoires', *Neurobiology of Learning and Memory* **70**, 119–136.

Griffiths, P. E. (1997), *What emotions really are: the problem of psychological categories*, The University of Chicago Press.

Hebb, D. O. (1949), *Organization of behavior: a neuropsychological theory*, John Wiley, New York.

Heider, F. ([1930]/1959), 'The function of the perceptual system', *Psychological Issues* pp. 371–394.

Hershberger, W. A. (1986), 'An approach through the looking glass', *Animal Learning and Behavior* **14**, 443–451.

Hikosaka, O. (1998), 'Neural systems for control of voluntary action–a hypothesis', *Advances in Biophysics* **35**, 81–102.

Holland, P. C. & Gallagher, M. (1999), 'Amygdala circuitry in attentional and representational processes', *Trends in Cognitive Sciences* **3**(2), 65–73.

Horvitz, J. (2000), 'Mesolimbical and nigrostriatal dopamine responses to salient non-reward events', *Neuroscience* **96**(4), 651–656.

Houk, J., Adams, J. & Barto, A. (1995), A model of how the basal ganglia generate and use neural signals that predict reinforcement, *in* J. D. J.C. Houk & D. Beiser, eds, 'Models of Information Processing in the Basal Ganglia', MIT Press, Cambridge.

Houwer, J. D., Thomas, S. & Baeyens, F. (2001), 'Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning', *Psychological bulletin* **127**(6), 853–869.

Hull, C. L. (1943), *Principles of behavior, an introduction to behavior theory*, Appleton Century, New York, NY.

Ikemoto, S. & Panksepp, J. (1999), 'The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking', *Brain Research Reviews* **31**, 6–41.

Introini-Collison, I. B., Dalmaz, C. & Mcgaugh, J. L. (1996), 'Amygdala β-Noradrenergic Influences on Memory Storage Involve Cholinergic Activation', *Neurobiology of Learning and Memory* **65**, 57–64.

Iversen, S., Kupfermann, I. & Kandel, E. R. (2000), Emotinal States and Feelings, *in* E. R. Kandel, J. H. Schwartz & T. M. Jessell, eds, 'Principles of Neural Science', 4th edn, McGraw-Hill.

Izard, C. (1971), *The Face of Emotion*, Appleton, New York.

Izard, C. (1977), *Human Emotions*, Plenum, New York.

Izard, C. (1993), 'Four Systems for Emotion Activation: Cognitive and Noncognitive Processes', *Psychological Review* **100**(1), 68–90.

Izard, C. (1994), 'Innate and Universal Facial Expressions: Evidence from developmental and cross cultural research', *Psychological Bulletin* **115**, 288–299.

James, W. (1884), 'What is an emotion?', *Mind* **9**, 188–205.

Johnson-Laird, P. & Oatley, K. (1992), Basic Emotions, Rationality, and Folk Theory, *in* N. L. Stein & K. Oatley, eds, 'Basic Emotions', Lawrence Erlbaum, Hove, UK.

Johnson-Laird, P. N. (1988), *Computer and the Mind: An Introduction to Cognitive Science*, Harvard University Press, Cambridge, MA.

Jürgens, U. (1976), 'Reinforcing concomitants of electrically elicited vocalizations', *Experimental Brain Research* **26**, 203–214.

Kaada, B. R. (1951), 'Somato–motor, autonomic, and electroencephalographic responses to electrical stimulation of 'rhinencephalic' and other structures in primates, cat, and dog. A study of responses from the limbic, subcallosal, orbito–insular, piriform and temporal cortex, hippocampus-fornix and amygdala', *Acta Physiologica Scandinavica* **23**((Suppl. 83)), 1–285.

Kihlstrom, J. (1987), 'The cognitive unconscious', *Science* **237**(4821), 1445–1452.

Kim, J. & Fanselow, M. (1992), 'Modality-specific retrograde amnesia of fear', *Science* **256**, 675–677.

Kleinginna, P. & Kleinginna, A. (1981), 'A categorized list of emotion definitions, with suggestions for a consensual definition', *Motivation and Emotion* **5**, 345–379.

Klüver, H. & Bucy, P. (1939), 'Preliminary analysis of functions of the temporal lobes in monkeys', *Archives of Neurology and Psychiatry* **42**, 979–1000.

Knowlton, B., Mangels, J. & Squire, L. (1996), 'A neostriatal habit learning system in humans', *Science* **273**, 1399–1402.

Koenig, N. & Howard, A. (2004), Design and Use Paradigms for Gazebo, An Open-Source Multi-Robot Simulator, *in* 'Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)', IEEE, Sendai, Japan, pp. 2149–2154.

Koffka, K. (1935), *Principles of Gestalt Psychology*, Harcourt Brace, New York.

Konidaris, G. & Barto, A. G. (2006), An Adaptive Robot Motivational System, *in* 'SAB', pp. 346–356.

Konorski, J. (1967), *Integrative Activity of the Brain: An Interdisciplinary Approach*, The University of Chicago, Chicago.

Kupfermann, I., Kandel, E. R. & Iversen, S. (2000), Motivational and Addictive States, *in* E. R. Kandel, J. H. Schwartz & T. M. Jessell, eds, 'Principles of Neural Science', 4th edn, McGraw-Hill.

Lazarus, R. (1991), *Emotion and Adaptation*, Oxford University Press, New York.

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

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

LeDoux, J. E. (1994), Emotional Processing, But Not Emotions, Can Occur Unconsciously, *in* P. Ekman & R. J. Davidson, eds, 'The Nature of Emotion: Fundamental Questions', Oxford University Press, New York, NY, pp. 291–292.

Ljungberg, T., Apicella, P. & Schultz, W. (1992), 'Responses of monkey dopamine neurons during learning of behavioral reactions', *Journal of Neurophysiology* **67**, 145–163.

Lorenz, K. (1973), *Foundations of Ethology*, Springer-Verlag, New York, NY.

Lutz, C. (1988), *Unnatural emotions: Everyday sentiments on a Micronesian atoll and their challenge to western theory*, University of Chicago Press, Chicago, IL.

MacLean, P. (1990), *The triune brain in evolution: Role in paleocerebral functions*, Plenum Press, New York.

Maes, P. (1989), 'How to Do the Right Thing', *Connection Science Journal*.

Maes, P. (1991), Situated Agents Can Have Goals, *in* P. Maes, ed., 'Designing Autonomous Agents', MIT Press, Cambridge.

Maes, P. (1995), Modeling Adaptive Autonomous Agents, *in* C. Langton, ed., 'Artificial Life', MIT Press, Cambridge.

Mandler, G. (1984), *Mind and body: The psychology of emotion and stress*, Norton, New York.

McClure, S. M., Daw, N. D. & Montague, P. R. (2003), 'A computational substrate for incentive salience', *Trends in neurosciences* **26**(8), 423–428.

McDonald, R. & White, N. (1993), 'A Triple dissociation of memory systems: hippocampus, amygdala, and dorsal striatum', *Behavioral Neuroscience* **107**(1), 3–22.

Meltzer, H. & Lowy, M. (1987), The serotonin hypothesis of depression, *in* H. Meltzer, ed., 'Psychopharmacology: the third generation of progress', Raven Press, New York, pp. 513–26.

Minguez, J. & Montano, L. (2004), 'Nearness Diagram (ND) Navigation: Collision Avoidance in Troublesome Scenarios', *IEEE Transactions on Robotics and Automation* **20**(1), 45–59.

Minsky, M. (1986), *The Society of Mind*, Simon and Schuster, New York.

Minsky, M. (2006), *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*, Simon and Schuster.

Mirenowicz, J. & Schultz, W. (1994), 'Importance of unpredictedness for reward responses in primate dopamine neurons', *Journal of Neurophysiology* **72**, 1024–1027.

Mirenowicz, J. & Schultz, W. (1996), 'Preferential activation of midbrain dopamine neurons by appetitive rather than aversive stimuli', *Nature* **379**, 449–451.

Montague, P. R., Hyman, S. E. & Cohen, J. D. (2004), 'Computational roles for dopamine in behavioural control', *Nature* **431**, 760–767.

Moss, R. & Dudley, C. (1984), The challenge studying the behavioral effects of neuropeptides, *in* L. Iversen, S. Iversen & S. Snyder, eds, 'Handbook of Psychopharmacology', Vol. 18, Plenum Press, New York.

Mowrer, O. H. (1960), *Learning theory and behavior*, Wiley, New York, NY.

Nicola, S., Surmeier, D. & Malenka, R. (2000), 'Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens', *Annual Review of Neuroscience* **23**, 185–215.

Nisbett, R. & Wilson, T. (1977), 'Telling more than we can know: Verbal reports on mental processes', *Psychological Review* **84**, 231–259.

Niv, Y., Daw, N. D., Joel, D. & Dayan, P. (2007), 'Tonic dopamine: opportunity costs and the control of response vigor', *Psychopharmacology* **191**, 507–520.

Oades, R. (1985), 'The role of noradrenaline in tuning and dopamine in switching between signals in the CNS', *Neuroscience and Biobehavioral Reviews* **9**, 261–282.

Oatley, K. & Johnson-Laird, P. N. (1987), 'Towards a cognitive theory of emotions', *Cognition and Emotion* **1**, 29–50.

Oatley, K. & Johnson-Laird, P. N. (1996), The communicative theory of emotions: empirical tests, mental models, and implications for social interaction, *in* L. L. Martin & A. Tesser, eds, 'Striving and feeling: Interactions among goals, affect, and self-regulation', Lawrence Erlbaum, Mahwah, NJ, pp. 363–393.

Öhman, A. (1986), 'Facethe beast and fear the face. Animal and social fears as prototypes for evolutionary analyses of emotion', *Psychophysiology* **23**, 123–145.

Ortony, A. & Turner, T. J. (1990), 'What's Basic About Basic Emotions?', *Psychological Review* **97**(3), 315–331.

Ortony, A., Clore, G. & Collins, A. (1988), *The Cognitive Structure of Emotions*, Cambridge University Press, Cambridge, England.

Packard, M. & McGaugh, J. (1996), 'Inactivation of Hippocampus or Caudate Nucleus with Lidocaine Differentially Affects Expression of Place and Response Learning', *Neurobiology of Learning and Memory* **65**, 65–72.

Packard, M. & Teather, L. (1998), 'Amygdala Modulation of Multiple Memory Systems: Hippocampus and Caudate-Putamen', *Neurobiology of Learning and Memory* **69**(2), 163–203.

Packard, M., Cahill, L. & McGaugh, J. (1994), 'Amygdala modulation of hippocampal-dependent and caudate-dependent memory processes', *Proceedings of the National Academy of Sciences* **91**, 8477–8481.

Panksepp, J. (1982), 'Toward a general psychobiological theory of emotions', *Behavioral and Brain Sciences* **5**, 407–467.

Panksepp, J. (1986), 'The neurochemistry of behavior', *Annual Review of Psychology* **37**, 77–107.

Panksepp, J. (1993), Neurochemical control of moods and emotions:Amino acids to neuropeptides, *in* M. Lewis & J. M. Haviland-Jones, eds, 'Handbook of Emotions', The Guilford Press, New York.

Panksepp, J. (1994), Basic Emotions Ramify Widely in the Brain, Yielding Many Concepts That Cannot Be Distinguished Unambiguously... Yet, *in* P. Ekman & R. Davidson, eds, 'The Nature of Emotion: Fundamental Questions', Oxford University Press, New York, NY, pp. 86–88.

Panksepp, J. (1995), 'The Emotional Brain and Biological Psychiatry', *Advances in Biological Psychiatry* **1**, 263–286.

Panksepp, J. (1998), *Affective neuroscience : the foundations of human and animal emotions*, Oxford University Press, New York.

Panksepp, J. (2000), Emotions as Natural Kinds within the Mammalian Brain, *in* M. Lewis & J. M. Haviland-Jones, eds, 'Handbook of Emotions', The Guilford Press, New York.

Panksepp, J. (2005), 'Affective consciousness: Core emotional feelings in animals and humans', *Consciousness and Cognition* **14**(1), 30–80.

Peciña, S., Berridge, K. C. & Parker, L. A. (1997), 'Pimozide does not shift palatability: separation of anhedonia from sensorimotor suppression by taste reactivity', *Pharmacology Biochemistry and Behavior* **58**, 801–811.

Peciña, S., Cagniard, B., Berridge, K. C., Aldridge, J. W. & Zhuang, X. (2003), 'Hyperdopaminergic Mutant Mice Have Higher "Wanting" But Not "Liking" for Sweet Rewards', *Journal of Neuroscience* **23**(28), 9395–9402.

Petersen, C., Caldwell, J., Jirikowski, G. & Insel, T., eds (1992), *Oxytocin in maternal, sexual, and social behaviors*, Vol. 652, Annals of the New York Academy of Sciences.

Picard, R. W. (1997), *Affective Computing*, MIT Press, Cambridge.

Plutchik, R. (1994), *The Psychology and Biology of Emotion*, HarperCollins, New York.

Plutchik, R. (2001), 'The Nature of Emotions', *American Scientist* **89**(4), 344–350.

Posner, M. & Badgaiyan, R. (1998), Attention and Neural Networks, *in* D. L. R. Parks & D. Long, eds, 'Fundamentals of Neural Network Modeling', MIT Press, Cambridge.

Redgrave, P., Prescott, T. & Gurney, K. (1999*a*), 'The basal ganglia: a vertebrate solution to the selection problem?', *Neuroscience* **89**(4), 1009–1023.

Redgrave, P., Prescott, T. & Gurney, K. (1999*b*), 'Is the short-latency dopamine response too short to signal reward error?', *Trends in Neurosciences* **22**, 146–151.

Reilly, S. N. (1996), Believable Social and Emotional Agents, PhD thesis, School of Computer Science, Carnegie Mellon University.

Reisenzein, R. (1994), 'Pleasure-arousal theory and the intensity of emotions', *Journal of Personality and Social Psychology* **67**, 525–539.

Reisenzein, R. & Hofmann, T. (1990), 'An investigation of dimensions of cognitive appraisal in emotion using a repertory grid technique', *Motivation and Emotion* **14**, 19–38.

Rescorla, R. & Wagner, A. (1972), A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement, *in* A. Black & W. Prokasy, eds, 'Classical Conditioning II: Current Research and Theory', Appleton Century Crofts, New York.

Robbins, T. & Everitt, B. (1992), 'Functions of dopamine in the dorsal and ventral striatum', *Seminars in Neuroscience* **4**, 119–128.

Robbins, T. W. & Everitt, B. J. (2006), 'A role for mesencephalic dopamine in activation: commentary on Berridge (2006)', *Psychopharmacology*. M3: 10.1007/s00213-006-0528-7.

Robbins, T. W. & Koob, G. F. (1980), 'Selective disruption of displacement behaviour by lesions of the mesolimbic dopamine system', *Nature* **285**, 409–412.

Robertson, H. A., Martin, I. L. & Candy, J. M. (1978), 'Differences in benzodiazepine receptor binding in Maudsley reactive and Maudsley non-reactive rats', *Journal of Pharmacology* **50**, 455–457.

Romo, R. & Schultz, W. (1990), 'Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements', *Journal of Neurophysiology* **63**, 592–606.

Roseman, I. J. (1984), Cognitive determinants of emotions: A structural theory, *in* P. Shaver, ed., 'Review of Personality and Social Psychology', Vol. 5, Sage, Beverly Hills, CA, pp. 11–36.

Roseman, I. J., Spindel, M. S. & Jose, P. E. (1990), 'Appraisal of emotion-eliciting events: Testing a theory of discrete emotions', *Journal of Personality and Social Psychology* **59**, 899–915.

Russell, J. A. (1980), 'A circumplex model of affect', *Journal of personality and social psychology* **39**(6), 1161–1178. NR: 71 reference(s) present, 71 reference(s) displayed RX: 446 (on Apr 25, 2007).

Russell, J. A. (2003), 'Core Affect and the Psychological Construction of Emotion', *Psychological Review* **110**(1), 145–172.

Salamone, J. D. & Correa, M. (2002), 'Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine', *Behavioural Brain Research* **137**(1-2), 3–25.

Salamone, J. D., Correa, M., Mingote, S. M. & Weber, S. M. (2005), 'Beyond the reward hypothesis: alternative functions of nucleus accumbens dopamine', *Current Opinion in Pharmacology* **5**(1), 34–41.

Scassellati, B. (1998), Building Behaviors Developmentally: A New Formalism, *in* 'Integrating Robotics Research: Papers from the 1998 AAAI Spring Symposium', AAAI Press.

Schacter, D., Chiu, C. & Ochsner, K. (1993), 'Implicit memory: A selective review', *Annual Review of Neuroscience* **16**, 159–182.

Scherer, K. (1993), 'Studying the emotion-antecedent appraisal process: An expert system approach', *Cognition and Emotion* **7**, 325–355.

Scherer, K. R. (1984), On the nature and function of emotion: A component process approach, *in* K. R. Scherer & P. Ekman, eds, 'Approaches to emotion', Lawrence Erlbaum, Hilllsdale, NJ, pp. 293–317.

Schneirla, T. (1959), An evolutionary and developmental theory of biphasic process underlying approach and withdrawal, *in* M. Jones, ed., 'Nebraska Symposium on Motivation', University of Nebraska, Lincoln.

Schultz, W. (1997), 'Dopamine neurons and their role in reward mechanisms', *Current Opinion in Neurobiology* **7**, 191–197.

Schultz, W. (1998), 'Predictive reward signal of dopamine neurons', *Journal of Neurophysiology* **80**, 1–27.

Schultz, W. (2002), 'Getting Formal with Dopamine and Reward', *Neuron* **36**(2), 241–263.

Schultz, W. (2006), 'Behavioral theories and the neurophysiology of reward', *Annual Review of Psychology* **57**, 87–115.

Schultz, W. & Romo, R. (1990), 'Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions', *Journal of Neurophysiology* **63**, 607–624.

Schultz, W., Dayan, P. & Montague, P. (1997), 'A neural substrate of prediction and reward', *Science* **275**, 1593–1599.

Schultz, W., Romo, R., Ljungberg, T., Mirenowicz, J., Hollerman, J. & Dickinson, A. (1995), Reward-related signals carried by dopamine neurons, *in* D. B. J.R. Houk, J.L. Davis, ed., 'Models of Information Processing in the Basal Ganglia', MIT Press, Cambridge.

Sherrington, C. (1906), *The Integrative Action of the Nervous System*, Charles Scribner's Sons, New York.

Smith, G. P. (1995), Dopamine and food reward, *in* A. M. Morrison & S. J. Fluharty, eds, 'Progress in psychobiology and physiological psychology', Vol. 15, Academic Press, New York, NY, pp. 83–144.

Squire, L. (1992), 'Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans', *Psychological Review* **99**, 195–231.

Squire, L. & Zola, S. (1996), 'Structure and function of declarative and non-declarative memory systems', *Proceedings of the National Academy of Sciences* **93**, 13515–13522.

Staddon, J. (1993), 'On Rate-Sensitive Habituation', *Adaptive Behavior*.

Strecker, R. & Jacobs, B. (1985), 'Substantia nigra dopaminergic unit activity in behaving cats: effect of arousal on spontaneous discharge and sensory evoked activity', *Brain Research* **361**, 339–350.

Sutton, R. (1988), 'Learning to predict by the method of temporal difference', *Machine Learning* **3**, 9–44.

Sutton, R. & Barto, A. (1981), 'Toward a modern theory of adaptive networks: expectation and prediction', *Psychological Review* **88**, 135–170.

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

Thagard, P. (2006), *Hot thought: mechanisms and applications of emotional cognition*, MIT Press, Cambridge, MA.

Thorndike, E. (1911), *Animal Intelligence: Experimental Studies*, Macmillan Press, New York.

Tinbergen, N. (1951), *The Study of Instinct*, Oxford University Press, New York.

Toates, F. (1986), *Motivational Systems*, Cambridge University Press, Cambridge, England.

Tomkins, S. S. (1962), *Affect, Imagery, consciousness: Vol.1 The Positive Affects*, Springer, New York.

Tomkins, S. S. (1963), *Affect, Imagery, consciousness: Vol.2 The Negative Affects*, Springer, New York.

Tomkins, S. S. (1984), Affect theory, *in* K. R. Scherer & P. Ekman, eds, 'Approaches to emotion', Erlbaum, Hillsdale, NJ, pp. 163–195.

Trulson, M. & Preussler, D. (1984), 'Dopamine-containing ventral tegmental area neurons in freely moving cats: activity during the sleep-waking cycle and effects of stress', *Experimental Neurology* **83**, 367–377.

Turner, T. J. & Ortony, A. (1992), 'Basic Emotions: Can Conflicting Criteria Converge?', *Psychological Review* **99**(3), 566–571.

Tyrell, T. (1994), 'An Evaluation of Maes' Bottom-Up Mechanism for Behavior Selection', *Adaptive Behavior* **2**(4), 307–348.

Velásquez, J. D. (1996), Cathexis: A Computational model for the Generation of Emotions and Their Influence in the Behavior of Autonomous Agents, Master's thesis, MIT.

Velásquez, J. D. (1997), Modeling Emotions and Other Motivations in Synthetic Agents, *in* 'Proceedings of the 1997 National Conference on Artificial Intelligence, AAAI-97', pp. 10–15.

Velásquez, J. D. (1998*a*), A Computational Framework for Emotion-Based Control, *in* 'Proceedings of SAB'98 workshop on Grounding Emotions in Adaptive Systems'.

Velásquez, J. D. (1998*b*), When Robots Weep: Emotional Memories and Decision-Making, *in* 'Proceedings of the 1998 National Conference on Artificial Ingelligence, AAAI-98', pp. 70–75.

Velásquez, J. D. (1999), Building Affective Robots, *in* 'Proceedings of the Second International Symposium on Humanoid Robots', Waseda University, Tokyo.

Walter, W. G. (1961), *The Living Brain*, Penguin Books, Victoria, Australia.

Watson, J. B. (1930), *Behaviorism*, University of Chicago Press, Chicago, IL.

Weiner, B. & Graham, S. (1984), An attributional approach to emotional development, *in* C. E. Izard, J. Kagan & R. B. Zajonc, eds, 'Emotions, cognition, and behavior', Cambridge University Press, New York, NY, pp. 167–191.

Weiskrantz, L. (1956), 'Behavioral changes associated with ablation of the amygdaloid complex in monkeys', *Journal of Comparative and Physiological Psychology* **49**, 381–391.

White, N. M. (1996), 'Addictive drugs as reinforcers: multiple partial actions on memory system', *Addiction* **91**, 921–949.

White, N. M. (1997), 'Mnemonic functions of the basal ganglia', *Current Opinion in Neurobiology* **7**, 164–169.

White, N. M. & McDonald, R. J. (2001), 'Multiple Parallel Memory Systems in the Brain of the Rat', *Neurobiology of Learning and Memory* **77**, 125–184.

Williams, D. R. & Williams, H. (1969), 'Auto-maintenance in the pigeon: Sustained pecking despite contingent non-reinforcement', *Journal of the Experimental Analysis of Behavior* **12**, 511–520.

Winkielman, P. & Berridge, K. C. (2004), 'Unconscious Emotion', *Current Directions in Psychological Science* **13**(3), 120–123.

Wise, R. (1996), 'Neurobiology of Addiction', *Current Opinion in Neurobiology* **6**, 243–251.

Wise, R. & Rompre, P. (1989), 'Brain dopamine and reward', *Annual Review of Psychology* **40**, 191–225.

Wise, R. A. (2002), 'Brain Reward Circuitry: Insights from Unsensed Incentives', *Neuron* **36**(2), 229–240.

Wyvell, C. L. & Berridge, K. C. (2000), 'Intra-Accumbens Amphetamine Increases the Conditioned Incentive Salience of Sucrose Reward: Enhancement of Reward "Wanting" without Enhanced "Liking" or Response Reinforcement', *Journal of Neuroscience* **20**(21), 8122–8130.

Young, A., Ahier, R., Upton, R., Joseph, M. & Gray, J. (1998), 'Increased extracellular dopamine in the nucleus accumbens of the rat during associative learning of neutral stimuli', *Neuroscience* **83**, 1175–1183.

Zajonc, R. B. (1985), 'Emotion and facial efference: A theory reclaimed', *Science* **228**, 15–21.

Zajonc, R. B. (1994), Evidence for Nonconscious Emotions, *in* P. Ekman & R. J. Davidson, eds, 'The Nature of Emotion: Fundamental Questions', Oxford University Press, New York, NY, pp. 293–297.