Who Needs Emotions? The Brain Meets the Robot

JEAN-MARC FELLOUS
MICHAEL A. ARBIB,
Editors

OXFORD UNIVERSITY PRESS

Who Needs Emotions?

SERIES IN AFFECTIVE SCIENCE

Series Editors

Richard J. Davidson

Paul Ekman

Klaus Scherer

The Nature of Emotion: Fundamental Questions
Edited by Paul Ekman and Richard J. Davidson

Boo! Culture, Experience, and the Startle Reflex
by Ronald Simons

Emotions in Psychopathology: Theory and Research
Edited by William F. Flack, Jr., and James D. Laird

What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)
Edited by Paul Ekman and Erika Rosenberg

Shame: Interpersonal Behavior, Psychopathology, and Culture
Edited by Paul Gilbert and Bernice Andrews

Affective Neuroscience: The Foundations of Human and Animal Emotions
by Jaak Panksepp

Extreme Fear, Shyness, and Social Phobia: Origins, Biological Mechanisms, and Clinical Outcomes
Edited by Louis A. Schmidt and Jay Schulkin

Cognitive Neuroscience of Emotion
Edited by Richard D. Lane and Lynn Nadel

The Neuropsychology of Emotion
Edited by Joan C. Borod

Anxiety, Depression, and Emotion
Edited by Richard J. Davidson

Persons, Situations, and Emotions: An Ecological Approach
Edited by Hermann Brandstätter and Andrzej Eliasz

Emotion, Social Relationships, and Health
Edited by Carol D. Ryff and Burton Singer

Appraisal Processes in Emotion: Theory, Methods, Research
Edited by Klaus R. Scherer, Angela Schorr, and Tom Johnstone

Music and Emotion: Theory and Research
Edited by Patrik N. Juslin and John A. Sloboda

Nonverbal Behavior in Clinical Settings
Edited by Pierre Philippot, Robert S. Feldman, and Erik J. Coats

Memory and Emotion
Edited by Daniel Reisberg and Paula Hertel

Psychology of Gratitude
Edited by Robert A. Emmons and Michael E. McCullough

Thinking about Feeling: Contemporary Philosophers on Emotions
Edited by Robert C. Solomon

Bodily Sensibility: Intelligent Action
by Jay Schulkin

Who Needs Emotions? The Brain Meets the Robot
Edited by Jean-Marc Fellous and Michael A. Arbib

Who Needs Emotions?

The Brain Meets the Robot

Edited by

JEAN-MARC FELLOUS &
MICHAEL A. ARBIB

OXFORD
UNIVERSITY PRESS

2005

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that further
Oxford University's objective of excellence
in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com

Library of Congress Cataloging-in-Publication Data
Who needs emotions? : the brain meets the robot / edited by Jean-Marc Fellous, Michael A. Arbib
p. cm.—(Series in affective science)
ISBN-13 978-0-19-516619-4
ISBN 0-19-516619-1
1. Emotions. 2. Cognitive neuroscience. 3. Artificial intelligence. 4. Robots.
I. Fellous, Jean-Marc. II. Arbib, Michael A. III. Series.
QP401.W48 2005
152.4—dc22 2004046936

9 8 7 6 5 4 3 2 1

Printed in the United States of America
on acid-free paper

Preface

For some, emotions are uniquely human attributes; for others, emotions can be seen everywhere from animals to machines and even the weather. Yet, ever since Darwin published *The Expression of the Emotions in Man and Animals*, it has been agreed that, no matter what may be their uniquely human aspects, emotions in some sense can be attributed to a wide range of animals and studied within the unifying framework of evolutionary theory. In particular, by relating particular facial expressions in an animal species to patterns of social behavior, we can come to more deeply appreciate how and why our own, human, social interactions can express our emotions; but what is "behind" these facial expressions? Part II of this book, "Brains," will probe the inner workings of the brain that accompany the range of human and animal emotions and present a range of unique insights gained by placing these brain mechanisms in an evolutionary perspective.

The last 50 years have seen not only a tremendous increase in the sophistication of neuroscience but also the truly revolutionary development of computer technology. The question "Can machines think?" long predates the computer age but gained new technical perspective with the development of that branch of computer science known as artificial intelligence (AI). It was long thought that the skillful playing of chess was a sure sign of intelligence, but now that Deep Blue has beaten Kasparov, opinion is divided as to whether the program is truly "intelligent" or just a "bag of tricks" exploiting a large database and fast computing. Either way, it is agreed that intelligence, whether human or otherwise, is not a unitary capability but rather a set of interacting capabilities. Some workers in AI are content to create the appearance of intelligence—behavior seen "from the outside"—while others

want their computer programs to parallel, at some level of abstraction, the structure of the human brain sufficiently to claim that they provide a "packet of intelligence" akin to that provided by particular neural circuits within the rich complexity of the human brain.

Part III of the book, "Robots," brings AI together with the study of emotion. The key division is between creating robots or computers that really have emotions and creating those that exhibit the appearance of emotion through, for example, having a "face" that can mimic human emotional expressions or a "voice" that can be given human-like intonations. To see the distinction, consider receiving a delightful present and smiling spontaneously with pleasure as against receiving an unsatisfactory present and forcing a smile so as not to disappoint the giver. For many technological applications—from computer tutors to video games—the creation of apparent emotions is all that is needed and certainly poses daunting challenges. Others seek to develop "cognitive architectures" that in some appropriately generalized sense may both explain human emotions and anchor the design of artificial creatures which, like humans, integrate the emotional and the rational in their behavior.

The aim of this book, then, is to represent the state of the art in both the evolutionary analysis of neural mechanisms of emotion (as well as motivation and affect) in animals as a basis for a deeper understanding of such mechanisms in the human brain as well as the progress of AI in creating the appearance or the reality of emotion in robots and other machines. With this, we turn to a brief tour of the book's contents.

Part I: Perspective. To highlight the differences of opinion that characterize the present dialog concerning the nature of emotion, we first offer a fictional dialog in which "Russell" argues for the importance of clear definitions to advance the subject, while "Edison" takes the pragmatic view of the inventor who just wants to build robots whose emotionality can be recognized when we see it. Both are agreed (a great relief to the editors) on the fruitfulness of sharing ideas between brain researchers and roboticists, whether our goal is to understand what emotions are or what they may become. Ralph Adolphs provides a perspective from social cognitive neuroscience to stress that we should attribute emotions and feelings to a system only if it satisfies various criteria in addition to mere behavioral duplication. Some aspects of emotion depend only on how humans react to observing behavior, some depend additionally on a scientific account of adaptive behavior, and some depend also on how that behavior is internally generated—the social communicative, the adaptive/regulatory, and the experiential aspects of emotion, respectively. He argues that correctly attributing emotions and feelings to robots would require not only that robots be situated in the world but also that they be constituted internally in respects that are relevantly similar to humans.

Part II: Brains. Ann E. Kelley provides an evolutionary perspective on the neurochemical networks encoding emotion and motivation. Cross-talk between cortical and subcortical networks enables intimate communication between phylogenetically newer brain regions, subserving subjective awareness and cognition (primarily cortex), and ancestral motivational systems that exist to promote survival behaviors (primarily hypothalamus). Neurochemical coding, imparting an extraordinary amount of specificity and flexibility within these networks, appears to be conserved in evolution. This is exemplified by examining the role of dopamine in reward and plasticity, serotonin in aggression and depression, and opioid peptides in pain and pleasure. However, Kelley reminds us that although these neurochemical systems generally serve a highly functional and adaptive role in behavior, they can be altered in maladaptive ways as in the case of addiction and substance abuse. Moreover, the insights gained raise the question of the extent to which human emotions can be abstracted from their specific neurochemical substrate, and the implications our answers may have for the study of robots.

Jean-Marc Fellous and Joseph E. LeDoux advance the view that, whereas humans usually think of emotions as feelings, they can be studied quite apart from feelings by looking at "emotional behavior." Thus, we may infer that a rat is "afraid" in a particular situation if it either freezes or runs away. Studies of fear conditioning in the rat have pinpointed the amygdala as an important component of the system involved in the acquisition, storage, and expression of fear memory and have elucidated in detail how stimuli enter, travel through, and exit the amygdala. Understanding these circuits provides a basis for discussing other emotions and the "overlay" of feelings that has emerged in human evolution. Edmund T. Rolls offers a related biological perspective, suggesting how a whole range of emotions could arise on the basis of the evolution of a variety of biological strategies to increase survival through adaptation based on positive and negative reinforcement. His hypothesis is that brains are designed around reward and punishment evaluation systems because this is the way that genes can build a complex system that will produce appropriate but flexible behavior to increase their fitness. By specifying goals rather than particular behavioral patterns of response, genes leave much more open the possible behavioral strategies that might be required to increase their fitness. Feelings and consciousness are then, as for Fellous and LeDoux, seen as an overlay that can be linked to the interaction of basic emotional systems with those that, in humans, support language. The underlying brain systems that control behavior in relation to previous associations of stimuli with reinforcement include the amygdala and, particularly well-developed in primates, the orbitofrontal cortex. The overlay in humans involves computation with many "if . . . then" statements, to implement a plan to obtain a reward. In this case, something akin to syntax

is required because the many symbols that are part of the plan must be correctly linked or bound.

Between them, these three chapters provide a strong evolutionary view of the role of the emotions in the brain's mediation of individual behavior but say little about the social dimension of emotion. Marc Jeannerod addresses this by emphasizing the way in which our social behavior depends on reading the expressions of others. This takes us back to Darwin's original concern with the facial expression of emotions but carries us forward by looking at ways in which empathy and emotional understanding may be grounded in brain activity shared between having an emotion and observing that emotion in others. Indeed, the activity of "mirror neurons" in the monkey brain, which are active both when the monkey executes a certain action and when it observes another executing a similar action, is seen by a number of researchers as providing the evolutionary grounding for both empathy and language. However, the utility of such shared representations demands other mechanisms to correctly attribute the action, emotion, or utterance to the appropriate agent; and the chapter closes with an analysis of schizophrenia as a breakdown in attribution of agency for a variety of classes of action and, in some cases, emotion.

Part III: Robots. Andrew Ortony, Donald A. Norman, and William Revelle, in their chapter, and Aaron Sloman, Ron Chrisley, and Matthias Scheutz, in theirs, contribute to the general analysis of a cognitive architecture of relevance both to psychological theorizing and to the development of AI in general and robots in particular. Ortony, Norman, and Revelle focus on the interplay of affect, motivation, and cognition in controlling behavior. Each is considered at three levels of information processing: the *reactive* level is primarily hard-wired; the *routine* level provides unconscious, uninterpreted expectations and automatized activity; and the *reflective* level supports higher-order cognitive functions, including meta-cognition, consciousness, self-reflection, and "full-fledged" emotions. Personality is then seen as a self-tunable system for the temporal patterning of affect, motivation, cognition, and behavior. The claim is that computational artifacts equipped with this architecture to perform unanticipated tasks in unpredictable environments will have emotions as the basis for achieving effective social functioning, efficient learning and memorization, and effective allocation of attention. Sloman, Chrisley, and Scheutz show how architecture-based concepts can extend and refine our pre-theoretical concepts of motivation, emotion, and affects. In doing so, they caution us that different information-processing architectures will support different classes of emotion, consciousness, and perception and that, in particular, different classes of robots may exhibit emotions very different from our own. They offer the CogAff schema as a general characterization of the types of component that may occur in a cognitive architecture and

sketch H-CogAff, an instance of the CogAff schema which may replicate human mental phenomena and enrich research on human emotions. They stress that robot emotions will emerge, as they do in humans, from the interactions of many mechanisms serving different purposes, not from a particular, dedicated "emotion mechanism."

Ronald C. Arkin sees emotions as a subset of motivations that provide support for an agent's survival in a complex world. He sees motivation as leading generally to the formulation of concrete goal-achieving behavior, whereas emotions are concerned with modulating existing behaviors in support of current activity. The study of a variety of human and nonhuman animal systems for motivation and emotion is seen to inspire schemes for behavior-based control for robots ranging from hexapods to wheeled robots to humanoids. The discussion moves from the sowbug to the praying mantis (in which fear, hunger, and sex affect the selection of motivated behaviors) to the use of canine ethology to design dog-like robots that use their emotional and motivational states to bond with their human counterparts. These studies ground an analysis of personality traits, attitudes, moods, and emotions.

Cynthia Breazeal and Rodney Brooks focus on human–robot interaction, examining how emotion-inspired mechanisms can enable robots to work more effectively in partnership with people. They demonstrate the cognitive and emotion-inspired systems of their robot, Kismet. Kismet's cognitive system enables it to figure out what to do, and its emotion system helps it to do so more flexibly in the human environment as well as to behave and interact with people in a socially acceptable and natural manner. They downplay the question of whether or not robots could have and feel human emotions. Rather, they speak of robot emotions in a functional sense, serving a pragmatic purpose for the robot that mirrors their natural analogs in human social interactions.

Emotions play a significant role in human teamwork. Ranjit Nair, Milind Tambe, and Stacy Marsella are concerned with the question of what happens to this role when some or all of the agents, that is, interacting intelligences, on the team are replaced by AI. They provide a short survey of the state of the art in multiagent teamwork and in computational models of emotions to ground their presentation of the effects of introducing emotions in three cases of teamwork: teams of simulated humans, agent–human teams, and pure agent teams. They also provide preliminary experimental results illustrating the impact of emotions on multiagent teamwork.

Part IV: Conclusions. One of the editors gets the final say, though some readers may find it useful to read our chapter as part of the opening perspective to provide a further framework for their own synthesis of the ideas presented in the chapters in Parts II and III. (Indeed, some readers may also

prefer to read Part III before Part II, to gain some sense of the state of play in "emotional AI" first and then use it to probe the biological database that Part II provides.)

Michael A. Arbib warns us to "Beware the Passionate Robot," noting that almost all of the book stresses the positive contribution of emotions, whereas personal experience shows that emotions "can get the better of one." He then enriches the discussion of the evolution of emotions by drawing comparisons with the evolution of vision and the evolution of language before returning to the issue of whether and how to characterize emotions in such a way that one might say a robot has emotions even though they are not empathically linked to human emotions. Finally, he reexamines the role of mirror neurons in Jeannerod's account of emotion, agency, and social coordination by suggesting parallels between their role in the evolution of language and ideas about the evolution of consciousness, feelings, and empathy.

In these ways, the book brings together the state of the art of research on the neuroscience and AI approaches to emotion in an effort to understand why humans and other animals have emotion and the various ways that emotion may factor into robotics and cognitive architectures of the future. The contributors to this book have their own answers to the question "Who needs emotions?" It is our hope that through an appreciation of these different views, readers will gain their own comprehensive understanding of why humans have emotion and the extent to which robots should and will have them.

Jean-Marc Fellous
La Jolla, CA

Michael A. Arbib
La Jolla and Los Angeles, CA

Contents

Contributors

Ralph Adolphs
Division of Humanities and Social Sciences
California Institute of Technology
Pasadena, CA 91125, USA
radolphs@hss.caltech.edu

Michael A. Arbib
Computer Science, Neuroscience, and USC Brain Project
University of Southern California
3614 Watt Way
Los Angeles, CA 90089-2520, USA
arbib@pollux.usc.edu

Ronald C. Arkin
Mobile Robot Laboratory
College of Computing
Georgia Institute of Technology
Atlanta, GA, 30332-0280, USA
arkin@cc.gatech.edu

Cynthia Breazeal
MIT Media Laboratory
20 Ames Street
E1S-449
Cambridge, MA 02139, USA
cynthia@media.mit.edu

Rodney Brooks
MIT Artificial Intelligence Laboratory
200 Technology Square
Cambridge, MA 02139, USA
brooks@csail.mit.edu

Ron Chrisley
Department of Informatics
University of Sussex
Falmer, BN1 9QH, United Kingdom
R.L.Chrisley@cogs.susx.ac.uk

Jean-Marc Fellous
Department of Biomedical Engineering
Duke University
136 Hudson Hall
P.O. Box 90281
Durham, NC 27708-0281, USA
fellous@duke.edu

Marc Jeannerod
Institut des Sciences Cognitives
67, boulevard Pinel
69675 Bron cedex, France
jeannerod@isc.cnrs.fr

Ann E. Kelley
Department of Psychiatry and Neuroscience Training Program
University of Wisconsin-Madison Medical School
6001 Research Park Boulevard
Madison, WI 53705, USA
aekelley@wisc.edu

Joseph E. LeDoux
Center for Neural Sciences
New York University
6 Washington Place
New York, NY 10003, USA
ledoux@cns.nyu.edu

Stacy Marsella
Information Sciences Institute
University of Southern California
4676 Admiralty Way, #1001
Marina del Rey, CA 90292, USA
marsella@isi.edu

Ranjit Nair
Computer Science Department
University of Southern California
941 W. 37th Place
Los Angeles, CA 90089, USA
nair@usc.edu

Donald A. Norman
Department of Computer Science
Northwestern University
1890 Maple Avenue,
Evanston, IL 60201-3150, USA
norman@northwestern.edu

Andrew Ortony
Departments of Computer Science and Psychology and School of Education
Northwestern University
2020 North Campus Drive
Evanston, IL 60208, USA
ortony@northwestern.edu

William Revelle
Department of Psychology
Northwestern University
2029 Sheridan Road
Evanston, IL 60208-2710, USA
revelle@northwestern.edu

Edmund T. Rolls
Department of Experimental Psychology
University of Oxford
South Parks Road
Oxford, OX1 3UD,
United Kingdom
Edmund.Rolls@psy.ox.ac.uk

Matthias Scheutz
Department of Computer Science
and Engineering
351 Fitzpatrick Hall
University of Notre Dame
Notre Dame, IN 46556, USA
Matthias.Scheutz.1@nd.edu

Aaron Sloman
School of Computer Science
University of Birmingham,
Birmingham, B15 2TT,
United Kingdom
A.Sloman@cs.bham.ac.uk

Milind Tambe
Computer Science Department and
Information Sciences Institute
University of Southern California
941 W. 37th Place
Los Angeles CA 90089, USA
tambe@usc.edu

This page intentionally left blank

PART I

PERSPECTIVES

This page intentionally left blank

1 "Edison" and "Russell"

Definitions versus Inventions in the Analysis of Emotion

JEAN-MARC FELLOUS AND
MICHAEL A. ARBIB

Editors' Note: Edison and Russell met at the Society for Neuroscience meeting. Russell, energized by his recent conversations with McCulloch and Pitts, discovered in himself a new passion for the logics of the brain, while Edison could not stop marveling at the perfection and complexity of this electrochemical machine. Exhausted by 5 days among the multitudes, they found themselves resting at a café outside the convention center and started chatting about their impressions of the meeting. Edison, now an established roboticist, and Russell, newly a theoretical neurobiologist, soon came to the difficult topic of emotion.

Russell suggested that "It would be useful to have a list of definitions of key terms in this subject—*drive*, *motivation*, and *emotion* for starters—that also takes account of logical alternative views. For example, I heard Joe LeDoux suggest that basic emotions did not involve feelings, whereas I would suggest that emotions do indeed include feelings and that 'emotions without feelings' might be better defined as drives!" Edison replied that he would rather build a useful machine than give it a logical definition but prompted Russell to continue and elaborate, especially on how his view could be of use to the robotics community.

RUSSELL: I confess that I had in mind definitions that best reflect on the study of the phenomenon in humans and other animals. However, I could also imagine a more abstract definition that could help you by providing criteria for investigating whether or not a robot or other machine exhibits, or might in the future exhibit, emotion. One could even investigate whether a community (the bees in a hive, the people of a country) might have emotion.

EDISON: One of the dangers in defining terms such as *emotion* is to bring the focus of the work on linguistic issues. There is certainly nothing wrong with doing so, but I don't think this will lead anywhere useful!

RUSSELL: There's nothing particularly linguistic in saying what you mean by *drive*, *motivation*, and *emotion*. Rather, it sets the standard for intellectual clarity. If one cannot articulate what one means, why write at all? However, I do understand—and may Whitehead forgive me—that we cannot ask for definitions in predicate logic. Nonetheless, I think to give at least an informal sense of what territory comes under each term is necessary and useful.

EDISON: Even if we did have definitions for *motivation* and *emotion*, I think history has shown that there couldn't be a consensus, so I assume that's not what you would be looking for. At best we could have "working definitions" that the engineer can use to get on with his work rather than definitions that constrain the field of research.

Still, I am worried about the problem of the subjectivity of the definitions. What I call *fear* (being electrocuted by an alternating current) is different from what you call *fear* (being faced with a paradox, such as defining a set of all sets that are not members of themselves!). We could compare definitions: I will agree with some of the definition of A, disagree with part of B, and so on. But this will certainly weaken the definition and could confuse everyone!

RUSSELL: I think researchers will be far more confused if they assume that they are talking about the same thing when they use the word *emotion* and they are not! Thus, articulating what one means seems to me crucial.

EDISON: In any case, most of these definitions will be based on a particular system—in my robot, fear cannot be expressed as "freezing" as it is for rats, but I agree with the fact that fear does not need to be "conscious." Then, we have to define *freezing* and *conscious*, and I am afraid we will get lost in endless debates, making the emotion definition dependent on a definition of *consciousness* and so on.

RUSSELL: But this is precisely the point. If one researcher sees emotions as essentially implying consciousness, then how can robots have emotions? One then wishes to press that researcher to understand if there is a sense of consciousness that can be ascribed to robots or whether robots can only have drives or not even that.

EDISON: If a particular emotion depends on consciousness, then a roboticist will have to think of what *consciousness* means for that particular robot. This will force the making of (necessarily simplifying) hypotheses that will go back to neuroscientists and force them to define *consciousness*. But how useful is a general statement such as "fear includes feelings, and hence consciousness"? Such a statement hides so many exceptions and particulars. Anyway, as a congressman once said "I do not need to define pornography, I know it when I see it." Wouldn't this apply to (human) emotions? I would argue that rather than defining *emotion* or *motivation* or *feelings*, we should instead ask for a clear explanation for what the particular emotion/motivation/feeling is "for" and ask for an operational view.

RUSSELL: All I ask is enough specificity to allow meaningful comparison between different approaches to humans, animals, and machines. Asking what an emotion/motivation/feeling is for is a fine start, but I do not think it will get you far! One still needs to ask "Do all your examples of emotion include feelings or not?" And if they include feelings, how can you escape discussions of consciousness?

EDISON: Why is this a need? The answer is very likely to be "no," and then what?

RUSSELL: You say you want to be "operational," but note that for the animal the operations include measurements of physiological and neurophysiological data, while human data may include not only comparable measurements (GSR, EEG, brain scans, etc.) but also verbal reports. Which of these measurements and reports are essential to the author's viewpoint? Are biology and the use of language irrelevant to our concerns? If they are relevant (and of course they are!), how do we abstract from these criteria those that make the discussion of emotion/motivation in machines nontrivial?

EDISON: It occurs to me that our difference of view could be essentially technical: I certainly have an engineering approach to the problem of emotion ("just do it, try things out with biology as guidance, generate hypotheses, build the machine and see if/how it works . . ."), while you may have a more theoretical approach ("first crisply define what you mean, and then implement the definition to test/refine it")?

RUSSELL: I would rather say that I believe in dialectic. A theory rooted in too small a domain may rob us of general insights. Thus, I am not suggesting that we try to find the one true definition of emotion a priori, only that each of us should be clear about what we think we mean or, if you prefer, about the ways in which we use key terms. Then we can move on to shared definitions and refine our thinking in the process. I think that mere tinkering can make the use of terms like *emotion* or *fear* vacuous.

EDISON: Tinkering! Yes! This is what evolution has done for us! Look at the amount of noise in the system! The problem of understanding the brain is a problem of differentiating signal from noise and achieving robustness and efficiency! Not that the brain is the perfect organ, but it is one pretty good solution given the constraints!

Ideally, I would really want to see this happen. The neuroscientist would say "For rats, the fear at the sight of a cat is for the preservation of its *self* but the fear response to a conditioned tone is to prepare for inescapable pain." And note, different kinds of *fear*, different neural substrates, but same word!

RUSSELL: Completely unsatisfactory! How do we define *self* and *pain* in ways that even begin to be meaningful for a machine? For example, a machine may overheat and have a sensor that measures temperature as part of a feedback loop to reduce overheating, but a high temperature reading has nothing to do with pain. In fact, there are interesting neurological data on people who feel no pain, others who know that they are feeling pain but do not care about it, as well as people like us. And then there are those unlucky few who have excruciating pain that is linked to no adaptive need for survival.

EDISON: I disagree! Overheating is not human pain for sure (but what about fever?) but certainly "machine" pain! I see no problem in defining *self* and *pain* for a robot.

The self could be (at least in part) machine integrity with all functions operational within nominal parameters. And pain occurs with input from sensors that are tuned to detect nonnominal parameter changes (excessive force exerted by the weight at the end of a robot arm).

RUSSELL: Still unsatisfactory. In psychology, we know there are people with multiple selves—having one body does not ensure having one self. Conversely, people who lose a limb and their vision in a terrorist attack still have a self even though they have lost "machine integrity." And my earlier examples were to make clear that "pain" and detection of parameter changes are quite different. If I have a perfect local anesthetic but smell my skin burning, then I feel no pain but have sensed a crucial parameter change. True, we cannot expect all aspects of human pain to be useful for the analysis of robots, but it does no good to throw away crucial distinctions we have learned from the studies of humans or other animals.

EDISON: Certainly, there may be multiple selves in a human. There may be multiple selves in machines as well! Machine integrity can (and should) change. After an injury such as the one you describe, all parameters of the robot have to be readjusted, and a new self is formed. Isn't it the case in humans as well? I would argue that the selves of a human before and after losing a limb and losing sight are different! You are not "yourself" anymore!

Inspired by what was learned with fear in rats, a roboticist would say "OK! My walking robot has analogous problems: encountering a predator—for a mobile robot, a car or truck in the street—and reacting to a low battery state, which signals the robot to prepare itself for functioning in a different mode, where energy needs to be saved." Those two robot behaviors are very similar to the rat behaviors in the operational sense that they serve the same kind of purpose. I think we might just as well call them "fear" and "pain." I would argue that it does not matter what I call them—the roboticist can still be inspired by their neural implementations and design the robotic system accordingly.

"Hmm, the amygdala is common to both behaviors and receives input from the hypothalamus (pain) and the LGN (perception). How these inputs are combined in the amygdala is unknown to neuroscientists, but maybe I should link the perceptual system of my robot and the energy monitor system. I'll make a subsystem that modulates perception on the basis of the amount of energy available: the more energy, the more objects perceptually analyzed; the less energy, only the most salient (with respect to the goal at hand) objects are analyzed."

The neuroscientist would reply: "That's interesting! I wonder if the amygdala computes something like salience. In particular, the hypothalamic inputs to the amygdala might modulate the speed of processing of the LGN inputs. Let's design an experiment." And the loop is closed!

RUSSELL: I agree with you that that interaction is very much worthwhile, but only if part of the effort is to understand what the extra circuitry adds. In particular, I note that you are still at the level of "emotions without feelings," which I would rather call "motivation" or "drive." At this level, we can ask whether the roboticist learns to make avoidance behavior more effective by studying animals. And it is interesting to ask if the roboticist's efforts will reveal the neural architecture as in some sense essential to all successful avoidance systems or as a biologically historical accident when one abstracts the core functionality away from the neuroanatomy, an abstraction that would be an important contribution. But does this increment take us closer to understanding human emotions as we subjectively know them or not?

EDISON: I certainly agree with that, and I do think it does! One final point: aren't the issues we are addressing—can a robot have emotion, does a robot need emotion, and so on—really the same issues as with animals and emotions—can an animal have emotion, does an animal need emotion?

RUSSELL: It will be intriguing to see how far researchers will go in answering all these questions and exploring the analogies between them.

Stimulated by this conversation, Edison and Russell returned to the poster sessions, after first promising to meet again, at a robotics conference.

This page intentionally left blank

2 Could a Robot Have Emotions?

Theoretical Perspectives from Social Cognitive Neuroscience

RALPH ADOLPHS

Could a robot have emotions? I begin by dissecting the initial question, and propose that we should attribute emotions and feelings to a system only if it satisfies criteria in addition to mere behavioral duplication. Those criteria require in turn a theory of what emotions and feelings are. Some aspects of emotion depend only on how humans react to observing behavior, some depend additionally on a scientific account of adaptive behavior, and some depend also on how that behavior is internally generated. Roughly, these three aspects correspond to the social communicative, the adaptive/regulatory, and the experiential aspects of emotion. I summarize these aspects in subsequent sections. I conclude with the speculation that robots could certainly interact socially with humans within a restricted domain (they already do), but that correctly attributing emotions and feelings to them would require that robots are situated in the world and constituted internally in respects that are relevantly similar to humans. In particular, if robotics is to be a science that can actually tell us something new about what emotions are, we need to engineer an internal processing architecture that goes beyond merely fooling humans into judging that the robot has emotions.

HOW COULD WE TELL IF A ROBOT HAD EMOTIONS AND FEELINGS?

Could a robot have emotions? Could it have feelings? Could it interact socially (either with others of its kind or with humans)?

Here, I shall argue that robots, unlike animals, could certainly interact socially with us in the absence of emotions and feelings to some limited extent; probably, they could even be constructed to have emotions in a narrow sense in the absence of feelings. However, such constructions would always be rather limited and susceptible to breakdown of various kinds. A different way to construct social robots, robots with emotions, is to build in feelings from the start—as is the case with animals. Before beginning, it may be useful to situate the view defended here with that voiced in some of the other chapters in this volume. Fellous and LeDoux, for example, argue, as LeDoux (1996) has done previously, for an approach to emotion which occurs primarily in the absence of feeling: emotion as behavior without conscious experience. Rolls has a similar approach (although neither he nor they shuns the topic of consciousness): emotions are analyzed strictly in relation to the behavior (as states elicited by stimuli that reinforce behavior) (Rolls, 1999).

Of course, there is nothing exactly wrong with these approaches as an analysis of complex behavior; indeed, they have been enormously useful. However, I think they start off on the wrong foot if the aim is to construct robots that will have the same abilities as people. Two problems become acute the more these approaches are developed. First, it becomes difficult to say what aspect of behavior is emotional and what part is not. Essentially any behavior might be recruited in the service of a particular emotional state, depending on an organism's appraisal of a particular context. Insofar as all behavior is adaptive and homeostatic in some sense, we face the danger of making the topic of emotion no different from that of behavior in general. Second, once a behaviorist starting point has been chosen, it becomes impossible to recover a theory of the conscious experience of emotion, of feeling. In fact, feeling becomes epiphenomenal, and at a minimum, this certainly violates our intuitive concept of what a theory of emotion should include.

I propose, then, to start, in some sense, in reverse—with a system that has the capacity for feelings. From this beginning, we can build the capacity for emotions of varying complexity and for the flexible, value-driven social behavior that animals exhibit. Without such a beginning, we will always be mimicking only aspects of behavior. To guide this enterprise, we can ask ourselves what criteria we use to assign feelings and emotions to other people. If our answer to this question indicates that more than the right appearances are required, we will need an account of how emotions, feelings, and social

behavior are generated within humans and other animals, an account that would provide a minimal set of criteria that robots would need to meet in order to qualify as having emotions and feelings.

It will seem misguided to some to put so much effort into a prior understanding of the mechanisms behind biological emotions and feelings in our design of robots that would have those same states. Why could we not simply proceed to tinker with the construction of robots with the sole aim of producing behaviors that humans who interact with them will label as "emotional?" Why not have as our aim solely to convince human observers that robots have emotions and feelings because they behave as though they do?

There are two initial comments to be made about this approach and a third one that depends more on situating robotics as a science. The attempt to provide a criterion for the possession of central mental or cognitive states solely by reproduction of a set of behavioral features is of course the route that behaviorism took (which simply omitted the central states). It is also the route that Alan Turing took in his classic paper, "Computing Machinery and Intelligence" (Turing, 1950). In that paper, Turing considered the question "Could a machine think?" He ended up describing the initial question as meaningless and recommended that it be replaced by the now (in)famous Turing test: provided a machine could fool a human observer into believing that it was a human, on the basis of its overt behavior, we should credit the machine with the same intelligence with which we credit the human.

The demise of behaviorism provides testament to the failure of this approach in our understanding of the mind. In fact, postulating by fiat that behavioral equivalence guarantees internal state equivalence (or simply omitting all talk of the internal states) also guarantees that we cannot learn anything new about emotions and feelings—we have simply defined what they are in advance of any scientific exploration. Not only is the approach nonscientific, it is also simply implausible. Suppose you are confronted by such a robot that exhibits emotional behavior indistinguishable from that of a human. Let us even suppose that it looks indistinguishable from a human in all respects, from the outside. Would you change your beliefs upon discovering that its actions were in fact remote-controlled by other humans and that all it contained in its head were a bunch of radio receivers to pick up radio signals from the remote controllers? The obvious response would be "yes;" that is, there is indeed further information that would violate your background assumptions about the robot. Of course, we regularly use behavioral observations alone in order to attribute emotions and feelings to fellow humans (these are all we usually have to go by); but we have critical background assumptions that they are also like us in the relevant internal respects, which the robot does not share.

This, of course, raises the question "What if the robot were not remote-controlled?" My claim here is that if we had solved the problem of how to build such an autonomously emotional robot, we would have done so by figuring out the answer to another question, raised above: "Precisely which internal aspects are relevant?" Although we as yet do not know the answer to this empirical question, we can feel fairly confident that neither will radio transmitters do nor will we need to actually build a robot's innards out of brain cells. Instead, there will have to be some complex functional architecture within the robot that is functionally equivalent to what the brain achieves. This situates the relevant internal details at a level below that of radio transmitters but above that of actual organic molecules.

A second, separate problem with defining emotions solely on the basis of overt behaviors is that we do not conceptually identify emotions with behaviors. We use behaviors as indicators of emotions, but it is common knowledge that the two are linked only dispositionally and that the attempt to create an exhaustive list of all the contingencies that would identify emotions with behaviors under particular circumstances is doomed to failure. To be sure, there are some aspects of emotional response, such as startle responses, that do appear to exhibit rather rigid links between stimuli and responses. However, to the extent that they are reflexive, such behaviors are not generally considered emotions by emotion theorists: emotions are, in a sense, "decoupled reflexes." The idea here is that emotions are more flexible and adaptive under more unpredictable circumstances than reflexes. Their adaptive nature is evident in the ability to recruit a variety of behavioral responses to stimuli in a flexible way. Fear responses are actually a good example of this: depending on the circumstances, a rat in a state of fear will exhibit a flight response and run away (if it has evaluated that behavioral option as advantageous) or freeze and remain immobile (if it has evaluated that behavioral option as advantageous). Their very flexibility is also what makes emotions especially suited to guide social behavior, where the appropriate set of behaviors changes all the time depending on context and social background.

Emotions and feelings are states that are central to an organism. We use a variety of cues at our disposal to infer that an organism has a certain emotion or feeling, typically behavioral cues, but these work more or less well in humans because everything else is more or less equal in relevant respects (other humans are constituted similarly internally). The robot that is built solely to mimic behavioral output violates these background assumptions of internal constituency, making the extrapolations that we normally make on the basis of behavior invalid in that case.

I have already hinted at a third problem with the Turing test approach to robot emotions: that it effectively blocks any connection the discipline could have with biology and neuroscience. Those disciplines seek to under-

stand (in part) the internal causal mechanisms that constitute the central states that we have identified on the basis of behavioral criteria. The above comment will be sure to meet with resistance from those who argue that central states, like emotions, are theoretical constructs (i.e., attributions that we make of others in order to have a more compact description of patterns in their behavior). As such, they need not correspond to any isomorphic physiological state actually internal to the organism. I, of course, do not deny that in some cases we do indeed make such attributions to others that may not correspond to any actual physical internal state of the same kind. However, the obvious response would be that if the central states that we attribute to a system are in fact solely our explanations of its behavior rather than dependent on a particular internal implementation of such behavior, they are of a different ontological type from those that we can find by taking the system apart. Examples of the former are functional states that we assign to artifacts or to systems generally that we are exploiting toward some use. For example, many different devices could be in the state "2 P.M." if we can use them to keep time; nothing further can be discovered about time keeping in general by taking them apart. Examples of the latter are states that can be identified with intrinsic physical states. Emotions, I believe, fall somewhere in the middle: you do not need to be made out of squishy cells to have emotions, but you do need more than just the mere external appearance of emotionally triggered behavior.

Surely, one good way to approach the question of whether or not robots can have these states is to examine more precisely what we know about ourselves in this regard. Indeed, some things could be attributed to robots solely on the basis of their behavior, and it is in principle possible that they could interact with humans socially to some extent. However, there are other things, notably feelings, that we will not want to attribute to robots unless they are internally constituted like us in the relevant respects. Emotions as such are somewhere in the middle here—some aspects of emotion depend only on how humans react to observing the behavior of the robot, some depend additionally on a scientific account of the robot's adaptive behavior, and some depend also on how that behavior is internally generated. Roughly, these three aspects correspond to the social communicative, the adaptive/regulatory, and the experiential aspects of an emotion.

WHAT IS AN EMOTION?

Neurobiologists and psychologists alike have conceptualized an emotion as a concerted, generally adaptive, phasic change in multiple physiological systems (including both somatic and neural components) in response to the value

of a stimulus (e.g., Damasio, 1999; Lazarus, 1991; Plutchik, 1980; see Scherer, 2000, for a review). An important issue, often overlooked, concerns the distinction between the emotional reaction (the physiological emotional response) and the feeling of the emotion (presumed in some theories to rely on a central representation of this physiological emotional response) (Damasio, 1999). It is also essential to keep in mind that an emotional response typically involves concerted changes in a very large number of somatic parameters, including endocrine, visceral, autonomic, and musculoskeletal changes such as facial expression, all of which unfold in a complex fashion over time.

Despite a long history of philosophical debate on this issue, emotions are indeed representational states: they represent the value or significance that the sets of sensory inputs and behavioral outputs have for the organism's homeostasis. As such, they involve mappings of body states in structures such as brain stem, thalamic, and cortical somatic and visceral sensory regions. It should be noted that it is not necessary to map an actual body state; only the result matters. Thus, it would be possible to have a "somatic image," in much the same way one has a visual image, and a concomitant feeling. Such a somatic image would supervene only on the neural representation of a body state, not on an actual body state.

In order to derive a framework for thinking about emotions, it is useful to draw upon two different theories (there are others that are relevant, but these two serve as a starting point). One theory, in line with both an evolutionary approach to emotion as well as aspects of appraisal theory, concerns the domain of information that specifies emotion processing. In short, emotions concern, or derive from, information that is of direct relevance to the homeostasis and survival of an organism (Damasio, 1994; Darwin, 1965; Frijda, 1986), that is, the significance that the situation has for the organism, both in terms of its immediate impact and in terms of the organism's plans and goals in responding to the situation (Lazarus, 1991). Fear and disgust are obvious examples of such emotions. The notion of homeostasis and survival needs also to be extended to the social world, to account for social emotions, such as shame, guilt, or embarrassment, that regulate social behavior in groups. It furthermore needs to be extended to the culturally learned appraisal of stimuli (different stimuli will elicit different emotions in people from different cultures to some extent because the stimuli have a different social meaning in the different cultures), and it needs to acknowledge the extensive self-regulation of emotion that is featured in adult humans. All of these make it extremely complex to define the categories and the boundaries of emotion, but they still leave relatively straightforward the paradigmatic issue with which emotion is concerned: the value of a stimulus or of a behavior—value to the organism's own survival or to the survival of its offspring, relatives, or larger social group.

This first point, the domain specificity of emotional information, tells us what distinguishes emotion processing from information processing in general but leaves open two further questions: how broadly should we construe this domain, and how is such specificity implemented? In regard to the former question, the domain includes social and basic emotions but also states such as pain, hunger, and any other information that has a bearing on survival. Is this too broad? Philosophers can and do worry about such distinctions, but for the present, we as neuroscientists can simply acknowledge that indeed the processing of emotions should (and, as it turns out, does) share mechanisms with the processing of thirst, hunger, pain, sex, and any other category of information that motivates behavior (Panksepp, 1998; Rolls, 1999). In regard to the latter question, the implementation of value-laden information will require information about the perceptual properties of a stimulus to be associated with information about the state of the organism perceiving that stimulus. Such information about the organism could be sensory (somatosensory in a broad sense, i.e., information about the impact that the stimulus has on homeostasis) or motor (i.e., information about the action plans triggered by the stimulus). This brings us to the second of the two emotion theories I mentioned at the outset.

The first emotion theory, then, acknowledges that emotion processing is domain-specific and relates to the value that a stimulus has for an organism, in a broad sense. The second concerns the cause-and-effect architecture of behavior, bodily states, and central states. Readers will be familiar with the theories of William James, Walter Cannon, and later thinkers, who debated the primacy of bodily states (Cannon, 1927; James, 1884). Is it that we are afraid first and then run away from the bear, or do we have an emotional bodily response to the bear first, the perception of which in turn constitutes our feeling afraid? James believed the latter; Cannon argued for the former. This debate has been very muddled for at least two reasons: the failure to distinguish emotions from feelings and the ubiquitous tendency for a single causal scheme.

It is useful to conceive of emotions as central states that are only dispositionally linked to certain physiological states of the body, certain behaviors, or certain feelings of which we are aware. An emotion is thus a neurally implemented state (or, better, a collection of processes) that operates in a domain-specific manner on information (viz., it processes biological value to guide adaptive behavior). However, the mechanism behind assigning value to such information depends on an organism's reactive and proactive responses to the stimulus. The proactive component prepares the organism for action, and the reactive component reflects the response to a stimulus. It is the coordinated web of action preparations, stimulus responses, and an organism's internal mapping of these that constitutes a central emotional state. Viewed this way,

an emotion is neither the cause nor consequence of a physiological response: it emerges in parallel with an organism's interaction with its environment, in parallel with physiological response, and in parallel with feeling. Behavior, physiological response, and feeling causally affect one another; and none of them in isolation is to be identified with the emotion, although we certainly use observations of them to infer an emotional state.

In addition to the question "What is an emotion?" there is a second, more fine-grained question: "What emotions are there?" While the majority of research on facial expression uses the emotion categories for which we have names in English (in particular, the "basic" emotions, e.g., happiness, surprise, fear, anger, disgust, and sadness) or, somewhat less commonly, a dimensional approach (often in terms of arousal/valence), there are three further frameworks that are worth exploring in more detail. Two of these arose primarily from animal studies. A scheme proposed by Rolls (1999) also maps emotions onto a two-dimensional space, as do some other psychological proposals; but in this case the dimensions correspond to the presentation or omission of reinforcers: roughly, presentation of reward (pleasure, ecstasy), presentation of punishment (fear), withholding of reward (anger, frustration, sadness), or withholding of punishment (relief). A similar, more psychological scheme has been articulated by Russell (2003) in his concept of "core affect," although he has a detailed scheme for how emotion concepts are constructed using such core affect as one ingredient. Another scheme, from Panksepp (1998), articulates a neuroethologically inspired framework for categorizing emotions; according to this scheme, there are neural systems specialized to process classes of those emotions that make similar requirements in terms of the types of stimulus that trigger them and the behaviors associated with them (specifically, emotions that fall under the four broad categories of seeking, panic, rage, and fear). Both of these approaches (Panksepp, 1998; Rolls, 1999) appear to yield a better purchase on the underlying neurobiological systems but leave unclear how exactly such a framework will map onto all the diverse emotions for which we have names (especially the social ones). A third approach takes a more fine-grained psychological analysis of how people evaluate an emotional situation and proposes a set of "stimulus evaluation checks" that can trigger individual components of an emotional behavior, from which the concerted response is assembled as the appraisal of the situation unfolds (Scherer, 1984, 1988). This latter theory has been applied to facial expressions with some success (Wehrle, Kaiser, Schmidt, & Scherer, 2000). While rather different in many respects, all three of these frameworks for thinking about emotion share the idea that our everyday emotion categories are probably not the best suited for scientific investigation.

It is worth considering the influences of culture on emotions at this point. Considerable work by cultural psychologists and anthropologists has shown

that there are indeed large and sometimes surprising differences in the words and concepts (Russell, 1991; Wierzbicka, 1999) that different cultures have for describing emotions, as well as in the social circumstances that evoke the expression of particular emotions (Fridlund, 1994). However, those data do not actually show that different cultures have different emotions, if we think of emotions as central, neurally implemented states. As for, say, color vision, they just say that, despite the same internal processing architecture, how we interpret, categorize, and name emotions varies according to culture and that we learn in a particular culture the social context in which it is appropriate to express emotions. However, the emotional states themselves are likely to be quite invariant across cultures (Panksepp, 1998; Russell, Lewicka, & Niit, 1989). In a sense, we can think of a basic, culturally universal emotion set that is sculpted by evolution and implemented in the brain, but the links between such emotional states and stimuli, behavior, and other cognitive states are plastic and can be modified by learning in a specific cultural context.

Emotional information processing depends on a complex collection of steps implemented in a large number of neural structures, the details of which have been recently reviewed. One can sketch at least some components of this architecture as implementing three serial processing steps: (1) an initial perceptual representation of the stimuli (or a perceptual representation recollected from memory), (2) a subsequent association of this perceptual representation with emotional response and motivation, and (3) a final sensorimotor representation of this response and our regulation of it. The first step draws on higher-order sensory cortices and already features some domain-specific processing: certain features of stimuli that have high signal value are processed by relatively specialized sectors of cortex, permitting the brain to construct representations of socially important information rapidly and efficiently. Examples include regions of extrastriate cortex that are specialized for processing faces or biological motion. Such modularity is most evident in regard to classes of stimuli that are of high value to an organism (and hence drove the evolution of relatively specialized neural systems for their processing), for example, socially and emotionally salient information. The second step draws on a system of structures that includes amygdala, ventral striatum, and regions in medial and ventral prefrontal cortex, all three of which are extensively and bidirectionally interconnected. This set of structures receives sensory information from the previously described step and (1) can participate in perceptual processing via feedback to those regions from which input was received (e.g., by attentional modulation of visual perception on the basis of the emotional/social meaning of the stimulus), (2) can trigger coordinated emotional responses (e.g., autonomic and endocrine responses as well as modulation of reflexes), and (3) can modulate other

cognitive processes such as decision making, attention, and memory. The third step finally encompasses an organism's internal representation of what is happening to it as it is responding to a socially relevant stimulus. This step generates social knowledge, allows us to understand other people in part by simulating what it is like to be them, and draws on motor and somatosensory-related cortices.

EMOTIONS AND SOCIAL COMMUNICATION

The idea that emotions are signals that can serve a role in social communication, especially in primates, was of course noted already by Darwin in his book *The Expressions of Emotions in Man and Animals* (Darwin, 1965). While perhaps the most evolutionarily recent aspect of emotion, social communication also turns out to be the one easiest to duplicate in robots. The easiest solution is to take an entirely pragmatic approach to the problem: to construct robots that humans will relate to in a certain, social way because the robots are designed to capitalize on the kinds of behavior and signal that we normally use to attribute emotional and social states to each other. Thus, a robot with the right external interface can be made to smile, to frown, and so on as other chapters in this volume illustrate (cf. Brezeal and Brooks, Chapter 10). In order to be convincing to people, these signals must of course be produced at the right time, in the right context, etc. It is clear that considerable sophistication would be required for a robot to be able to engage socially with humans over a prolonged period of time in an unconstrained context. Indeed, as mentioned earlier, the strong intuition here would be that if all we pay attention to is the goal of fooling human observers (as Turing did in his paper and as various expert systems have done since then), then sooner or later we will run into some unanticipated situation in which the robot will reveal to us that it is merely designed to fool us into crediting it with internal states so that we can interact socially with it; that is, sooner or later, we should lose our faith in interacting with the robot as with another person and think of the machine as simply engaging us in a clever deception game. Moreover, as noted at the beginning of this chapter, such an approach could perhaps help in the investigation of the different perceptual cues humans use to attribute emotions to a system, but it seems misguided if we want to investigate emotions themselves. It is conceivable that we might someday design robots that convince humans with whom they interact that they have emotions. In that case, we will have either learned how to build an internal architecture that captures some of the salient functional features of biological emotion reviewed here, or designed a system that happens to be able to fool humans into (erroneously) believing that it has emotions.

The direction in which to head in order to construct artificial systems that are resilient to this kind of breakdown and that can tell us something new about emotion itself is to go beyond the simulation of mere external behavior and to pay attention to the mechanisms that generate such behavior in real organisms. Robotics has in fact recently taken such a route, in large part due to the realization that its neglect results in systems whose behavior is just too rigid and breaks down in unanticipated cases. The next steps, I believe, are to look at feelings, then at emotions, and finally the social behavior that they help regulate. Roughly, if you build in the feelings, the emotions and the social behavior follow more easily.

The evidence that social communication draws upon feeling comes from various avenues. Important recent findings are related to simulation, as reviewed at length in Chapter 6 (Jeannerod). Data ranging from neurophysiological studies in monkeys (Gallese & Goldman, 1999) to lesion studies in humans (Adolphs, 2002) support the idea that we figure out how other people feel, in part, by simulating aspects of their presumed body state and that such a mechanism plays a key role in how we communicate socially. Such a mechanism would simulate in the observer the state of the person observed by estimating the motor representations that gave rise to the behavior. Once we have generated the state that we presume the other person to share, a representation of this actual state in ourselves could trigger conceptual knowledge. Of course, this is not the only mechanism whereby we obtain information about the mental states of others; inference-based reasoning strategies and a collection of abilities dubbed "theory of mind" participate in this process as well.

The simulation hypothesis has recently received considerable attention due to experimental findings that appear to support it. In the premotor cortex of monkeys, neurons that respond not only when the monkey prepares to perform an action itself but also when it observes the same visually presented action performed by another have been reported (Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; Gallese & Goldman, 1999; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996). Various supportive findings have also been obtained in humans: observing another's actions results in desynchronization in motor cortex as measured with magnetoencephalography (Hari et al., 1998) and lowers the threshold for producing motor responses when transcranial magnetic stimulation is used to activate motor cortex (Strafella & Paus, 2000); imitating another's actions via observation activates premotor cortex in functional imaging studies (Iacoboni et al., 1999); moreover, such activation is somatotopic with respect to the body part that is observed to perform the action, even in the absence of any overt action on the part of the subject (Buccino et al., 2001). It thus appears that primates construct motor representations suited to performing the same action

that they visually perceive someone else perform, in line with the simulation theory.

The specific evidence that simulation may play a role also in recognition of the actions that accompany emotional states comes from disparate experiments. The experience and expression of emotion are correlated (Rosenberg & Ekman, 1994) and offer an intriguing causal relationship: production of emotional facial expressions (Adelman & Zajonc, 1989) and other somatovisceral responses (Cacioppo, Berntson, & Klein, 1992) results in changes in emotional experience. Producing a facial expression to command influences the feeling and autonomic correlates of the emotional state (Levenson, Ekman, & Friesen, 1990) as well as its electroencephalographic correlates (Ekman & Davidson, 1993). Viewing facial expressions in turn results in expressions on one's own face that may not be readily visible but can be measured with facial electromyography (Dimberg, 1982; Jaencke, 1994) and that mimic the expression shown in the stimulus (Hess & Blairy, 2001); moreover, such facial reactions to viewing facial expressions occur even in the absence of conscious recognition of the stimulus, for example to subliminally presented facial expressions (Dimberg, Thunberg, & Elmehed, 2000). Viewing the facial expression of another can thus lead to changes in one's own emotional state; this in turn would result in a remapping of one's own emotional state, that is, a change in feeling. While viewing facial expressions does indeed induce changes in feeling (Schneider, Gur, Gur, & Muenz, 1994; Wild, Erb, & Bartels, 2001), the mechanism could also operate without the intermediate of producing the facial expression, by direct modulation of the somatic mapping structures that generate the feeling (Damasio, 1994, 1999).

There is thus a collection of findings that provide strong support for the idea that expressing emotional behaviors in oneself and recognizing emotional behaviors in others automatically engage feelings. There are close correlations, following brain damage, between impairments in emotion regulation, social communication, and the ability to feel emotions. These correlations prompt the hypothesis that social communication and emotion depend to some extent on feelings (Adolphs, 2002).

Some have even proposed that emotions can occur only in a social context, as an aspect (real or vicarious) of social communication (Brothers, 1997). To some extent, this issue is just semantic, but emphasizing the social communicative nature of emotions does help to distinguish them from other motivational states with which they share much of the same neural machinery but that we would not normally include in our concept of emotion: such as hunger, thirst, and pain. Certainly, emotions play a very important role in social behavior, and some classes of emotions—the so-called social or moral emotions, such as embarrassment, jealousy, shame, and pride—can exist only in a social context. However, not all instances of all emotions are social: one can be afraid of falling off a cliff in the absence of any social context. Con-

versely, not all aspects of social communication are emotional: the lexical aspects of language are a good example.

EMOTION AND FEELING

What is a feeling? It would be impossible to do justice to this question within the scope of this chapter. Briefly, feelings are one (critical) aspect of our conscious experience of emotions, the aspect that makes us aware of the state of our body—and through it, often the state of another person's body. Sadness, happiness, jealousy, and sympathy are examples. We can be aware of much more than feelings when we experience emotions, but without feelings we do not have an emotional experience at all.

It is no coincidence that the verb *to feel* can be both transitive and intransitive. We feel objects in the external environment, and their impact on us modulates how we feel as a background awareness of the state of our body. Feeling emotions is no different: it consists in querying our body and registering the sensory answer obtained. It is both action and perception. This view of feeling has been elaborated in detail by writers such as Antonio Damasio (1999) and Jaak Panksepp (1998). Although they emphasize somewhat different aspects (Damasio the sensory end and Panksepp the action/motor end), their views converge with the one summarized above. It is a view that is finding resonance from various theorists in their accounts of consciousness in general: it is enactive, situated in a functional sense, and dependent on higher cortical levels querying lower levels in a reverse hierarchical fashion. One way of describing conscious sensory experience, for example, is as a skill in how we interact with the environment in order to obtain information about it. Within the brain itself, conscious sensory experience likewise seems to depend on higher-level processing regions sending signals to lower regions to probe or reconstruct sensory representations at those lower levels (cf. Pascual-Leone & Walsh, 2001, for a good example of such a finding). Feeling emotions thus consists of a probe, a question, and an input registered in response to that probe (Damasio, 1999). When we feel sad, for example, we do not become aware of some property of a mental representation of sadness; rather, the distributed activities of asking ourselves how we feel together with the information we receive generate our awareness that we feel sad.

What components does such a process require? It requires, at a minimum, a central model of ourselves that can be updated by such information and that can make information available globally to other cognitive processes. Let us take the features itemized below as prerequisites of possessing feelings (no doubt, all of them require elaboration and would need to be supplemented depending on the species).

- A self-model that can query certain states of the system itself as well as states of the external environment.
- Such a model is updated continuously; in fact, it depends on input that is related to its expectations. It thus maps prior states of the model and expectations against the information obtained from sensory organs. It should also be noted that, certainly in higher animals, the model is extremely detailed and includes information from a vast array of sources.
- The state of the self-model is made available to a host of other cognitive processes, both automatic and volitional. It thus guides information processing globally.
- The way in which states of the self-model motivate behaviors is arranged such that, globally, these states signal motivational value for the organism: they are always and automatically tied to survival and maintenance of homeostasis.

COULD A ROBOT HAVE EMOTIONS?

Our initial question points toward another: what is our intent in designing robots? It seems clear (in fact, it is already the case) that we can construct robots that behave in a sufficiently complex social fashion, at least under some restricted circumstances and for a limited time, that they cause humans with whom they interact to attribute emotions and feelings to them. So, if our purpose is to design robots toward which humans behave socially, a large part of the enterprise consists in paying attention to the cues on the basis of which human observers attribute agency, goal directedness, and so on. While a substantial part of such an emphasis will focus on how we typically pick out biological, goal-directed, intentional behavior, action, and agency in the world, another topic worth considering is the extent to which human observers could, over sufficient time, learn to make such attributions also on the basis of cues somewhat outside the normal range. That is, it may well be that even robots that behave somewhat differently from actual biological agents can be given such attributions; but in this case, the slack in human–computer social interaction is taken up by the human rather than by the computer. We can capitalize on the fact that humans are quite willing to anthropomorphize over all kinds of system that fall short of exhibiting actual human behavior.

What has concerned me in this chapter, however, is a different topic: not how to design robots that could make people believe that they have emotions, but how to construct robots that really do have emotions, in a sense autonomous from the beliefs attributed by a human observer (and in

the sense that we could find out something new about emotion without presupposing it). The former approach can tell us something about how humans attribute emotions on the basis of behavior; the latter can tell us something about how emotions actually regulate the behavior of a system. I have ventured that the former approach can never lead to real insight into the functions of emotion (although it can be useful for probing human perception and judgment), whereas the latter indeed forces us to grapple precisely with an account of what emotion and feeling are. I have further argued that taking the latter approach in fact guarantees success also for the former. This of course still leaves open the difficult question of exactly how we could determine that a system has feelings. I have argued that this is an empirical question; whatever the criteria turn out to be, they will involve facts about the internal processing architecture, not just passing the Turing test.

Building in self-representation and value, with the goal of constructing a system that could have feelings, will result in a robot that also has the capacity for emotions and for complex social behavior. This approach would thus not only achieve the desired design of robots with which humans can interact socially but also hold out the opportunity to teach us something about how feeling, emotion, and social behavior depend on one another and about how they function in humans and other animals.

I have been vague about how precisely to go about building a system that has feelings, aside from listing a few preliminary criteria. The reason for this vagueness is that we at present do not have a good understanding of how feelings are implemented in biological systems, although recent data give us some hints. However, the point of this chapter has been less to provide a prescription for how to go about building feeling robots than to suggest a general emphasis in the design of such robots. In short, neuroscientific investigations of emotions and feelings in humans and other animals should go hand-in-hand with designing artificial systems that have emotions and feelings: the two enterprises complement one another.

Acknowledgment. Supported in part by grants from the National Institutes of Health and the James S. McDonnell Foundation.

References

Adelman, P. K., & Zajonc, R. B. (1989). Facial efference and the experience of emotion. *Annual Review of Psychology*, *40*, 249–280.

Adolphs, R. (2002). Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Reviews*, *1*, 21–61.

Brothers, L. (1997). *Friday's footprint*. New York: Oxford University Press.
Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V. V., Seitz, R. J., Zilles, K., Rizzolatti, G., & Freund, H.-J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: An fMRI study. *European Journal of Neuroscience*, *13*, 400–404.
Cacioppo, J. T., Berntson, G. G., & Klein, D. J. (1992). What is an emotion? The role of somatovisceral afference, with special emphasis on somatovisceral "illusions." In M. S. Clark (Ed.), *Emotion and social behavior* (Vol. 14, pp. 63–98). Newbury Park, CA: Sage.
Cannon, W. B. (1927). The James-Lange theory of emotions: A critical examination and an alternative theory. *American Journal of Psychology*, *39*, 106–124.
Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Grosset/Putnam.
Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York: Harcourt Brace.
Darwin, C. (1965). *The expression of the emotions in man and animals*. Chicago: University of Chicago Press. (Original work published 1872)
Dimberg, U. (1982). Facial reactions to facial expressions. *Psychophysiology*, *19*, 643–647.
Dimberg, U., Thunberg, M., & Elmehed, K. (2000). Unconscious facial reactions to emotional facial expressions. *Psychological Science*, *11*, 86–89.
Ekman, P., & Davidson, R. J. (1993). Voluntary smiling changes regional brain activity. *Psychological Science*, *4*, 342–345.
Fridlund, A. J. (1994). *Human facial expression*. New York: Academic Press.
Frijda, N. H. (1986). *The emotions*. New York: Cambridge University Press.
Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*, 593–609.
Gallese, V., & Goldman, A. (1999). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*, 493–500.
Hari, R., Forss, N., Avikainen, S., Kirveskari, E., Salenius, S., & Rizzolatti, G. (1998). Activation of human primary motor cortex during action observation: a neuromagnetic study. *Proceedings of the National Academy of Sciences of the USA*, *95*, 15061–15065.
Hess, U., & Blairy, S. (2001). Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. *International Journal of Psychophysiology*, *40*, 129–141.
Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science*, *286*, 2526–2528.
Jaencke, L. (1994). An EMG investigation of the coactivation of facial muscles during the presentation of affect-laden stimuli. *Journal of Psychophysiology*, *8*, 1–10.
James, W. (1884). What is an emotion? *Mind*, *9*, 188–205.
Lazarus, R. S. (1991). *Emotion and adaptation*. New York: Oxford University Press.
LeDoux, J. (1996). *The emotional brain*. New York: Simon and Schuster.
Levenson, R. W., Ekman, P., & Friesen, W. V. (1990). Voluntary facial action gen-

erates emotion-specific autonomic nervous system activity. *Psychophysiology*, *27*, 363–384.

Panksepp, J. (1998). *Affective neuroscience*. New York: Oxford University Press.

Pascual-Leone, A., & Walsh, V. (2001). Fast backprojections from the motion to the primary visual area necessary for visual awareness. *Science*, *292*, 510–512.

Plutchik, R. (1980). *Emotion: a psychoevolutionary synthesis*. New York: Harper and Row.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*, 131–141.

Rolls, E. T. (1999). *The brain and emotion*. New York: Oxford University Press.

Rosenberg, E. L., & Ekman, P. (1994). Coherence between expressive and experiential systems in emotion. *Cognition and Emotion*, *8*, 201–230.

Russell, J. A. (1991). Culture and the categorization of emotions. *Psychological Bulletin*, *110*, 426–450.

Russell, J. A., Lewicka, M., & Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology*, *57*, 848–856.

Scherer, K. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to emotion*. Hillsdale, NJ: Erlbaum.

Scherer, K. R. (1988). Criteria for emotion-antecedent appraisal: A review. In V. Hamilton, G. H. Bower, & N. H. Frijda (Eds.), *Cognitive perspectives on emotion and motivation* (pp. 89–126). Dordrecht: Martinus Nijhoff.

Scherer, K. R. (2000). Psychological models of emotion. In J. C. Borod (Ed.), *The neuropsychology of emotion* (pp. 137–162). New York: Oxford University Press.

Schneider, F., Gur, R. C., Gur, R. E., & Muenz, L. R. (1994). Standardized mood induction with happy and sad facial expressions. *Psychiatry Research*, *51*, 19–31.

Strafella, A. P., & Paus, T. (2000). Modulation of cortical excitability during action observation: A transcranial magnetic stimulation study. *Experimental Brain Research*, *11*, 2289–2292.

Turing, A. (1950). Computing machinery and intelligence. Reprinted in Anderson, A. (1964). *Minds and machines*. Englewood Cliffs, NJ: Prentice-Hall.

Wehrle, T., Kaiser, S., Schmidt, S., & Scherer, K. R. (2000). Studying the dynamics of emotional expression using synthesized facial muscle movements. *Journal of Personality and Social Psychology*, *78*, 105–119.

Wierzbicka, A. (1999). *Emotions across languages and cultures*. Paris: Cambridge University Press.

Wild, B., Erb, M., & Bartels, M. (2001). Are emotions contagious? Evoked emotions while viewing emotionally expressive faces: Quality, quantity, time course and gender differences. *Psychiatry Research*, *102*, 109–124.

This page intentionally left blank

PART II

BRAINS

This page intentionally left blank

3 Neurochemical Networks Encoding Emotion and Motivation

An Evolutionary Perspective

ANN E. KELLEY

Specific and phylogenetically ancient motivational systems exist in the brain that have evolved over the course of millions of years to ensure adaptation and survival. These systems are engaged by perception of environmental events or stimuli, and when so engaged generate specific affective states (positive or negative emotions) that are powerful drivers of behavior. Positive emotions generally serve to bring the organism in contact with potentially beneficial resources—food, water, territory, mating or other social opportunities. Negative emotions serve to protect the organism from danger—mainly to ensure fight-or-flight responses, or other appropriate defensive strategies such as submissive behavior or withdrawal, protection of territory or kin, and avoidance of pain. Brain systems monitor the external and internal world for signals, and control the ebb and flow of these motivational states. Their elaboration and expression, when elicited by appropriate stimuli, are instantiated in complex but highly organized neural circuitry. Cross talk between cortical and subcortical networks enables intimate communication between phylogenetically newer brain regions, subserving subjective awareness and cognition (primarily cortex), and ancestral motivational systems that exist to promote survival behaviors (primarily hypothalamus). Neurochemical coding, imparting an extraordinary amount of

specificity and flexibility within these networks, appears to be conserved in evolution. This is exemplified by examining the role of dopamine in reward and plasticity, serotonin in aggression and depression, and opioid peptides in pain and pleasure. Moreover, across the course of thousands of years, humans, through interactions with plant alkyloids, have discovered how to facilitate or blunt emotions with psychoactive drugs. Thus, while neurochemical systems mediating emotion generally serve a highly functional and adaptive role in behavior, they can be altered in maladaptive ways in the case of addiction.

> In attempting to understand the elements out of which mental phenomena are compounded, it is of the greatest importance to remember that from the protozoa to man there is nowhere a very wide gap either in structure or in behavior.
>
> —Bertrand Russell (*The Analysis of Mind*, 1921)

Emotions are necessary for the survival of the individual and the species. Therefore, a simple answer to the title of this book is that all organisms on earth need emotional systems, in their broadest biological definition. Emotional systems enable animals to more effectively explore and interact with their environment, eat, drink, mate, engage in self-protective and defensive behaviors, and communicate. Thus, a robot designed to survive in the world as successfully as its living counterparts undoubtedly would require an equivalent system, one that instills urgency to its actions and decisions—in short, one that motivates and directs. Along with exquisitely designed perceptual, cognitive, and motor networks, evolution has enabled built-in affective mechanisms that in essence constitute a powerful, readily available energizer that ensures efficiency and maximizes survival. The basic premise of this chapter is that emotions are derived from complex, neurochemically coded systems, structured by evolution, that are present in one form or another from single-celled bacteria to primates. Of course, human subjective awareness of a negative emotion such as dejection or humiliation and a crayfish displaying a submissive posture following a struggle with a conspecific are vastly different events; yet one is struck by shared features that characterize neurochemical coding and behavioral mechanisms throughout the evolutionary development of affective systems. Within the rich array of diverse molecules, proteins, neurotransmitters, receptors, and neurohormones in living organisms—some of which have become specialized for emotion—there is a striking phylogenetic conservation of chemical signaling molecules, many of which have played apparently related roles

throughout evolution. In the lobster, serotonin biases dominant behavior, acting as a "gain-setting" device in aggressive conspecific encounters; in humans, serotonin is thought to be a key modulator of mood and control of impulse and aggression. Nuclear transcription factors such as cyclic adenosine monophosphate (cAMP) response element binding protein (CREB), by interacting with genes that encode synaptic modeling molecules, enable plasticity and flexibility of motivated behavior in both fruit flies and mammals. Dopamine receptors likely play a role in reward learning in honeybees, mollusks, mice, and primates. This richness and complexity of behavioral and affective coding presents a great puzzle for behavioral neuroscientists, but the challenge for computational neuroscientists or roboticists modeling emotion is even more daunting. Computational modeling has tackled certain processes, such as sensation, learning, and motor control, with some success; but to incorporate an organism's genome and the combinatorial encoding enabled by its protein products and to relate this to emotional states introduces a different and much more formidable level of complexity. Can knowledge of chemical signaling and transmission inform theories about emotion? Can emotional processes be modeled by machines?

PHYLOGENETIC DEVELOPMENT OF MOTIVATIONAL–EMOTIONAL SYSTEMS

Most chapters or treatises on emotion attempt to define what is meant by such terms as *emotion*, *affect*, and *feelings*. This is a traditional sticking point in the science of emotion as it is notoriously difficult to define what one means by a "feeling"; historically, such endeavors have often led to the philosophy of subjective experience (Russell, 1921) or invited ridicule and the temporary demise of mental science (Watson, 1924). However, in recent decades, a number of testable theories of emotion within the domains of psychology and neuroscience have been developed (Buck, 1999; Damasio, 1996; Ekman & Davidson, 1994; Izard, 1993; MacLean, 1990; Panksepp, 1991; Tomkins, 1982). Buck (1999) nicely summarizes a common thread in these viewpoints: "Rather than stemming from higher-order cognitive appraisal processes, emotions are seen to be based on biologically structured systems that are phylogenetic adaptations, that is, are innate" (p. 302). The conceptual framework of the present chapter is based on ideas emerging from these theorists and on present knowledge of anatomy, neurochemistry, gene expression patterns, molecular evolution, and function of basic brain motivational circuits. It is clear that much of what we conceive of as emotional processing can be accounted for by a growing understanding of motivational circuits and chemical mechanisms within the brain.

It is useful to begin with two important premises: first, that specific and phylogenetically ancient motivational systems exist in the brain and have evolved over the course of millions of years to ensure adaptation and survival and, second, that these systems are engaged by perception of environmental events or stimuli, that is, information, and when so engaged generate specific affective states (positive or negative emotions) that are temporarily powerful drivers and/or sustainers of behavior. Positive emotions generally bring the organism in contact with potentially beneficial resources: food, water, territory, mating, or other social opportunities. Negative emotions protect the organism from danger: mainly ensuring fight-or-flight responses or other appropriate defensive strategies such as submissive behavior or withdrawal, protection of territory or kin, and avoidance of pain. Brain systems monitor the external and internal (bodily) worlds for signals and control the ebb and flow of these emotions (see Fig. 3.1).

Regarding the first premise, the vertebrate brain contains multiple selective systems that are adapted for specific purposes, such as mating, social communication, and ingestion. Corresponding systems exist in the invertebrate brain. These were termed "special purpose" systems by Buck (1999; in contrast to *general purpose systems*, see below) and, within an anatomical framework, *behavioral control columns* by Swanson (2000). A typical example is a system designed to procure water under conditions of dehydration. Sensory information indicating a need for water (dry mouth, stimulation of volume receptors, osmoreceptors) is conveyed via specifically designed anatomical and neurochemical routes (e.g., neural information converging on the periventricular nucleus of the hypothalamus and the neurohormone angiotensin II detected in the subfornical organ). Hypothalamic output pathways connect to the motor system, and the motivated, thirsty animal seeks and procures water. Depending on how thirsty the animal is, the behavior is more or less vigorous and sustained. Other complex neurochemically, anatomically, and hormonally coded systems, discussed in detail below, exist to optimize survival of the individual and the species, ranging from opioids signaling distress calls in rat pups separated from their mother to sex steroids directing sexual differentiation and reproductive behavior. Thus, hunger, thirst, sex, aggression, the need for air and water, and the need for shelter or territory—what Paul MacLean (1969) calls "the primary affects"—are specific drive states that exist to goad the organism to seek the stimuli that will address its basic survival. Among these are the needs to breathe, to have freedom of movement, to rid the body of filth and excrement, and to rest or sleep. Descriptive words for the primary affects associated with many of these basic needs come readily to mind, for example, *hunger*, *thirst*, *suffocation*, *fatigue*, *pain*.

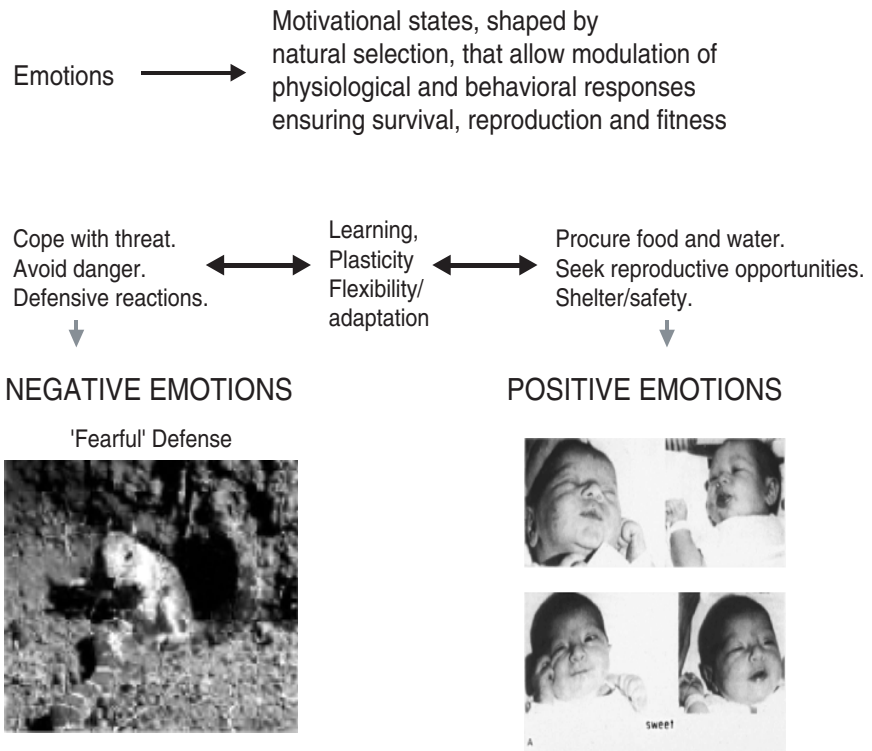


Figure 3.1. Emotions serve as adaptive states that energize and direct survival behaviors, as discussed in the text. Emotions with negative valence (fear, anger, aggression) protect the organism from danger; an example of defensive burying by the ground squirrel faced with threat is shown (photograph by John Cooke, from Coss & Owings, 1989, with permission). Emotions with positive valence are generally associated with appetitive behaviors such as food seeking, sex, and social bonding; shown are facial expressions from neonates given sucrose solution on the tongue (from Steiner, 1973, with permission). Although the potential for species-specific affective behaviors is hard-wired in brain circuits, motivational–emotional systems are capable of flexibility and plasticity due to experience.

However, these are not activated at all times (with the exception of breathing); only in response to particular conditions, states, or needs will motivational circuits be utilized. Buck (1999) develops a very useful notion concerning the concepts of motivation and emotion. Motivation, he postulates, is "*potential* for behavior that is built into a system of behavioral control" [my italics]. It exists whether activated or not; in contrast, emotion is

the *readout* of that system when activated, that is, the *manifestation* of the potential. For example, all organisms have instinctive, built-in mechanisms for defensive behavior in the face of threat or danger; when threat is present, the systems are activated and species-specific defensive behavior ensues.

The latter point leads to the second premise, that these mechanisms are activated by specific environmental (internal or external) stimuli or sensory conditions and amplified and energized by *affect* or *emotion*. Indeed, the origin of the word *emotion*, derived from the Latin *movere* and *e*, meaning "to move out," suggests action; and early uses of the term referred to a moving, stirring, or agitation in the physical sense: "Thunder caused so great an Emotion in the air" (1708 quote from the *Oxford English Dictionary*). Neural and chemical systems exist for aggression and self-defense, but these are manifested, or "moved out," only under appropriate conditions. Indeed, Young (1943), one of the first 20th century students of emotion and motivation, proposed that the most important aspects were energy, regulation, and direction. Tomkins (1982) conceptualized affects as more general mechanisms than drives and hypothesized that a separate affect mechanism exists to amplify or "assist" other mechanisms of behavior. For example, certain aspects of the physical emotional responses associated with both fear and sexual arousal—increased heart rate, blood pressure, respiration, skin conductance—are not specific but rather generalized mechanisms lending urgency to the drive system.

> The affect system is, therefore, the primary motivational system because without its amplification, nothing else matters, and with its amplification, anything else *can* matter. It combines urgency and generality. It lends power to memory, to perception, to thought and to action less than to drives. (Tomkins, 1982, p. 355)

If a mother is walking with a child by her side in a parking lot, she may say, in a normal voice, "Watch out for the cars." However, if she sees a car about to hit her child, she will scream "Watch out" with a level of physical intensity that will clearly be processed by the child in a different way. In both cases, the basic verbal message communicated is similar, but in the latter case, tremendous urgency amplifies the message and changes the context.

In his extensive psychobiological theory of emotion, Buck (1999) expands on Tompkins' theories but argues that it is not necessary to postulate a separate mechanism for affects and that basic drives indicative of bodily needs have their own powerful motivational force associated with them. Buck's thesis is very pertinent to the ideas presented in this chapter, that more general affect systems evolved from more specific motivational mechanisms and can be engaged by higher-level social, cognitive, and (in the case of humans) moral systems. He suggests that motivation and emotion are two sides of the same coin, that, as noted above, emotion is simply the readout

of motivation or the "manifestation of motivational potential," which can be expressed in different ways, such as autonomic activity, social or communicative behavior, and subjective experience. Specific neurochemical systems have evolved to enable these readouts and to render the organism's behavior and subjective state exquisitely sensitive to changes. When injected into the brain, minute amounts of the neuropeptide angiotensin, a major hormone involved in regulation of thirst and sodium appetite, induces immediate and vigorous drinking in a nonthirsty rat (Schulkin, 1999). Gonadal hormones such as estrogen and progesterone, acting on brain sites preserved through evolution, trigger the potential for female sexual behavioral response (Pfaff, 1980). A monkey has the ability to discriminate, via its choice behaviors, drugs that specifically activate the dopamine system from drugs that activate noradrenergic systems (Tidey & Bergman, 1998). Further, the latter case is an example of how animals can choose to artificially amplify emotions through psychoactive drugs, a point we will return to at the end of the chapter. Organisms have the ability to sense ongoing interoceptive changes associated with these different readouts. Thus, the notion that neurochemically and genetically specified neural systems mediate particular species-specific behaviors and behavioral states is a powerful model for explaining the evolutionary development of emotional systems in the brain.

THE NATURAL HISTORY OF MOTIVATIONAL–EMOTIONAL SYSTEMS

The simplest forms of motivational–emotional systems are termed *taxes* (plural of *taxis*) and *tropisms*, or simple movements in response to a stimulus. Motivational–emotional systems evolved from the movements of the earliest organisms on earth at the beginning of life approximately 4 billion years ago. Well before the advent of multicelled organisms and insects, bacteria displayed what is known as *chemotaxis* (movement toward a beneficial stimulus and away from a noxious stimulus). The work of Julius Adler (1966, 1969) with *Escherichia coli* has elegantly traced the genetic and molecular origins of this behavior, and with a little imagination, one can observe the ancient roots of motivation in the behavior of these organisms. These and other motile bacteria will swim toward organic and inorganic attractants such as oxygen, glucose, hydrophilic amino acids, and salt (in the right concentration) and swim away from repellants such as ethanol, certain fatty acids, and hydrophobic amino acids (Adler, Hazelbauer, & Dahl, 1973; Qi & Adler, 1989; Tso & Adler, 1974). Like primordial nervous systems, these cells possess sensory reception, an integrating and transmitting mechanism, and an excitation/effector pathway (see Fig. 3.2).

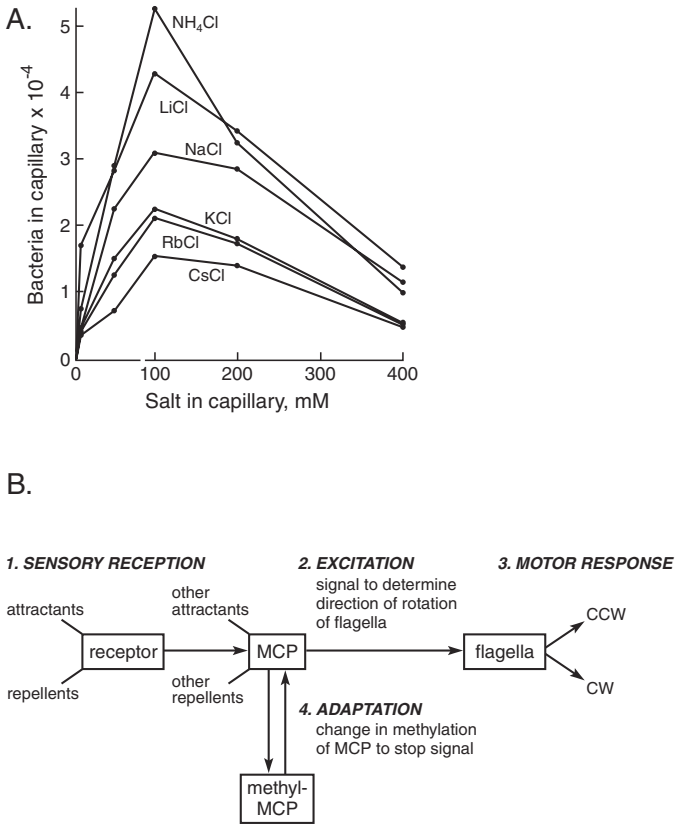


Figure 3.2. (*A*) Chemotaxis toward monovalent cation salts in *Escherichia coli* bacteria. Bacteria are attracted into capillaries, each containing a salt. Salts are attractants only in certain concentrations, usually near 100 mM. (From Qi & Adler, 1989, with permission.) (*B*) Schematic diagram of the mechanism of bacterial chemotaxis. (From Adler, 1990.) cw = clockwise; ccw = counter clockwise; mcp = methyl-accepting protein.

Certain genes code for these various steps, and mutants in different genes produce different deficits in chemotaxic behavior. The whole process is regulated by chemoreceptors, signal-transducing proteins and calcium, and activation of a motor structure, the flagella, via transient methylation of a membrane-bound protein (methyl-accepting chemotaxis protein). The parallels to complex approach–avoidance behaviors and their underlying bases in invertebrates and vertebrates are compelling. In his 1966 *Science* article, Adler noted the relevance of this phenomenon to modern interpretations of the neuroscience of motivation.

> Modern studies of biology have revealed universality among living things. For example, all organisms have much in common when it comes to their metabolism and genetics. Is it not possible that all organisms also share common mechanisms for responding to stimuli by movement? Just as the higher organisms' machinery for metabolism and genetics appears to have evolved from processes already present in the lowest forms, so it is possible that the nervous systems and behavior of higher organisms evolved from chemical reactions that can be found even in the most primitive living things. From this point of view one may hope that a knowledge of chemotaxis in bacteria might contribute to our understanding of neurobiology and psychology. (Adler, 1966)

Many organisms show instinctive behaviors that are highly adaptive and part of repertoires of behaviors that may differ in form in different species but have the common purpose of optimizing survival. Reflexes are the simplest form in that central integration is not needed to accomplish the movement; for example, rapid withdrawal from a painful stimulus is processed by local spinal circuits in the vertebrate nervous system. Reflexes, taxes, and tropisms are all adaptive innate response mechanisms but are devoid of the organizing and energizing attributes of more complex instinctive behaviors, which are linked to core power-generating mechanisms such as heart rate and respiration. Cofer and Appley (1964) state: "starting with this motivating core (the endogenous energy of the instinct proper) a complex and modifiable program of appetitive behavior may be developed." The ethological concept of instinctive behavior, based on detailed observations of animals in their natural environments, has contributed to the notions of fixed action patterns and innate releasing mechanisms. Certain relevant stimuli, termed *releasers*, will elicit particular patterns of behaviors with no prior experience. The face of a parent eliciting a smile in a young infant, the retrieval of a wayward nest egg by the herring gull with its beak, fighting in the male stickleback fish upon intrusion of another male, and the smell of a male rat inducing characteristic enticing behaviors in a female rat in estrus (hopping, darting, ear wiggling) are examples of complex, specific behaviors elicited under particular circumstances. My cat has never seen crustaceans but, upon noticing my son's new pet crayfish, immediately engaged in characteristic feline predatory stalking behavior. Another good example of fixed action patterns is taste reactivity. Many species, including human neonates, show stereotypical facial and oral behaviors when a sweet (positive hedonic) or bitter/sour (negative hedonic) stimulus is applied to the tongue (Steiner, Glaser, Hawilo, & Berridge, 2001; and see Fig. 3.1).

Although instinctual behaviors in animals may not reflect the complexity of human emotions, the origin of the word *instinct*, from the Latin *instiguere* meaning "to incite, to impel," reminds us of the Latin origins of the word *emotion* ("to move out") and suggests a conceptual link between instinct and emotion. An early observer of behavior in animals, McDougall, postulated this close relationship between instinct and emotion. He conceptualized each instinct as

> an inherited or innate psychophysical disposition to perceive, and to pay attention to, objects of a certain class, to experience emotional excitement of a particular class, and to experience an emotional excitement of a particular quality upon perceiving such an object and to act in regard to it in a particular manner, or at least, to experience an impulse to such action. (McDougall, 1908)

Impressive examples of the primitive roots of complex motivated behavior are found in the wonderful observations of lizard behavior by Paul MacLean (1990), who worked in the Laboratory of Brain Evolution and Behavior at the National Institute of Mental Health throughout the middle part of the last century. Based on the writings and extensive observations of ethologists, as well as his own work, MacLean brings to our attention the daily behavioral patterns of the rainbow lizard, a six-inch lizard from West Africa, and the giant komodo dragon, an Indonesian lizard that grows up to 10 feet in length. The daily routines, subroutines, and use of signature displays for social communications are described in detail. In the morning, the typical male rainbow lizard emerges from his safe, protected niche, warms himself, attends to his toilet, and then goes off to forage and feed on insects. If he has established territorial rights, he will display brilliant red and blue colors rather than drab brown. Depending on what and who he encounters, he wards off male intruders into his space by very particular signals (nodding and pushups) and, if left in peace by other males, engages in courtship and possibly copulation with a willing female, exemplified by neck biting and wrapping his leg around her to facilitate mating. At the end of the day, the lizard retires and the next day the routine repeats itself. One sees the fixed, routine patterning as well as the modifiability and flexibility in the expression of these behaviors. For example, in the male blue spiny lizard defending its territory, there are degrees of aggressive display depending on the nature of the encounter. If the intruder merely approaches, there is a "warning, take-notice" display. If the intruder does not heed this, there is a "challenge" display, in which the lizard expands various aspects of his body to make it larger and exposes the blue coloration on his belly. If the intruder still fails to retreat, the tenant rushes for him, tail-lashes, and bites the tail of the offending conspecific (sometimes tails are lost). One way or another, the encounter ends with one member engaging in a sub-

missive bow and retreating. Thus, the particular behaviors aimed at maximal adaptation are fixed, yet their expression is modifiable and sensitive to ongoing stimulus conditions and outcomes. Here, we see the reptilian roots of motivational–emotional systems and their functions in the domains of positive affect (foraging, mating, sunning) and negative affect (aggressive defense, submission, pain).

The notion of drives deserves mention before we explore the brain systems that mediate affect. The concept of instinct is derived from two main notions: that of essentially fixed, innate behavioral programs and that of drive, or satisfaction of bodily needs. However, the term *drive*, like *emotion*, has a somewhat checkered past in the history of psychology and motivation, particularly in traditional learning theory. Drive theory essentially postulated that learning was based on satisfaction of needs (Hull, 1943); however, modern interpretations of learning often discount drive as an explanatory concept, based on many examples of learning and flexible behavior in the absence of satisfaction of any obvious need state (e.g., Berridge, 2001; Bindra, 1978; Bolles, 1972; Dickinson & Balleine, 1994). It is certainly true that drive as conceptualized in a physiological sense (a biological need) cannot account for the diversity of motivated and learned behaviors. However, if we broaden our definition to include motivated, adaptive behaviors that are beneficial to the organism in its environment, it is still a useful concept. For example, most people would not argue that many mammalian species have an innate "drive" to move about and explore, to seek social stimulation, or to learn about spatial surroundings—behaviors that are not obviously mediated by any deficit state such as hunger or thirst but that clearly maximize fitness and availability of resources.

BRAIN CIRCUITS AND THE REGULATION OF MOTIVATED BEHAVIOR

The foregoing account suggests that there are specific brain networks that subserve motivations and emotions. In recent decades, knowledge concerning these networks has advanced at a rapid pace in terms of the detailed understanding of their organization, connectivity, neurochemical and neurohumoral integration, and molecular biology. The purpose of this section is to provide a condensed overview of the key elements and basic organization of these networks. A number of excellent in-depth reviews of anatomy related to motivated behavior exist, to which the reader is referred for more detailed information as well as theoretical implications of brain neuroarchitecture (Risold, Thompson, & Swanson, 1997; Swanson, 2000; Petrovich, Canteras, & Swanson, 2001; Saper, 2000, 2002).

Motivated behavior requires the processing of external and internal sensory information and the coordination and execution of autonomic, endocrine, and somatomotor outputs. The notion of the *limbic system*, a set of related neural structures beneath the neocortical mantle, has profoundly influenced the study of the neural basis of emotion. It is indeed the limbic system concept, with its main focus on the hypothalamus and its connections, that has provided the intellectual framework for thinking about the brain and emotion. Although the conceptual basis of the limbic system has been questioned on a number of grounds and has certainly undergone considerable revisions in recent years (LeDoux, 2000), it is nevertheless a highly useful heuristic, or at least a historical starting point, for grasping the complex organization of pathways underlying the behaviors discussed above. The classic paper by Papez in 1937 ("A proposed mechanism of emotion") is widely acknowledged as seminal to modern affective neuroscience. Papez built his thesis on a number of anatomical and behavioral observations and proposed that a set of anatomically connected structures, including the hypothalamus, anterior thalamic nuclei, cingulate gyrus, hippocampus, and their interconnections, subserved emotional expression and viscero-endocrine responses. MacLean (1949, 1958) developed this notion further and took a distinctly evolutionary point of view, synthesizing aspects of comparative anatomy, paleontology, and ethology. He first postulated the term *limbic system* and drew attention not only to the circuits delineated by Papez but also to their relation to the hypothalamus and proposed that these pathways were phylogenetically quite old compared with the neocortex. MacLean (1990) is perhaps best known for his espousal of the notion of the *triune brain*, which is particularly interesting with regard to the viewpoint of the present chapter. He proposed that the mammalian brain was essentially composed of three formations, which together represent different levels of development in evolution: the protoreptilian brain (represented in lizards and other reptiles and composed of the diencephalic/brain-stem core as well as the basal ganglia), the paleomammalian brain (represented in earlier mammals and composed of limbic structures such as the hippocampus, amygdala, and related structures like the septum), and the neomammalian brain (reaching its most extensive development in later mammals and primates and composed of the neocortex). The general idea is that many basic behaviors necessary for survival—feeding, reproduction, social-communicative behaviors—are hard-wired in striatal–hypothalamic–brain-stem circuits. As natural selection proceeded and animals adapted across millions of years to more and different environments, further behavioral flexibility (embodied in more complex forms of learning, cognition, and ultimately language) was enabled, in a hierarchical fashion, through the progressive expansion of the limbic and neocortical mantle (see Fig. 3.3). MacLean's work has formed a useful structure–function framework for more recent thinking about the evolution of emotional

systems. For the purposes of the present discussion, we will focus on the prototypical mammalian brain.

Modern neuroanatomical tracing methods combined with more advanced cellular and chemical marking techniques have allowed investigators to accrue an enormous amount of information about how the brain is organized. However, as noted by Swanson (2000), such an array of detail provides little insight without a synthetic perspective based on unifying and simplifying principles. In a recent series of extensively detailed papers by Swanson and colleagues, a model of brain architecture and function has been proposed that is based on converging lines of evidence from neurodevelopment, gene expression patterns, circuit connectivity, and function and provides striking insight into the basic organizational patterns that have

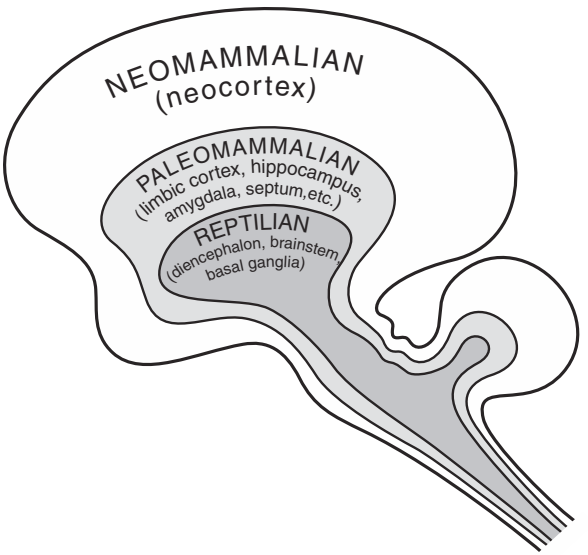


Figure 3.3. The triune brain, as conceptualized by Paul MacLean. MacLean (1990) proposed that the mammalian brain is composed of three main anatomical formations, which represent different levels of development in evolution: the protoreptilian brain (represented in lizards and other reptiles and composed of the diencephalic/brain-stem core as well as the basal ganglia), the paleomammalian brain (represented in earlier mammals and composed of limbic structures), and the neomammalian brain (reaching its most extensive development in later mammals and primates and composed of the neocortex). Behaviors necessary for survival—feeding, reproduction, social–communicative behaviors—are hard-wired in protoreptilian circuits. As natural selection proceeded, further behavioral flexibility was enabled, in a hierarchical fashion, through the progressive expansion of the limbic and neocortical mantle.

emerged through decades of study (Petrovich, Canteras, & Swanson, 2001; Risold, Thompson, & Swanson, 1997; Swanson, 2000). One important feature of this model and very relevant to the current chapter is the notion of behavior control columns. Swanson (2000) proposes that very specific and highly interconnected sets of nuclei in the hypothalamus are devoted to the elaboration and control of specific behaviors necessary for survival: spontaneous locomotor behavior, exploration, and ingestive, defensive, and reproductive behaviors. Animals with chronic transections above the hypothalamus can more or less eat, drink, reproduce, and show defensive behaviors, whereas if the brain is transected below the hypothalamus, the animal displays only fragments of these behaviors, enabled by motor pattern generators in the brain stem. Stimulation and lesion studies during the first half of the 20th century indicated that the motor instructions for species-specific motivated behaviors were instantiated within the hypothalamic circuitry and its brain-stem motor targets. Indeed, such investigations were the hallmark of early physiological psychology. Hess's (1957) extensive treatise on hypothalamic stimulation in the cat provides numerous compelling examples. Aggressive, exploratory, ingestive, and oral responses as well as sleep and many autonomic responses (defecation, blood pressure, respiratory changes, and pupillary dilation) were observed upon electrical stimulation of various hypothalamic sites. The affective component associated with the displays is also vividly described by Hess:

> Perhaps the most striking example is one type of behavior of the cat, in which it looks like it were being threatened by a dog. The animal s*pits*, *snorts*, *or growls* at the same time; *the hair stands on end* and *its tail becomes bushy*; *its pupils widen . . . ears lie back* (to frighten the nonexistent enemy) . . . when the stimulation is maintained or intensified, the cat makes an "actual" attack. The cat turns towards a person standing in its vicinity and leaps on him or strikes a well-aimed blow at him with its paw. This can only mean that the somatic movement is accompanied by a corresponding *psychic attitude*. (Hess, 1957, p. 23, original italics)

Many instances of evoked motivated behavior by direct electrical chemical stimulation—eating, drinking, grooming, attack, sleep, maternal behavior, hoarding, copulation—have been described in the literature. Such examples of remarkably specific evoked responses suggest that the "potential" for behaviors that are often associated with emotion (using Buck's term) is hardwired within the hypothalamic and brain-stem circuitry.

The major divisions of the behavior control column constitute a rostral segment containing nuclei involved in ingestive and social behaviors (reproductive and defensive) and a more caudal segment involved in general for-

aging/exploratory behaviors. Within the rostral column reside nuclei mediating sexual dimorphic behaviors: medial preoptic nucleus, ventrolateral ventromedial nucleus, and ventral premammillary nucleus (these areas contain high levels of estrogen receptor mRNA). Other nuclei, the anterior nucleus, dorsomedial part of the ventromedial nucleus, dorsal premammillary nucleus, mediate defensive responses including defense of territory (and have abundant levels of androgen receptor mRNA). The main hypothalamic controllers for food and water intake are found in the periventricular zone and include the ventromedial and dorsomedial nuclei, the descending part of the paraventricular nucleus, the subfornical organ, and the arcuate nucleus. The more caudal segment of the column includes the mammillary body, the ventral tegmental area, and the reticular part of the substantia nigra. This area, whose efferents ultimately reach parts of voluntary motor circuits via the thalamus and superior colliculus, is proposed to mediate forward locomotion and may also play a role in eye, head, and upper body orientation to salient environmental stimuli. The lateral hypothalamus is not specifically included in Swanson's behavioral control column scheme but probably plays a critical role in arousal, control of behavioral state, and reward-seeking behavior. The lateral hypothalamus has long posed an enigma to investigators, not the least because rats will press a lever thousands of times per hour to deliver electrical stimulation to this region (Olds, 1958).

Central to this basic model of motivated behavior is an appreciation of the main inputs to these hypothalamic systems, the features of its organization with regard to other major brain regions, and its targets. As elaborated above, motivational–emotional systems are triggered into action by specific signals—energy deficits, osmotic imbalances, olfactory cues, threatening stimuli—that impinge on the system and initiate (as well as terminate) activity in specific brain pathways, thereby effecting responses. In higher mammals, these signals reach the behavioral control column in multiple ways. Multiple sensory inputs from the external world reach the hypothalamus both directly and indirectly (Risold, Thompson, & Swanson, 1997). For example, it receives direct input from the retina; olfactory and pheromonal information is conveyed via a massive projection from the medial amygdala and bed nucleus of the stria terminalis. Cues relating to territory and identification of individuals as prey or predators arrive to the rostral control column; those important for visceral (including pain, temperature, and heart rate) and gustatory processing reach the hypothalamus principally through the brain-stem nucleus of the solitary tract and parabrachial nuclei, which bring in information about taste and visceral sensations. This information influences the periventricular zone involved in ingestion as well as the lateral hypothalamus. Metabolic and humoral information (circulating levels of glucose, salt, fatty acids, hormones such as insulin and leptin, angiotensin,

and gonadal and adrenal steroids) influence the hypothalamus via circumventricular organs such as the arcuate nucleus, which has dense receptors for circulating chemical signals. The spinohypothalamic tract carries somatosensory information (mostly to the lateral hypothalamus). Thus, many neural and chemical sensory inputs to the behavioral control columns have been identified, and it is clear that the architecture is elegantly designed for complex coordination of adaptive motivated behavior.

Returning to Swanson's model, a second route for critically important inputs to the behavioral control column is via the cerebral cortex, including massive direct and indirect afferents from such areas as the hippocampus, amygdala, prefrontal cortex, striatum, and pallidum. Via these inputs, the "reptilian core" has access to the highly complex computational, cognitive, and associative abilities of the cerebral cortex. For example, hippocampal inputs from the subiculum innervate the caudal aspect of the column involved in foraging and provide key spatial information to control navigational strategies; place cells are found in regions of the mammillary bodies as well as the hippocampus, anterior thalamus, and striatum (Blair, Cho, & Sharp, 1998; Ragozzino, Leutgeb, & Mizumori, 2001). The amygdala's role in reward valuation and learning, particularly in its lateral and basolateral aspects (which are intimately connected with the frontotemporal association cortex), can influence and perhaps bias lateral hypothalamic output. Indeed, recent studies have supported this notion; disconnection of the amygdalo–lateral hypothalamic pathway does not abolish food intake but alters subtle assessment of the comparative value of the food based on learning or sensory cues (Petrovich, Setlow, Holland, & Gallagher, 2002); in some of our recent work, inactivation of the amygdala prevents expression of ingestive behavior mediated by striatal–hypothalamic circuitry (Will, Franzblau, & Kelley, 2004). The potential for cellular plasticity in cortical and striatal regions is greatly expanded compared to brain-stem and hypothalamic systems. Indeed, gene expression patterns can reveal this expansion in evolutionary development. An example from our material (Fig. 3.4) shows that the cortex and striatum are rich in the protein product of the gene *zif268*, which plays an important role in glutamate- and dopamine-mediated plasticity (Keefe & Gerfen, 1996; Wang & McGinty, 1996). Levels of this gene product are much lower in the brain stem and diencephalon. Thus, the phylogenetically most recently developed and expanded brain region, the "neomammalian" cerebral cortex, is intricately wired to communicate with and influence the ancestral behavioral control columns and capable of complex cellular plasticity based on experience.

As the origin of the term would suggest, motivation must ultimately result in behavioral actions. The Canadian physiological psychologist Gordon Mogenson and colleagues (1980) drew attention to this matter in their land-

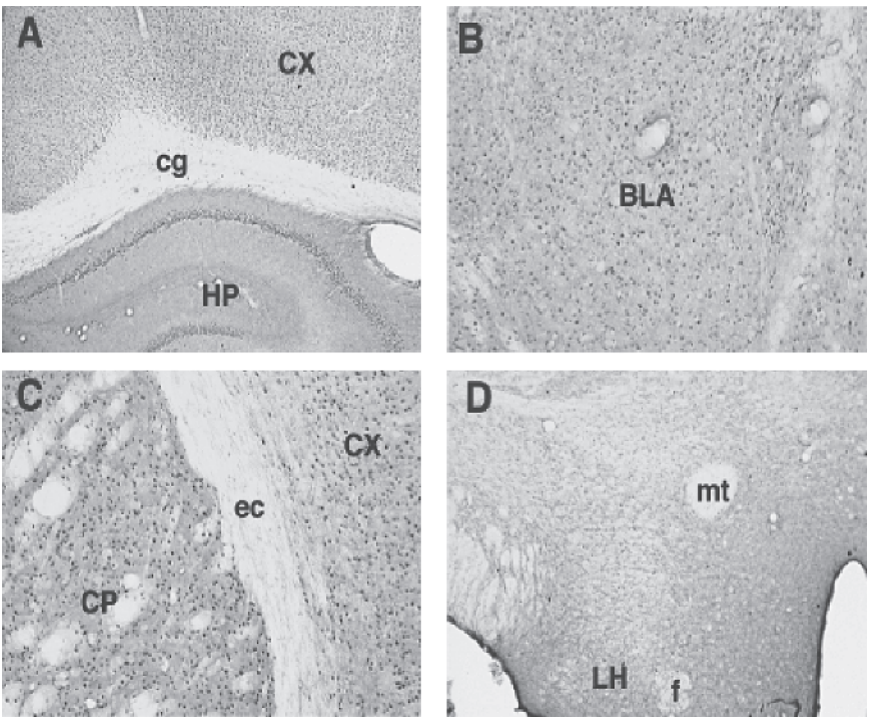


Figure 3.4. Immunostained sections of rat brain show expression of the immediate early gene *zif268*, which has been implicated in cellular plasticity. The *zif268* gene is regulated by dopamine and glutamate and may mediate long-term alterations underlying learning and memory. Each black dot represents nuclear staining in a cell. Note strong expression in cortical, hippocampal, striatal, and amygdalar areas (*A*, *B*, *C*) and much weaker expression in diencephalic areas (*D*). This gene and others like it may be preferentially expressed in corticolimbic and striatal circuits, which play a major role in plasticity. BLA, Basolateral amygdala; cg, cingulum; CP, caudate-putamaten; CX, cortex; ec, external capsule; f, fornix; HP, hippocampus; LH, lateral hypothalamus; mt; mammillothalamic tract.

mark paper "From Motivation to Action." Actions occur when the motor outputs of these systems are signaled, whether via autonomic output (heart rate, blood pressure), visceroendocrine output (cortisol, adrenaline, release of sex hormones), or somatomotor output (locomotion, instrumental behavior, facial/oral responses, defensive or mating postures). During coordinated expression of context-dependent motivated behaviors, various combinations of these effector systems are utilized. Indeed, all the behavioral control columns project directly to these motor effector routes. However, in mammals, conscious,

voluntary control of actions is further enabled by superimposition of cortical systems on the basic sensory-reflexive network. Moreover, there is extensive reciprocal communication between the cerebral hemispheres and motor effector networks. An additional major principle for organization of the behavioral control columns is that they project massively back to the cerebral cortex/voluntary control system directly or indirectly via the dorsal thalamus (Risold, Thompson, & Swanson, 1997; Swanson, 2000). For example, nearly the entire hypothalamus projects to the dorsal thalamus, which in turn projects to widespread regions of the neocortex. Moreover, recently characterized neuropeptide-coded systems have revealed that orexin/hypocretin- and melanin concentrating hormone–containing cells within the lateral hypothalamus project directly to widespread regions within the neocortex, amygdala, hippocampus, and ventral striatum and may be very important for behavioral state regulation and arousal (Espana, Baldo, Kelley, & Berridge, 2001; Peyron et al., 1998). This feed-forward hypothalamic projection to the cerebral hemispheres is an extremely important anatomical fact for grasping the notions elaborated above, that intimate access of associative and cognitive cortical areas to basic motivational networks enables the generation of emotions or the manifestation of "motivational potential." Thus, in the primate brain, this substantial reciprocal interaction between phylogenetically old behavioral control columns and the more recently developed cortex subserving higher-order processes such as language and cognition has enabled a two-way street for emotion. Not only can circuits controlling voluntary motor actions, decision making, and executive control influence and modulate our basic drives, but activity within the core motivational networks can impart emotional coloring to conscious processes. A flat map anatomical diagram from the work of Swanson (2000), showing some of the pathways described here, is provided in Figure 3.5.

EVOLUTIONARY DEVELOPMENT OF NEUROTRANSMITTER SYSTEMS

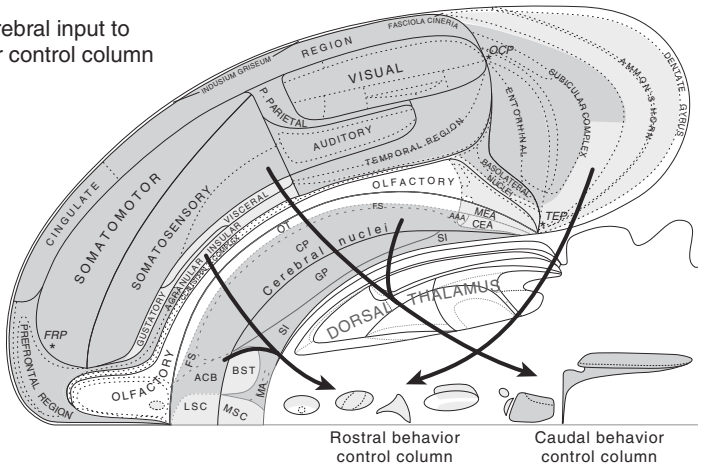
Neurochemical signaling pathways involved in emotional processing in the mammalian brain have evolved over the billions of years since the origins of life. Within the constraints of genetic evolution, nervous systems became more complex and enabled progressively greater possibilities for the animal in its relationship with its environment. Chemical signaling played a critical role in this connectivity and adaptation. Neurotransmitter signaling networks and their corresponding receptor molecules, particularly the biogenic amines, small neuropeptides, and neuropeptide hormones, became specialized for particular behaviors or motivational states (Niall, 1982;

Walker, Brooks, & Holden–Dye, 1996). Neurotransmitters are released from axon terminals, cross the synapse, and bind to postsynaptic receptor sites to effect a cascade of intracellular biochemical events. Uptake sites on presynaptic terminals are proteins that regulate the synaptic level of neurotransmitter by binding released transmitter and transporting it back into the terminal. These molecules play a role in adaptive behaviors to a surprisingly conserved degree across species and phyla. Subjective states in humans which are associated with such feelings as joy, fear, anxiety, and maternal love are derived from the actions of truly primordial chemical systems. Following the origins of bacterial life, eukaryotic cells appeared approximately 2 billion years ago, primitive multicelled organisms appeared around 800 million years ago, and vertebrates are estimated to have diverged from invertebrates around 500–600 million years ago. All extant mammals, birds, and reptiles are derived from stem reptiles that lived approximately 200–300 million years ago. Neurotransmitter development followed this evolutionary path. All neurons, throughout the animal kingdom, contain at least one releasable substance (usually an amine, peptide, amino acid, or acetylcholine) and utilize either ligand-gated ion channels or second messengers such as G proteins, AMP, phospholipase C, and calcium to communicate their signal postsynaptically. Second-messenger systems appeared quite early in evolution, perhaps to add a longer time scale and greater flexibility in neural communication.* For example, the yeast alpha-mating factor (a peptide pheromone) is a member of the G protein–coupled receptor superfamily (Darlison & Richter, 1999), and G protein–coupled receptors are found throughout arthropods, flatworms, and mollusks (Walker, Brooks, & Holden-Dye, 1996).† Calcium, a ubiquitous second messenger, plays this role even in bacteria (Tisa & Adler, 1992). Ligand-gated channels, complex membrane-bound proteins that allow fast chemical transmission via gating of the flow of cations and anions in and out of the cell (such as that involving γ-aminobutyric acid, acetylcholine, and glutamate), are present in all animal species studied thus far. Chemical compounds can act in several ways: strictly as transmitters that convey specific information via their effect on postsynaptic receptors, as modulators of the postsynaptic receptor so as to alter other incoming signals, or as signals acting at sites distal from release sites, thus acting as neurohormones.

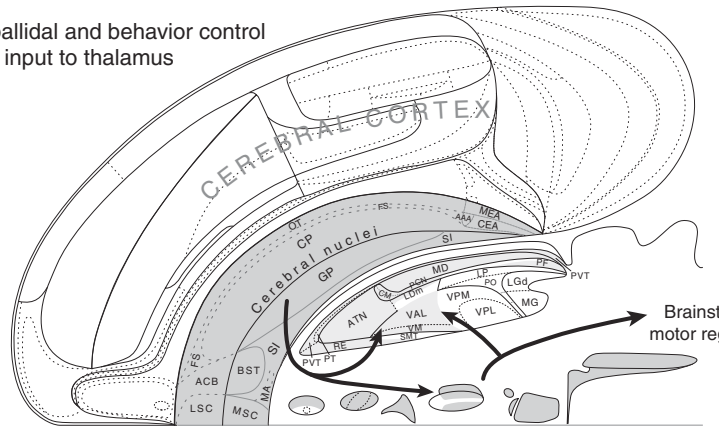
*Ligand-gated ion channels are proteins that allow rapid flux of ions such as sodium or potassium in and out of the neuron, depending on the binding of neurotransmitter to its receptor. Second messengers are molecules that aid in the transduction of the chemical signal to an electrical signal.

†G protein–coupled receptors are receptors for neurotransmitters that utilize specific membrane-bound proteins—G proteins—that activate certain critical intracellular second messenger enzymes, such as cyclic AMP.

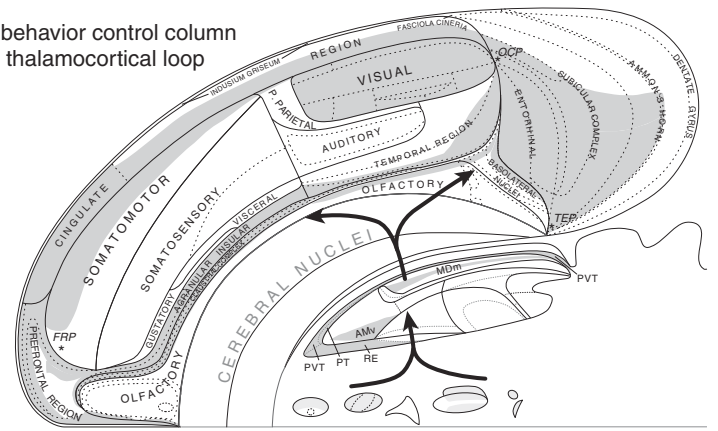
A. Total cerebral input to behavior control column



B. Striatopallidal and behavior control column input to thalamus



C. Rostral behavior control column input to thalamocortical loop



The time scale of these processes can vary from milliseconds to months and even years, in the case of long-term plasticity.

Modern genetic sequencing techniques combined with advances in bioinformatics have allowed novel insights into assessments of gene nucleotide sequence homology throughout evolution and the animal kingdom. Comparison of sequence relationships in genes between different species yields evidence of both diversity and conservation of neurochemical signaling and function. For example, acetylcholine and its corresponding nicotinic and muscarinic receptors occur across species from the platyhelminths (flatworms) and nematodes to vertebrates, functioning as a chemical signal in

Figure 3.5 (facing page). Flat map of general forebrain organization, according to Swanson (2000), showing major pathways subserving emotion and motivation, as discussed in the text. At the bottom of each figure, the "behavior control columns" are depicted; the rostral segment governs ingestive, reproductive, and defensive behaviors, while the more caudal segment directs exploratory and foraging behaviors. (*A*) Nearly the entire cerebral hemispheres project to the behavior control column. Cerebral inputs to the rostral segment are shown in light gray and those to the caudal segment, in darker gray. (*B*) The entire basal ganglia (striatopallidum) gives rise to a branched projection to the dorsal thalamus and behavior control column, which in turn generates a branched projection to both the dorsal thalamic and brain-stem motor regions. The part of the dorsal thalamus innervated by the basal ganglia and behavior control columns is shown in lighter gray. Keep in mind that this part of the thalamus projects massively back to the entire cerebral cortex. (C) The thalamocortical projection, indicated in darker gray, is influenced by the rostral behavior control column (arising mainly from the medial dorsal nucleus). AAA, anterior amygdalar area; ACB, nucleus accumbens; AMv, anteromedial nucleus, ventral part; ATN, anterior thalamic nuclei; BST, bed nuclei stria terminalis; CEA, central nucleus amygdala; CM, central medial nucleus; CP, caudoputamen; FRP, frontal pole; FS, striatal fundus; GP, globus pallidus; LGd, dorsal lateral geniculate nucleus; LP, lateral posterior nucleus; LSC, lateral septal complex; MA, magnocellular (preoptic) nucleus; MDm, mediodorsal nucleus, medial part; MEA, medial nucleus amygdala; MG, medial geniculate nucleus; MSC, medial septal complex; OCP, occipital pole; OT, olfactory tubercle; PCN, paracentral nucleus; PF, parafascicular nucleus; PO, posterior complex thalamus; PT, paratenial nucleus, PVT, paraventricular nucleus thalamus; RE, nucleus reuniens; SMT, submedial nucleus thalamus; SI, substantia innominata; TEP, temporal pole; VAL, ventral anterior–lateral complex; VM, ventral medial nucleus; VPL, ventral posterolateral nucleus; VPM, ventral posteromedial nucleus. (Adapted from Swanson, 2000, with permission.)

sensory neurons, interneurons, and motor neurons (Changeux et al., 1998). Serotonin (5-hydroxytryptamine [5-HT]) is a further example of a substance with an important role in various physiological and behavioral processes. In the mammalian brain, over 15 different receptors for 5-HT have been cloned and sequenced; some interact directly with ion channels and others with G protein–coupled second-messenger systems (Peroutka & Howell, 1994). A high degree of sequence homology exists between many of these and those characterized for lower invertebrates such as *Drosophila* and *Aplysia*, as shown in Figure 3.6.

Dopamine (DA) receptors are also widely studied, and five subtypes have been cloned (Jackson & Westland-Danielsson, 1994; Missale et al., 1998). Interestingly, there appears to be an insect homologue for the mammalian dopamine D_1 receptor, which has been implicated in memory and plasticity; a high degree of transmembrane domain homology exists between the *Drosophila Ddop-1* gene and the mammalian *D1/D5* gene (Blenau & Baumann, 2001). Much is now known about families of neuropeptide genes and their receptors (Hoyle, 1999). For example, the nonapeptide family, which includes vasopressin and oxytocin, peptides critical for neural control of social communication, that is, territorial, reproductive, and parenting behavior, provides a particularly good example of biochemical evolution. This system in mammals has its ancestral roots in invertebrates, with function in reproduction in some cases being conserved. For example, oxytocin has multiple roles in maternal behavior in mammals, including infant attachment (Insel & Young, 2000); a member of this family, conopressin, regulates ejaculation and egg laying in the snail (Van Kesteren et al., 1995); and the related vasotocin regulates birthing behavior and egg laying in sea turtles (Figler et al., 1989). The neuropeptide Y (NPY) superfamily is also widely distributed in evolution. This system is a good example of peptide superfamilies where there is considerable sequence homology for the presynaptic peptide across species but much greater diversity in the evolution of its receptors (Hoyle, 1999). Since peptide receptors are generally much larger than transmitter peptides, it is likely that there was much greater chance for mutations and gene duplication in receptors with receptor function being maintained. In mammals, NPY is involved in hypothalamic feeding mechanisms; in a recent study of *Caenorhabditis elegans*, one single-base mutation in the *npr-1* gene, coding for a receptor structurally related to the mammalian NPY receptor, was enough to dramatically alter the feeding behavior of these worms (de Bono & Bargmann, 1998).

It is important to note that although I have emphasized interesting sequence homologies coding for various chemical signaling molecules across the evolution of species, there are many instances where a peptide or protein has been conserved but evolves to serve multiple and often unrelated

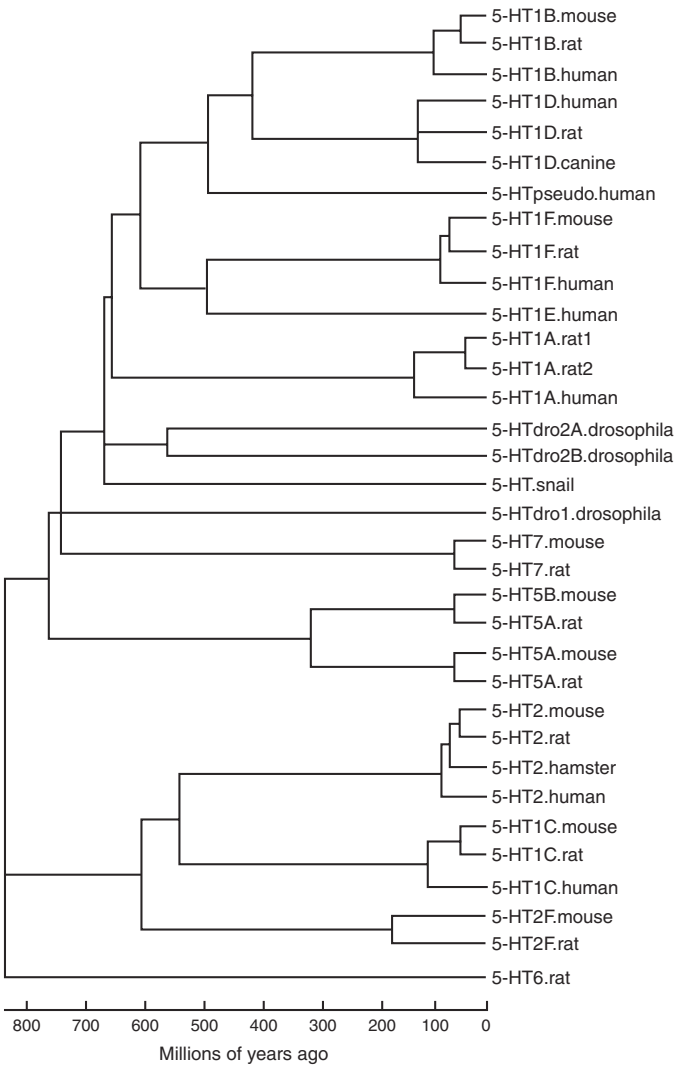


Figure 3.6. Phylogenetic tree of the serotonin (5-HT) receptor, showing strong homology across many species and the ancient nature of neuronal signaling proteins involved in motivated behavior. Evolutionary distance between receptor populations is indicated by the length of each branch of the tree. (From Peroutka & Howell, 1994, with permission.)

functions. Niall (1982) notes that gene duplication is only one means of diversification; another is development of a new or different function for a peptide hormone. For example, he notes that prolactin enables fish to adapt to varying salt concentrations; in mammals, it became involved in the control of lactation. Moreover, many so-called pituitary hormones are made in many brain and gut regions, possibly serving various functions in these different structures. Medawar (1953) noted that "endocrine evolution is not evolution of hormones but an evolution to the uses to which they were put." Thus, although there are many intriguing examples of conservation of function across phyla, it is important to appreciate the diversity of signaling functions as well.

NEUROCHEMISTRY AND PHARMACOLOGY OF EMOTIONS

The above account provides an organizational framework for understanding the hard-wiring of motivational circuits, how they are affected by sensory stimuli, and how they have the ability both to effect behavioral responses via direct motor outputs and to feed forward to influence higher cortical regions and perhaps generate awareness. Communication between the billions of synapses as well as general modulation of these systems is accomplished via chemical signaling; but how and where do these substances act to produce changes in emotion, mood, and behavioral state? Given the space limitation here, I cannot possibly describe in detail the vast array of neurotransmitters and neuromodulators that contribute to the functional role of these systems. Instead, I have chosen several candidate systems that represent compelling examples of chemical signaling systems that mediate motivation and emotion and that have parallel links to related functions across phyla.

Dopamine: Reward and Plasticity

A great amount of attention has been given to the catecholamine DA in a variety of species. In mammals, DA is proposed to play a major role in motor activation, appetitive motivation, reward processing, and cellular plasticity and certainly can be thought of as one candidate molecule that plays a major role in emotion. Like the other catecholamines norepinephrine and epinephrine, DA is synthesized from the amino acid tyrosine and involves several biosynthetic steps employing the enzymes tyrosine hydroxylase and dihydroxyphenylalanine (DOPA) decarboxylase. Receptors for DA exist in two major classes: D_1-like (D_1 and D_5 receptors) and D_2-like (D_2, D_3, D_4).

The molecular pharmacological characteristics of these receptors are well established, as is the presynaptic DA transporter or uptake site (Jackson & Westland-Danielsson, 1994; Vallone, Picetti, & Borrelli, 2000). The DA receptors belong to a large gene family of G protein–coupled, seven-transmembrane domain-spanning receptors that are linked to intracellular second-messenger systems such as cAMP (Missale et al., 1998). In the mammalian brain, DA is contained in specific pathways that have their origins in the substantia nigra and ventral tegmental area of the midbrain and ascend to innervate widespread areas of striatal, limbic, and cortical regions such as the striatum, prefrontal cortex, amygdala, and other forebrain regions. Thus, the DA system targets many of the cortical and striatal regions noted above.

A number of interesting hypotheses have been developed concerning the role of DA within motivational–emotional systems, primarily based on research in rodents and primates. Perhaps the most important notion along these lines is that DA plays a major role in motor activation, reward, and reinforcement. Through studies that quantify DA activity via microdialysis, voltammetry, electrophysiological recordings, pharmacological manipulations, and lesion studies, it has been shown that DA is activated by many natural and drug rewards and that its blockade or removal severely impairs an animal's ability to respond to rewards or reward-related cues (secondary reinforcers) in the environment (Wise & Rompré, 1989; Berridge & Robinson, 1998; Horvitz, 2000; Salamone, Cousins, & Snyder, 1997). For example, DA in the ventral striatum plays a critical role in both male and female sexual behavior (Becker, Rudick, & Jenkins, 2001; Pfaus et al., 1990), and rewarding stimuli such as highly palatable food or reward-associated stimuli strongly activate DA release (Bassareo & Di Chiara, 1999; Wilson, Nomikos, Collu, & Fibiger, 1995). Drugs of abuse have the common property of activating DA, and humans and other animals self-administer drugs that increase brain DA (Di Chiara, 1998; see also below). Anticipatory situations when animals are expecting a reward appear to engage DA neuronal activation; for example, placing an animal in a context where it has previously received food, sex, or drugs can increase DA cell firing or extracellular levels of DA (Blackburn, Phillips, Jakubovic, & Fibiger, 1989; Pfaus & Phillips, 1991; Ito et al., 2000; Schultz, Apicella, & Ljungberg, 1993). In humans, cues associated with drugs such as heroin or cocaine or even with playing a video game can activate DA systems or areas heavily innervated by DA (Childress et al., 1999; Koepp et al., 1998; Sell et al., 1999; Volkow, Fowler, & Wang, 2002). Compelling evidence for DA playing a necessary role in motivation derives from the fact that rats deprived selectively of all forebrain DA will starve to death unless fed artificially; these animals have the capability of moving and eating but appear unable to maintain a critical level of

motivation arousal necessary for ingestive behavior (Marshall, Richardson, & Teitelbaum, 1974; Ungerstedt, 1971).

In addition to mediating the processing of ongoing incentive stimuli in an organism's environment, DA signals appear to be an integral part of learning and plasticity in the many forebrain regions that they influence (Di Chiara, 1998; Cardinal, Parkinson, Hall, & Everitt, 2002; Sutton & Beninger, 1999). Indeed, several major hypotheses concerning the functional role of DA primarily emphasize its role in associative or incentive learning or the ability of the organism to learn about beneficial or potentially beneficial stimuli in its environment and react appropriately. For example, Robinson and Berridge have proposed that DA is important for attributing incentive salience to neural representations of rewards, through a process that enables an environmental stimulus to be attractive, or "wanted," and to elicit voluntary approach behavior (Berridge & Robinson, 1998; Robinson & Berridge, 1993). The work of Schultz (2000), utilizing single-cell recording in the awake, behaving monkey, suggests that DA neurons fire to predicted rewards and track expected and unexpected environmental events, thereby encoding "prediction errors," that is, information about future events of potential salience or value to the animal. Studies show that neurons in prefrontal–striatal networks are sensitive to reward expectations and activated in association with motor responses to specific learned cues or events (Pratt & Mizumori, 2001; Schoenbaum, Chiba, & Gallagher, 1998; Schultz, Tremblay, & Hollerman, 2000). Work in the prefrontal cortex is particularly interesting in this regard. Prefrontal networks are equipped with the ability to hold neural representations in memory and use them to guide adaptive behavior; DA and particularly D_1 receptors are essential for this ability (Williams & Goldman-Rakic, 1995). In rats, D_1 receptor activation in the prefrontal cortex is necessary for active retention of information that guides future behavior in a foraging task and modulates hippocampal inputs to the prefrontal cortex (Seamans, Floresco, & Phillips, 1998). Thus, DA may "prime" and ultimately reinforce motor strategies that result in adaptive, beneficial behavior. Electrophysiological work indicates that DA is able to hold or gate neurons in a primed "up state" and to facilitate the potential for the network to learn new information and initiate plasticity (Lewis & O'Donnell, 2000; Wang & O'Donnell, 2001). Our own work has shown that activation of D_1 receptors in corticostriatal networks is essential for hungry rats' ability to learn an instrumental response for food (Baldwin, Sadeghian, & Kelley, 2002; Smith-Roe & Kelley, 2000).

As mentioned earlier, monoamines are abundant throughout phylogenetic development. Dopamine, DA receptors, and associated proteins such as transporters and synthetic and phosphorylating enzymes have been found in all species thus far examined, including nematodes, mollusks, crustaceans, insects, and vertebrates (Cardinaud et al., 1998; Kapsimali et al., 2000;

Walker, Brooks, & Holden-Dye, 1996). One study showed a 70% homology between cloned *Drosophila* D_1/D_5 receptors and their human counterpart as well as stimulation of cAMP production by DA (Gotzes, Balfanz, & Baumann, 1994). Although the functional role of DA has not been examined in lower species as extensively as in vertebrates, there is certainly some evidence that it may influence cellular function, adaptive behaviors, and possibly plasticity in many animals. In *Drosophila*, the DA system appears to regulate development, feeding, sexual behavior, and possibly learning (Neckameyer, 1996, 1998; Tempel, Livingstone, & Quinn, 1984; Yellman, Tao, He, & Hirsh, 1997). In honeybees, DA receptors have been well characterized and proposed to play a role in motor behavior (Blenau & Baumann, 2001; Kokay & Mercer, 1996). Menzel and colleagues (1999) have used classical conditioning in the honeybee to demonstrate a potential analogue for a DA role in reward learning. Appetitive conditioning to sucrose (olfactory conditioning to the proboscis extension reflex) is impaired with depletion of biogenic amines and restored by DA. Further studies of reward learning have shown that a neuron, termed VUMmx1, shows similar "reward prediction" properties to the mammalian homologue (Menzel, 2001). In *Aplysia*, DA appears to be a transmitter in a central pattern generator important for feeding (Kabotyanski, Baxter, Cushman, & Byrne, 2000; Diaz-Rios, Oyola, & Miller, 2002), and recent investigations show that dopaminergic synapses mediate neuronal changes during operant conditioning of the buccal reflex (Nargeot, Baxter, Patterson, & Byrne, 1999; Brembs et al., 2002). These examples suggest that throughout evolutionary development of species DA has retained a role of reward–motor coupling. Its expanded capacity to modulate and modify the activity of cortical networks involved in cognition, motor planning, and reward expectation is apparent in mammalian species.

Serotonin: Aggression and Depression

A further example of monoamine modulation of motivated behaviors and affective processing is the serotonergic system. Serotonin (5-HT), an indoleamine synthesized from the amino acid tryptophan, has been widely implicated in many behavioral functions, including behavioral state regulation and arousal, motor pattern generation, sleep, learning and plasticity, food intake, mood, and social behavior. In terms of anatomy in the mammalian brain, serotonergic systems are widespread; their cell bodies reside in midbrain and pontine regions, and there are extensive descending and ascending projections. Descending projections reach brain-stem and spinal motor and sensory regions, while the ascending inputs project to widespread regions in the cortex, limbic system, basal ganglia, and hypothalamus—indeed, the

serotonergic system is proposed to be "the most expansive neurochemical network in the vertebrate CNS" (Jacobs & Azmitia, 1992). It is a system that is highly conserved across phyla; serotonin is an important neurotransmitter in many invertebrates, and over 15 subtypes of 5-HT receptors have arisen through molecular evolution, most of which interact with G proteins (Peroutka & Howell, 1994; Saudou & Hen, 1994). This extensive development of receptor subtypes suggests a great diversity of signaling within serotonin systems across phyla. Serotonin is a particularly interesting and appropriate chemical signal to examine in the context of the evolution of motivated behavior and emotion. Over the past 25 years or so, its pharmacology, physiology, and molecular biology have been extensively studied in crustaceans, insects, mollusks, worms, and mammals. Moreover, dysfunction of serotonin in humans has been implicated in many psychiatric disorders, such as depression, anxiety, obsessive–compulsive behavior, and alcoholism (Nemeroff, 1998). Serotonin-selective reuptake inhibitors (SSRIs) are among the most commonly prescribed psychiatric drugs. Thus, understanding the functions of 5-HT through analysis of its role in behavior and brain function may provide important insights into the neurochemical basis of human emotion and its disorders.

While serotonin clearly has roles of a very diverse nature, there is an interesting common thread that weaves through much of the research on this substance. Work from a variety of approaches leads to the general consensus that 5-HT plays a critical role in the modulation of aggression and agonistic social interactions in many animals and possibly regulation of aggressive behavior and mood in nonhuman primates and humans (Insel & Winslow, 1998). Experiments on crustaceans such as the lobster and crayfish clearly implicate serotonin, as well as octopamine (a phenol analogue of norepinephrine), in the control of behaviors concerned with the maintenance of social hierarchies. The fighting behavior of these phylogenetically ancient animals, highly successful predators and scavengers, has been extensively studied in artificial aquatic environments.

> In a typical scenario, intensity of fighting increases in a step-wise fashion beginning with threat displays upon first contact, followed by phases of ritualized aggression, restrained use of claws, and in rare instances ending in periods of unbridled combat. The presence of such a structured behavioral system, combined with an opportunity to bring the analysis to the level of individual neurons, thus offers unique opportunities for exploring fundamental issues of interactions between aggression, dominance, and amine neurochemistry. (Huber et al., 2001, p. 272)

In lobsters and crayfish, serotonin and octopamine direct or bias the readout of stereotypical motor programs that cause dominant or subordinate postures, respectively. Infusions of serotonin into the hemolymph induce aggressive, dominant postures, even in formerly subordinate animals. In groups of individuals, social status becomes established as a hierarchy develops. Thus, social behavior can become conditioned and show plasticity; indeed, the social status can determine the extent of the response to serotonin (Yeh, Fricke, & Edwards, 1996), and infusion of serotonin into a subordinate animal actually increases its willingness to fight (Huber et al., 1997).

In mammals, there is extensive evidence for involvement of serotonin in modulation of aggression. As for other amines, a variety of methods have been used to manipulate the serotonin system, including pharmacology, lesions, microdialysis, and genetic knockout strategies. Also, it is very important to note that the number and variety of serotonin receptors suggest that this modulation is very complex (appraisal of the literature indicates that both too little and too much tone in serotonergic neurons can disrupt aggression); moreover, treatments that globally affect serotonin affect multiple receptor systems and may not reveal any clear function. Early studies indicated that in mice and rats depletion of serotonin with drugs induced a temporary increase in aggression; for example, rats that normally ignore mice would engage in mouse-killing behavior (Vergnes, Depaulis, & Boehrer, 1986). More recent work with receptor-specific drugs indicated that treatment of rats with 5-HT1B or 5-HT1A agonists tended to reduce aggression in a rat resident–intruder model of aggressive behavior (Miczek, Mos, & Olivier, 1989). Moreover, 5-HT1B knockout mice show increased levels of aggression (Saudou et al., 1994), and the 5-HT1A receptor is strongly implicated in expression of anxiety-like behavior in animal tests (Gross et al., 2002).

In nonhuman primates, the link between aggression and serotonin is quite compelling, although mainly based on an indirect measure of serotonin, cerebral spinal fluid (CSF) measures of 5-hydroxyindoleacetic acid (5-HIAA), the metabolite serotonin. Moreover, this work reveals an important relationship between 5-HT, aggression, and social relationships among conspecifics, much as in crustaceans. For example, Higley et al. (1996) and Mehlman et al. (1994) examined the relationship between CSF 5-HIAA and behavior in free-ranging monkeys in naturalistic environments. These studies show a clear correlation between lowered CSF 5-HIAA levels and increased levels of impulsive aggression. For example, among rhesus monkeys living in social colonies, animals with the lowest quartile of CSF 5-HIAA had high levels of unprovoked, escalated aggression and a higher risk of injury or death. These animals would initiate aggression at inappropriate targets, such as high-ranking males, and demonstrate impaired impulse control in other behaviors, such

as tree jumping (Doudet et al., 1995; Higley et al., 1996; Mehlman et al., 1994). In another study of monkey social groups, treatment with drugs that enhanced or reduced brain serotonin levels clearly revealed behavioral patterns where animals with high serotonin showed lower levels of aggression and enhanced social skills (Raleigh & McGuire, 1991).

In humans, there is convincing evidence for serotonin dysfunction in a variety of disorders and disordered behavior. In psychiatric populations, there is a well-established link between abnormally low central 5-HT levels and increased aggressive or antisocial behavior, alcoholism, and impaired impulsive control (Mann et al., 1996; Virkkunen and Linnoila, 1992). Patients with reduced serotonin function have been shown to have higher rates of major depression and suicide attempts or completed suicide (Coccaro et al., 1989; Mann et al., 1996). Thus, research from studies on humans and other animals clearly implicates an important and complex role for central serotonin and its receptors in the control of behavioral state. In nonhuman animals, this has been demonstrated in the realm of control of aggression and social status or interactions; in humans, this involvement is expanded to regulation of mood and emotions, particularly control of negative mood or affect. Since serontin-containing neurons innervate nearly all regions of the neuraxis in higher mammals, this role is also a particularly good example of the anatomical and functional evolution of a neurochemical system: in crustaceans, serotonin plays a specific role in social status and aggression; in primates, with the system's expansive development and innervation of the cerebral cortex, serotonin has come to play a much broader role in cognitive and emotional regulation.

Opioid Peptides: Pain and Pleasure

The opioid peptides and their receptors are a further example of neurochemical modulation of affect. Since the discovery of endogenous opioid peptides and their receptors nearly three decades ago (Lord, Waterfield, Hughes, & Kosterlitz, 1977; Pert & Snyder, 1973), there has been enormous interest in understanding the functional role of these compounds in the brain. Opioids, which comprise multiple families of peptides such as the endorphins, enkephalins, and dynorphins as well as their multiple receptor subtypes (mu, delta, kappa), are found in various networks throughout the brain but particularly within regions involved in emotional regulation, responses to pain and stress, endocrine regulation, and food intake (LaMotte, Snowman, Pert, & Snyder, 1978; Mansour et al., 1987). This distribution as well as extensive empirical work has led to the notion that opioids play a major role in diverse biological processes such as pain modulation, affect and emotion,

response to stress, and reinforcement (Van Ree et al., 2000). Much of the investigation of central opioids has been fueled by an interest in understanding the nature of addiction. Indeed, when naturally occurring opioid compounds were discovered, there was much excitement about the possibility that studies of endorphins and enkephalins would lead to the development of nonaddictive pain medications or improved treatment of narcotic addictions. However, it was clear from the intense research that soon followed these discoveries that endogenous opioids had very similar physiological profiles to exogenous opiate drugs such as morphine and heroin (i.e., tolerance and dependence). Nevertheless, studies of opioid peptide systems and the effects of exogenous opiate drugs have provided important insights into the nature of physical pain and psychological distress.

Although opioids mediate diverse functions in different brain regions and these functions may differ across species, several commonalities characterize them. Increased opioid function is clearly associated with positive affective states—for example, relief of pain; feelings of euphoria, well-being, or relaxation; feelings or behavior associated with social attachment; and pleasurable states associated with highly palatable foods. Herman and Panksepp (1981) and Panksepp et al. (1980) have conducted pioneering work on the role of opioids in behaviors related to social attachment and separation. A considerable body of research demonstrates that activation of opioid receptors promotes maternal behavior in mothers and attachment behavior and social play in juveniles. Separation distress, exhibited by archetypal behaviors and calls in most mammals and birds, is reduced by opiate agonists and increased by opiate antagonists in many species (Panksepp, 1998); maternal separation in rat pups also causes an opiate-mediated analgesia (Kehoe & Blass, 1986b). This distress behavior in the young serves as a powerful determinant of maternal behavior; upon such calls, mothers characteristically come back to and comfort their young. It has been theorized that touch, a powerful signal of care, activates endogenous opiate signals; for example, motivation for allogrooming in primates appears to be mediated by opiates (Graves, Wallen, & Maestripieri, 2002; Keverne, Martensz, & Tuite, 1989; Martel et al., 1993).

Perhaps the most remarkable effect of opiates is the reduction or elimination of pain. Pain is generally conceptualized to have both a physical and an affective component; often, we can describe the physical sensation induced by a painful stimulus, but additionally it induces a negative emotional state. Opiate drugs can act on both components of pain, probably at the spinal and cortical levels; they clearly augment the pain threshold but also induce statements in patients such as "I still feel the pain, but I don't mind it as much." Pain clearly serves as an enormously adaptive component in protecting the organism from further danger and eliciting escape responses. Why

then have a system that acts to reduce pain sensation? Endogenous opioids appear to modulate pain responses, perhaps dampening the transmission in pain fibers and pathways. One theory suggests that animals need to have a pain-modulation system in the face of acute threat or danger, which is temporarily activated but later deactivated when danger has subsided and recuperative behaviors take over (Bolles & Fanselow, 1980; Fanselow & Sigmundi, 1986). Thus, if an animal is injured during attack or flight, activation of opioid systems can reduce pain and facilitate adaptive escape. When the environment becomes safe, recuperative behaviors, such as rest, "licking one's wounds," and so on, are elicited. There is a large literature supporting the contention that endogenous opioids are activated during situations of physical or psychological stress (Bolles & Fanselow, 1982; Drolet et al., 2001). For example, exposure of a rat to shock or a human to a painful stimulus raises the pain threshold, but even exposure to non-physical stimuli such as conditioned fear cues, novelty, or cat smell can provoke opioid-mediated analgesia in rats (Bolles & Fanselow, 1982; Lester & Fanselow, 1985; Siegfried, Netto, & Izquierdo, 1987). A further interesting example is that imminent parturition in pregnant female rats results in a progressive opiate-mediated analgesia, which subsides after birth; human birth is also associated with an increase in circulating endorphins (Fajardo et al., 1994; Gintzler, 1980).

In addition to modulation of social bonding and pain, central opioids appear to play a key role in the affective response to palatable food. Many years ago, it was shown that morphine induces voracious eating in rats (Martin et al., 1963). Since that time, there has been extensive research showing that opioid activation of specific brain sites increases feeding, while antagonism of central opiate receptors with drugs such as naltrexone reduces feeding (reviews, Levine et al., 1985; Cooper & Kirkham, 1993; Kelley et al., 2002). A major facet of current hypotheses concerning opioid modulation of food intake is that opioids specifically regulate palatability and positive hedonic evaluation of food. For example, in humans, experimental work shows that naltrexone or naloxone reduces subjective ratings of food pleasantness while leaving feelings of hunger and taste recognition unchanged and reduces preference for sweet, high-fat foods (Fantino, Hosotte, & Apfelbaum 1986; Drenowski et al., 1992). In the taste reactivity test, noted earlier, morphine enhances taste palatability (Pecina & Berridge, 1995). In our own work in rats, we have found that opioids, particularly those with a preference for the mu receptor, potently increase the intake of normal chow as well as sucrose, salt, saccharine, and fat when injected into the nucleus accumbens. We have also found that rats thus treated will work harder and longer in an operant task for sugar pellets, even when not food-restricted (Zhang, Balmadrid, & Kelley, 2003). We have hypothesized that opioid-mediated mechanisms in the nucleus accumbens (and undoubtedly other brain regions)

mediate food "liking" or the pleasurable affective state induced by calorically dense foods. Thus, it seems that the positive emotional state induced by tasty, energy-dense foods is in part mediated by brain opioids. It interesting to speculate that this system may be responsible for the effect that "comfort foods" have on mood and general emotional state in humans. Supportive of this notion is the finding that in humans and other animals consumption of high-fat or sweet foods induces analgesia (Kanarek, Przypek, D'Anci, & Marks-Kaufman, 1997), suggesting that their consumption can literally reduce pain. Consumption of chocolate or sugar activates brain circuits encoding emotion and increases pleasurable feelings (Small et al., 2001).

In sum, opioid peptide-coded neural networks in striatal, limbic, and brain-stem regions appear to be fundamental substrates for certain affects. Lowered opioid peptide levels may signal distress, pain, and aversive motivation; enhanced peptide levels appear to be associated with safety and contentment. It is therefore not surprising that for thousands of years humans have chosen to activate this system artificially with opium, heroin, cannabinoids, and alcohol, all of which interact strongly with opioid systems.

Addictive Drugs and Artificial Stimulation of Emotions

Neurochemically coded brain circuits have evolved to serve as critical internal signals in guiding adaptive behavior and in maximizing fitness and survival. We have seen from the above account that the development of emotional–motivational systems in mammals has its molecular roots in ancestral behaviors of organisms millions and even billions of years ago. These systems enable animals to seek stimuli that enhance availability of resources (food, mating opportunities, safety, shelter) and to avoid danger or defend against predators. In humans, derangement or imbalance in these systems can lead to poor coping skills, emotional and mental distress, and psychopathologies such as anxiety, depression, and obsessive–compulsive disorder. For thousands of years, humans have used drugs that artificially stimulate these emotional systems. In the context of the present chapter, it is of interest to consider the use of drugs by humans within this evolutionary framework. Considerable advances have been made in understanding the neurobiological concomitants of addiction; however, it is only relatively recently that researchers have considered drug use and addiction from an evolutionary perspective (Nesse & Berridge, 1997; Panksepp, Knutson, & Burgdorf, 2002; Sullivan & Hagen, 2002).

Drugs serving as reinforcers are not a uniquely human phenomenon. Many species, such as rats, mice, and nonhuman primates, will directly self-administer most drugs that are used or abused by humans, such as alcohol,

heroin and other opiates, cannabinoids, nicotine, cocaine, amphetamine, and caffeine. Animals will perform an operant response—say, pressing a lever—in order to obtain an intravenous infusion of these compounds, in some cases (such as cocaine) to the point of death, ignoring other essential rewards such as food and water (Aigner & Balster, 1978). It is remarkable that even invertebrates prefer stimuli that are associated with exposure to drugs; for example, crayfish show positive place conditioning to psychostimulants (Panksepp & Huber, 2004; see Fig. 3.7), and 5-day-old rat pups learn to prefer odors that have been associated with morphine (Kehoe & Blass, 1986a). These behavioral findings suggest that there are common chemical and molecular substrates that rewarding drugs tap into across the animal kingdom.

Evidence supporting this hypothesis is mounting through the use of powerful molecular biological and genetic techniques. For example, cocaine acts primarily through its effects on the DA and 5-HT transporters, presynaptic uptake membrane proteins that control the levels of these transmitters in the synapse. These universal high-affinity monoamine transporters are found in nearly every species studied, for example, in C. *elegans* (Jayanthi et al., 1998). The DA transporter (DAT) protein has been characterized in *Drosophila* and shown to be the target for cocaine-stimulated behaviors in the fruit fly (Porzgen et al., 2001); DA is necessary for the activating effects of cocaine, nicotine, and ethanol in the fly (Bainton et al., 2000). In a notable study, it was found that, as for rodents, both D_1 and glutamate *N*-methyl 1-D-as partate (NMDA) receptors are involved in the cocaine response in the fruit fly; Torres and Horowitz (1998) comment that, "therefore as in rats, the NMDA (and D-1) receptor pathways in this arthropod represent obligatory targets for the behavioral effects of psychostimulants." This is remarkable given that the major substrates in cellular and behavioral plasticity with regard to learning and memory are the D_1 and NMDA receptors. A further example is provided by the protein DARPP-32. This intracellular signal-transduction protein (DA-regulated phosphoprotein) is an essential regulator of DA and glutamate signaling and plays a key role in cellular plasticity, learning, and addiction in mammalian models (Greengard et al., 1998). DARPP-32 immunoreactivity is also found in the lizard and turtle brain (Smeets, Lopez, & Gonzalez, 2001, 2003). A further example is the nicotinic receptor, an endogenous receptor for acetylcholine so named for its high affinity for nicotine binding. Different functional subunits that have been characterized in numerous species derive from primordial proteins over 1000 million years old (Changeux et al., 1998).

Thus, findings are accumulating that identify conserved genomic substrates and chemical pathways for psychoactive drug action across phyla. This knowledge addresses proximate causations of behavior (Tinbergen, 1963), or "how" drugs act in the brain to stimulate emotions. However, we are left

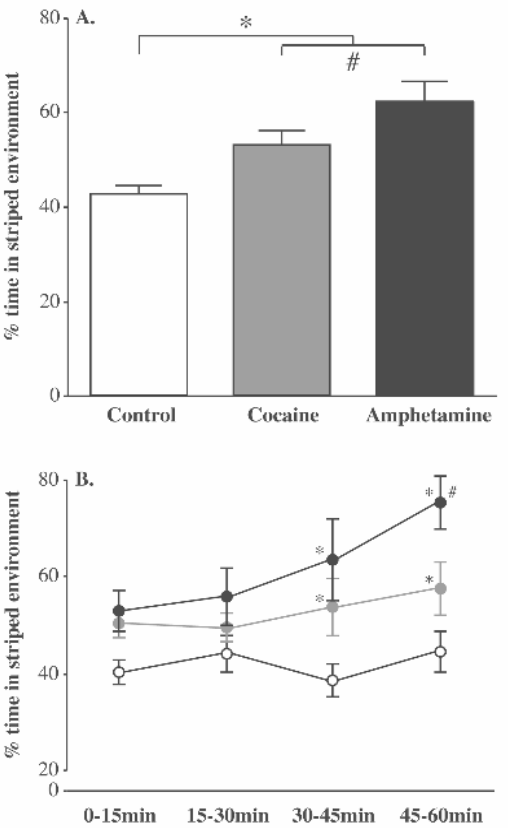


Figure 3.7. Crayfish prefer an environment associated with the psychostimulants amphetamine or cocaine. A: Crayfish were infused with the drugs and then placed in a striped visual environment; control infusions were associated with a plain visual environment. On a test trial following conditioning, no drug was given and the animals were allowed to swim through the aquarium. B: Significantly more time was spent in the presence of psychostimulant-paired contextual cues. * = significantly different with respect to control; # = significant difference between amphetamine and cocaine values. (From Panksepp & Huber, 2004, with permission.)

wondering about the ultimate or functional causations of behavior, or "why" drug use and addiction have evolved as major human behaviors. Clearly, chemical systems mediating emotions and adaptive survival behaviors did not evolve so that humans could discover the benefits of pleasurable drug states. The field of Darwinian medicine explores the mechanisms of natural selection that lead to vulnerability to disease (including addiction), and here some insights have been provided. Certain genotypes may have conferred

benefits that also presented vulnerabilities. For example, genes related to the DA system that may have enhanced novelty seeking may have provided advantages in seeking and finding new habitats and resources. In ancestral environments, such genetic quirks would be beneficial or at the very least not deleterious; however, in modern environments, with availability of pure drugs such as cocaine, disproportionate susceptibility among individuals may occur. Gerald and Higley (2002) have proposed a fascinating model for genetic susceptibility to alcohol dependence in relation to variations in serotonin function. Their research shows that monkeys with lower levels of brain 5-HT tend to be less affiliative and social, to be more aggressive and impulsive, and to have a higher mortality in the wild. These monkeys drink excessive amounts of alcohol compared to monkeys with high 5-HT levels. Thus, heritable traits that may have been advantageous in certain contexts could contribute to susceptibility to alcoholism and excessive alcohol intake.

Ultimately, it is critical to address the remarkable similarities between plant alkaloids and nervous system chemicals and receptors in animals. Figure 3.8 shows examples of cannabinoid and opiate receptors in the mammalian brain. Sullivan and Hagen (2002) ponder this question and propose that psychotropic substance seeking is an adaptation reflective of a coevolutionary relationship between psychotropic plant substances and humans that is millions of years old. Plants containing *allelochemicals* (toxic metabolites used by plants to discourage herbivores and pathogens) were widespread in the ancestral environment, and these alkaloids were often chemical analogues of vertebrate and invertebrate neurotransmitters.

> this "deep time" relationship is self-evident both in the extant chemical–ecological adaptations that have evolved in mammals to metabolize psychotropic plant substances and in the structure of plant defensive chemicals that have evolved to mimic the structure, and interfere with the function, of mammalian neurotransmitters. (Sullivan & Hagen, 2002)

Taking an anthropological point of view, these authors suggest that extensive evidence of substance use in antiquity may have been a mundane, ubiquitous activity similar to how we use caffeine in the present. These authors propose that there may have been selective and relatively specific benefits of plant use, particularly before the advent of agriculture. The use of the coca plant can be traced at least as far back as 7000 years ago, and Sullivan and Hagen (2002) cite archeological evidence that the betel nut (containing arecoline, a muscarinic agonist) was chewed 13,000 years ago in Timor and 10,700 years ago in Thailand. These authors suggest that in a foraging environment humans may have exploited these neurotransmitter analogue chemicals to enhance energy and fitness, particularly for nutritionally constrained

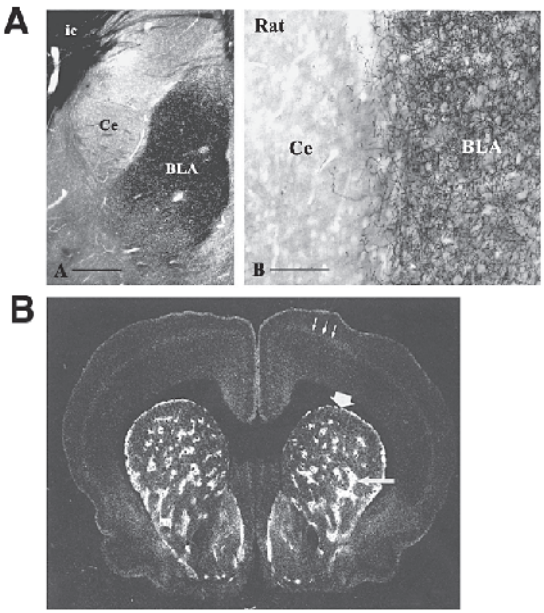


Figure 3.8. Receptors that selectively bind opiates and cannabinoids are present in the mammalian brain, perhaps indicating a coevolutionary relationship between humans and plant alkyloids, as discused in the text. (*A*) Strong expression of cannabinoid receptors in the basolateral nucleus of the amygdala in rat brain, an area involved in emotion regulation. On the left is a low-power view and on the right is a high-power view of sections stained for cannabinoid receptor immunoreactivity. BLA, Basolateral amygdala; Ce, central nucleus; ic, internal capsule. (From Katona et al., 2001, with permission.) (*B*) Localization of opiate receptor binding in the striatum of rat brain, utilizing 3H-naloxone autoradiography. Light staining against dark field indicates dense, patchy distribution of mu opiate receptor distribution in the dorsal and ventral striatum, areas important for learning and reinforcement processes. Small arrow in cortex indicates mu binding in layer k of cortex; larger arrow indicates intense binding in the subcallosal streak and patchy areas called "striosomes." (From Delfs et al., 1994, with permission.)

neurotransmitters (the monoamines and acetylcholine). This could bring a clear benefit in times of privation and resource scarcity. Behavioral, nutritional, and energetic advantages have been ascribed to ethanol consumption, present in low levels in ripe and fermenting fruit, which have been consumed by frugivore primates for 40 million years (Dudley, 2002).

Whatever the ultimate explanation for drug-seeking behavior, it is clear that there is a close evolutionary relationship between certain plant alkaloids

and brain neurotransmitters. Many of these compounds bind specifically to brain receptors and are able to induce feelings of positive emotion or pleasure, and relieve negative emotional states such as anxiety and depression. In the present ecological environment, the overabundance and availability of high quantities of pure drugs have resulted in maladaptive consequences of uncontrolled use and addiction.

CONCLUSIONS

The present chapter has provided a framework for thinking about the evolution of brain neurotransmitter systems that mediate motivational processes and emotional expression. Emotions (or their equivalent state) are required to activate adaptive behavior, from single-cell organisms to humans. Their elaboration and expression, when elicited by appropriate stimuli, are instantiated in complex but highly organized neural circuitry. A major feature of this circuitry, at least in mammalian brains, is reciprocal and feed-forward links between core motivational systems within the hypothalamus and higher-order corticostriatal and limbic structures. This cross-talk between cortical and subcortical networks enables intimate communication between phylogenetically newer brain regions, subserving subjective awareness and cognition, with ancestral motivational systems that exist to promote survival behaviors. Neurochemical coding, imparting an extraordinary amount of specificity and flexibility within these networks, appears to be conserved in evolution; several examples with monoamines and peptides have been provided above. Across the course of thousands of years, humans, through interactions with plant alkyloids, have discovered how to facilitate or blunt emotions with psychoactive drugs. Thus, while emotional systems generally serve a highly functional and adaptive role in behavior, they can be altered in maladaptive ways in the case of addiction. Future research will undoubtedly generate more insight into the chemical, genetic, and organizational nature of motivational–emotional systems.

References

Adler, J. (1966). Chemotaxis in bacteria. *Science*, *153*, 708–716.

Adler, J. (1969). Chemoreceptors in bacteria. *Science*, *166*, 1588–1597.

Adler, J. (1990). The sense of "smell" in bacteria: Chemotaxis in *E. coli*. In K. Colbow (Ed.), *R. H. Wright lectures on olfaction*. Burnaby, Canada: Simon Fraser University.

Adler, J., Hazelbauer, G. L., & Dahl, M. M. (1973). Chemotaxis toward sugars in *Escherichia coli*. *Journal of Bacteriology*, *115*, 824–847.

Aigner, T. G., & Balster, R. L. (1978). Choice behavior in rhesus monkeys: Cocaine versus food. *Science, 201*, 534–535.

Bainton, R. J., Tsai, L. T., Singh, C. M., Moore, M. S., Neckameyer, W. S., & Heberlein, U. (2000). Dopamine modulates acute responses to cocaine, nicotine and ethanol in *Drosophila*. *Current Biology, 10*, 187–194.

Baldwin, A. E., Sadeghian, K., & Kelley, A. E. (2002). Appetitive instrumental learning requires coincident activation of NMDA and dopamine D_1 receptors within the medial prefrontal cortex. *Journal of Neuroscience, 22*, 1063–1071.

Bassareo, V., & Di Chiara, G. (1999). Modulation of feeding-induced activation of mesolimbic dopamine transmission by appetitive stimuli and its relation to motivational state. *European Journal of Neuroscience, 11*, 4389–4397.

Becker, J. B., Rudick, C. N., & Jenkins, W. J. (2001). The role of dopamine in the nucleus accumbens and striatum during sexual behavior in the female rat. *Journal of Neuroscience, 21*, 3236–3241.

Berridge, K. C. (2001). Reward learning. In *The psychology of learning and motivation* (pp. 223–277). New York: Academic Press.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews, 28*, 309–369.

Bindra, D. (1978). How adaptive behavior is produced: A perceptual–motivational alternative to response-reinforcement. *Behavioral and Brain Sciences, 1*, 41–91.

Blackburn, J. R., Phillips, A. G., Jakubovic, A., & Fibiger, H. C. (1989). Dopamine and preparatory behavior: II. A neurochemical analysis. *Behavioral Neuroscience, 103*, 15–23.

Blair, H. T., Cho, J., & Sharp, P. E. (1998). Role of the lateral mammillary nucleus in the rat head direction circuit: A combined single unit recording and lesion study. *Neuron, 21*, 1387–1397.

Blenau, W., & Baumann, A. (2001). Molecular and pharmacological properties of insect biogenic amine receptors: Lessons from *Drosophila melanogaster* and *Apis mellifera*. *Archives of Insect Biochemistry and Physiology, 48*, 13–38.

Bolles, R. C. (1972). Reinforcement, expectancy and learning. *Psychological Review, 79*, 394–409.

Bolles, R. C., & Fanselow, M. S. (1980). A perceptual–defensive–recuperative model of fear and pain. *Behavioral and Brain Sciences, 3*, 291–301.

Bolles, R. C., & Fanselow, M. S. (1982). Endorphins and behavior. *Annual Review of Psychology, 33*, 87–101.

Brembs, B., Lorenzetti, F. D., Reyes, F. D., Baxter, D. A., & Byrne, J. H. (2002). Operant reward learning in *Aplysia*: Neuronal correlates and mechanisms. *Science, 296*, 1706–1709.

Buck, R. (1999). The biological affects: A typology. *Psychological Review, 106*, 301–336.

Cardinal, R. N., Parkinson, J. A., Hall, J., & Everitt, B. J. (2002). Emotion and motivation: The role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience and Biobehavioral Reviews, 26*, 321–352.

Cardinaud, B., Gilbert, J. M., Liu, F., Sugamori, K. S., Vincent, J. D., Niznik, H. B., & Vernier, P. (1998). Evolution and origin of the diversity of dopamine receptors in vertebrates. *Advances in Pharmacology*, *42*, 936–940.

Changeux, J. P., Bertrand, D., Corringer, P. J., Dehaene, S., Edelstein, S., Lena, C., Le Novere, N., Marubio, L., Picciotto, M., & Zoli, M. (1998). Brain nicotinic receptors: Structure and regulation, role in learning and reinforcement. *Brain Research Reviews*, *26*, 198–216.

Childress, A. R., Mozley, P. D., McElgin, W., Fitzgerald, J., Reivich, M., & O'Brien, C. P. (1999). Limbic activation during cue-induced cocaine craving. *American Journal of Psychiatry*, *156*, 11–18.

Coccaro, E. F., Siever, L. J., Klar, H. M., Maurer, G., Cochrane, K., Cooper, T. B., Mohs, R. C., & Davis, K. L. (1989). Serotonergic studies in patients with affective and personality disorders. Correlates with suicidal and impulsive aggressive behavior. *Archives of General Psychiatry*, *46*, 587–599.

Cofer, C. N., & Appley, M. H. (1964). *Motivation: Theory and research*. New York: Wiley.

Cooper, S. J., & Kirkham, T. C. (1993). Opioid mechanisms in the control of food consumption and taste preferences. In A. Herz (Ed.), *Handbook of experimental pharmacology* (pp. 239–262). Berlin: Springer-Verlag.

Coss, R. G., & Owings, D. H. (1989). Rattler battlers. *Natural History*, *48*, 30–35.

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*, 1413–1420.

Darlison, M. G., & Richter, D. (1999). Multiple genes for neuropeptides and their receptors: Co-evolution and physiology. *Trends in Neurosciences*, *22*: 81–88.

de Bono, M., & Bargmann, C. I. (1998). Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in C. *elegans*. *Cell*, *94*, 679–689.

Delfs, J. M., Kong, H., Mestek, A., Chen, Y., Yu, L., Reisine, T., & Chesselet, M. F. (1994). Expression of mu opioid receptor mRNA in rat brain: An *in situ* hybridization study at the single cell level. *Journal of Comparative Neurology*, *345*, 46–68.

Diaz-Rios, M., Oyola, E., & Miller, M. W. (2002). Colocalization of gamma-aminobutyric acid-like immunoreactivity and catecholamines in the feeding network of *Aplysia californica*. *Journal of Comparative Neurology*, *445*, 29–46.

Di Chiara, G. (1998). A motivational learning hypothesis of the role of mesolimbic dopamine in compulsive drug use. *Journal of Psychopharmacology*, *12*, 54–67.

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning Behavior*, *22*, 1–18.

Doudet, D., Hommer, D., Higley, J. D., Andreason, P. J., Moneman, R., Suomi, S. J., & Linnoila, M. (1995). Cerebral glucose metabolism, CSF 5-HIAA levels, and aggressive behavior in rhesus monkeys. *American Journal of Psychiatry*, *152*, 1782–1787.

Drewnowski, A., Krahn, D. D., Demitrack, M. A., Nairn, K., & Gosnell, B. A. (1992).

Taste responses and preferences for sweet high-fat foods: Evidence for opioid involvement. *Physiological Behavior, 51*, 371–379.

Drolet, G., Dumont, E. C., Gosselin, I., Kinkead, R., Laforest, S., & Trottier, J. F. (2001). Role of endogenous opioid system in the regulation of the stress response. *Progress in Neuropsychopharmacology and Biological Psychiatry, 25*, 729–741.

Dudley, R. (2002). Fermenting fruit and the historical ecology of ethanol ingestion: Is alcoholism in modern humans an evolutionary hangover? *Addiction, 97*, 381–388.

Ekman, P., & Davidson, R. J. (1994). *The nature of emotion: Fundamental questions*. New York: Oxford University Press.

Espana, R. A., Baldo, B. A., Kelley, A. E., & Berridge, C. W. (2001). Wake-promoting and sleep-suppressing actions of hypocretin (Orexin): Basal forebrain sites of action. *Neuroscience, 106*, 699–715.

Fajardo, M. C., Florido, J., Villaverde, C., Oltras, C. M., Gonzalez-Ramirez, A. R., & Gonzalez-Gomez, F. (1994). Plasma levels of beta-endorphin and ACTH during labor and immediate puerperium. *European Journal of Obstetrics, Gynecology, and Reproductive Biology, 55*, 105–108.

Fanselow, M. S., & Sigmundi, R. A. (1986). Species-specific danger signals, endogenous opioid analgesia, and defensive behavior. *Journal of Experimental Psychology: Animal Behavior Processes, 12*, 301–309.

Fantino, M., Hosotte, J., & Apfelbaum, M. (1986). An opioid antagonist, naltrexone, reduces preference for sucrose in humans. *American Journal of Physiology, 251*, R91–R96.

Figler, R. A., MacKenzie, D. S., Owens, D. W., Licht, P., & Amoss, M. S. (1989). Increased levels of arginine vasotocin and neurophysin during nesting in sea turtles. *General and Comparative Endocrinology, 73*, 223–232.

Gerald, M. S., & Higley, J. D. (2002). Evolutionary underpinnings of excessive alcohol consumption. *Addiction, 97*, 415–425.

Gintzler, A. R. (1980). Endorphin-mediated increases in pain threshold during pregnancy. *Science, 210*, 193–195.

Gotzes, F., Balfanz, S., & Baumann, A. (1994). Primary structure and functional characterization of a *Drosophila* dopamine receptor with high homology to human $D_{1/5}$ receptors. *Receptors and Channels, 2*, 131–141.

Graves, F. C., Wallen, K., & Maestripieri, D. (2002). Opioids and attachment in rhesus macaque (*Macaca mulatta*) abusive mothers. *Behavioral Neuroscience, 116*, 489–493.

Greengard, P., Nairn, A. C., Girault, J. A., Ouimet, C. C., Snyder, G. L., Fisone, G., Allen, P. B., Fienberg, A., & Nishi, A. (1998). The DARPP-32/protein phosphatase-1 cascade: A model for signal integration. *Brain Research Reviews, 26*, 274–284.

Gross, C., Zhuang, X., Stark, K., Ramboz, S., Oosting, R., Kirby, L., Santarelli, L., Beck, S., & Hen, R. (2002). Serotonin 1A receptor acts during development to establish normal anxiety-like behavior in the adult. *Nature, 416*, 396–400.

Herman, B. H., & Panksepp, J. (1981). Ascending endorphin inhibition of distress vocalization. *Science, 211*, 1060–1062.

Hess, W. R. (1957). *The functional organization of the diencephalon*. New York: Grune & Stratton.

Higley, J. D., Mehlman, P. T., Higley, S. B., Fernald, B., Vickers, J., Lindell, S. G., Taub, D. M., Suomi, S. J., & Linnoila, M. (1996). Excessive mortality in young free-ranging male nonhuman primates with low cerebrospinal fluid 5-hydroxyindoleacetic acid concentrations. *Archives of General Psychiatry*, *53*, 537–543.

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient nonreward events. *Neuroscience*, *96*, 651–656.

Hoyle, C. H. (1999). Neuropeptide families and their receptors: Evolutionary perspectives. *Brain Research*, *848*, 1–25.

Huber, R., Panksepp, J. B., Yue, Z., Delago, A., & Moore, P. (2001). Dynamic interactions of behavior and amine neurochemistry in acquisition and maintenance of social rank in crayfish. *Brain, Behavior and Evolution*, *57*, 271–282.

Huber, R., Smith, K., Delago, A., Isaksson, K., & Kravitz, E. A. (1997). Serotonin and aggressive motivation in crustaceans: Altering the decision to retreat. *Proceedings of the National Academy of Sciences of the USA*, *94*, 5939–5942.

Hull, C. L. (1943). *Principles of behavior*. New York: Appleton.

Insel, T. R., & Winslow, J. T. (1998). Serotonin and neuropeptides in affiliative behaviors. *Biological Psychiatry*, *44*, 207–219.

Insel, T. R., & Young, L. J. (2000). Neuropeptides and the evolution of social behavior. *Current Opinion in Neurobiology*, *10*, 784–789.

Ito, R., Dalley, J. W., Howes, S. R., Robbins, T. W., & Everitt, B. J. (2000). Dissociation in conditioned dopamine release in the nucleus accumbens core and shell in response to cocaine cues and during cocaine-seeking behavior in rats. *Journal of Neuroscience*, *20*, 7489–7495.

Izard, C. E. (1993). Four systems for emotion activation: Cognitive and noncognitive processes. *Psychological Review*, *100*, 68–90.

Jackson, D. M., & Westland-Danielsson, A. (1994). Dopamine receptors: Molecular biology, biochemistry, and behavioral aspects. *Pharmacology and Therapeutics*, *64*, 291–369.

Jacobs, B. L., & Azmitia, E. C. (1992). Structure and function of the brain serotonin system. *Physiological Reviews*, *72*, 165–229.

Jayanthi, L. D., Apparsundaram, S., Malone, M. D., Ward, E., Miller, D. M., Eppler, M., & Blakely, R. D. (1998). The *Caenorhabditis elegans* gene *T23G5.5* encodes an antidepressant- and cocaine-sensitive dopamine transporter. *Molecular Pharmacology*, *54*, 601–609.

Kabotyanski, E. A., Baxter, D. A., Cushman, S. J., & Byrne, J. H. (2000). Modulation of fictive feeding by dopamine and serotonin in *Aplysia Journal of Neurophysiology*, *83*, 374–392.

Kanarek, R. B., Przypek, J., D'Anci, K. E., & Marks-Kaufman, R. (1997). Dietary modulation of mu and kappa opioid receptor–mediated analgesia. *Pharmacology, Biochemistry and Behavior*, *58*, 43–49.

Kapsimali, M., Dumond, H., Le Crom, S., Coudouel, S., Vincent, J. D., & Vernier, P. (2000). Evolution and development of dopaminergic neurotransmitter systems in vertebrates. *Journal of the Society of Biology*, *194*, 87–93.

Katona, I., Rancz, E. A., Acsady, L., Ledent, C., Mackie, K., Hajos, N., & Freund, T. F. (2001). Distribution of CB1 cannabinoid receptors in the amygdala and their role in the control of GABAergic transmission. *Journal of Neuroscience*, *21*, 9506–9518.

Keefe, K. A., & Gerfen, C. R. (1996). D_1 dopamine receptor–mediated induction of zif268 and c-fos in the dopamine-depleted striatum: Differential regulation and independence from NMDA receptors. *Journal of Comparative Neurology*, *367*, 165–176.

Kehoe, P., & Blass, E. M. (1986a). Behaviorally functional opioid systems in infant rats: I. Evidence for olfactory and gustatory classical conditioning. *Behavioral Neuroscience*, *100*, 359–367.

Kehoe, P., & Blass, E. M. (1986b). Opioid-mediation of separation distress in 10-day-old rats: Reversal of stress with maternal stimuli. *Developmental Psychobiology*, *19*, 385–398.

Kelley, A. E., Bakshi, V. P., Haber, S. N., Steininger, T. L., Will, M. J., & Zhang, M. (2002). Opioid modulation of taste hedonics within the ventral striatum. *Physiology and Behavior*, *76*, 365–377.

Keverne, E. B., Martensz, N. D., & Tuite, B. (1989). Beta-endorphin concentrations in cerebrospinal fluid of monkeys are influenced by grooming relationships. *Psychoneuroendocrinology*, *14*, 155–161.

Koepp, M. J., Gunn, R. N., Lawrence, A. D., Cunningham, V. J., Dagher, A., Jones, T., Brooks, D. J., Bench, C. J., & Grasby, P. M. (1998). Evidence for striatal dopamine release during a video game. *Nature*, *393*, 266–268.

Kokay, I. C., & Mercer, A. R. (1996). Characterisation of dopamine receptors in insect (*Apis mellifera*) brain. *Brain Research*, *706*, 47–56.

LaMotte, C. C., Snowman, A., Pert, C. B., & Snyder, S. H. (1978). Opiate receptor binding in rhesus monkey brain: Association with limbic structures. *Brain Research*, *155*, 374–379.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience*, *23*, 155–184.

Lester, L. S., & Fanselow, M. S. (1985). Exposure to a cat produces opioid analgesia in rats. *Behavioral Neuroscience*, *99*, 756–759.

Levine, A. S., Morley, J. E., Gosnell, B. A., Billington, C. J., & Bartness, T. J. (1985). Opioids and consummatory behavior. *Brain Research Bulletin*, *14*, 663–672.

Lewis, B. L., & O'Donnell, P. (2000). Ventral tegmental area afferents to the prefrontal cortex maintain membrane potential "up" states in pyramidal neurons via D_1 dopamine receptors. *Cerebal Cortex*, *10*, 1168–1175.

Lord, J. A., Waterfield, A. A., Hughes, J., & Kosterlitz, H. W. (1977). Endogenous opioid peptides: Multiple agonists and receptors. *Nature*, *267*, 495–499.

MacLean, P. (1949). Psychosomatic disease and the "visceral brain." *Psychosomatic Medicine*, *11*, 338–353.

MacLean, P. D. (1958). The limbic system with respect to self preservation and the preservation of the species. *Journal of Nervous and Mental Disease*, *127*, 1–11.

MacLean, P. D. (1969). The hypothalamus and emotional behavior. In W. J. H. Nauta (Ed), *The hypothalamus*. Springfield, IL: Thomas.

MacLean, P. D. (1990). *The triune brain in evolution*. New York: Plenum.

Mann, J. J., Malone, K. M., Psych, M. R., Sweeney, J. A., Brown, R. P., Linnoila, M., Stanley, B., & Stanley, M. (1996). Attempted suicide characteristics and cerebrospinal fluid amine metabolites in depressed inpatients. *Neuropsychopharmacology*, *15*, 576–586.

Mansour, A., Kachaturian, H., Lewis, M. E., Akil, H., & Watson, S. J. (1987). Autoradiographic differentiation of mu, delta, and kappa opioid receptors in the rat forebrain and midbrain. *Journal of Neuroscience*, *7*, 2445–2464.

Marshall, J. F., Richardson, J. S., & Teitelbaum, P. (1974). Nigrostriatal bundle damage and the lateral hypothalamic syndrome. *Journal of Comparative and Physiological Psychology*, *87*, 808–830.

Martel, F. L., Nevison, C. M., Rayment, F. D., Simpson, M. J., & Keverne, E. B. (1993). Opioid receptor blockade reduces maternal affect and social grooming in rhesus monkeys. *Psychoneuroendocrinology*, *18*, 307–321.

Martin, W. R., Wikler, A., Eades, C. G., & Pescor, F. T. (1963). Tolerance to and physical dependence on morphine in rats. *Psychopharmacologia*, *4*, 247–260.

McDougall, W. (1908). *An introducton to social psychology*. London: Methuen.

Medawar, P. (1953). Some immunological and endocrinological problems raised by the evolution of viviparity in vertebrates. *Symposia of the Society for Experimental Biology and Medicine*, *7*, 320–338.

Mehlman, P. T., Higley, J. D., Faucher, I., Lilly, A. A., Taub, D. M., Vickers, J., Suomi, S. J., & Linnoila, M. (1994). Low CSF 5-HIAA concentrations and severe aggression and impaired impulse control in nonhuman primates. *American Journal of Psychiatry*, *151*, 1485–1491.

Menzel, R. (2001). Searching for the memory trace in a mini-brain, the honeybee. *Learning and Mem*ory, *8*, 53–62.

Menzel, R., Heyne, A., Kinzel, C., Gerber, B., & Fiala, A. (1999). Pharmacological dissociation between the reinforcing, sensitizing, and response-releasing functions of reward in honeybee classical conditioning. *Behavioral Neuroscience*, *113*, 744–754.

Miczek, K. A., Mos, J., & Olivier, B. (1989). Brain 5-HT and inhibition of aggressive behavior in animals: 5-HIAA and receptor subtypes. *Psychopharmacology Bulletin*, *25*, 399–403.

Missale, C., Nash, S. R., Robinson, S. W., Jaber, M., and Caron, M. G. (1998). Dopamine receptors: From structure to function. *Physiological Reviews*, *78*, 189–225.

Mogenson, G. J., Jones, D. L., & Yim, C. Y. (1980). From motivation to action: Functional interface between the limbic system and the motor system. *Progress in Neurobiology*, *14*, 69–97.

Nargeot, R., Baxter, D. A., Patterson, G. W., & Byrne, J. H. (1999). Dopaminergic synapses mediate neuronal changes in an analogue of operant conditioning. *Journal of Neurophysiology*, *81*, 1983–1987.

Neckameyer, W. S. (1996). Multiple roles for dopamine in *Drosophila* development. *Developmental Biology*, *176*, 209–219.

Neckameyer, W. S. (1998). Dopamine and mushroom bodies in *Drosophila*: Experience-dependent and -independent aspects of sexual behavior. *Learning and Memory*, *5*, 157–165.

Nemeroff, C. B. (1998). Psychopharmacology of affective disorders in the 21st century. *Biological Psychiatry*, *44*, 517–525.

Nesse, R. M., & Berridge, K. C. (1997). Psychoactive drug use in evolutionary perspective. *Science*, *278*, 63–66.

Niall, H. D. (1982). The evolution of peptide hormones. *Annual Review of Physiology*, *44*, 615–624.

Olds, J. (1958). Self-stimulatoin of the brain: Its use to study local effects of hunger, sex and drugs. *Science*, *127*, 315–324.

Panksepp, J. (1991). Emotional circuits of the mammalian brain: Implications for biological psychiatry. In K. T. Strongman (Ed.), *International review of studies on emotion* (Vol. 1, pp. 59–99). Chichester, UK: Wiley.

Panksepp, J. (1998). *Affective neuroscience*. New York: Oxford University Press.

Panksepp, J., Herman, B. H., Vilberg, T., Bishop, P., & DeEskinazi, F. G. (1980). Endogenous opioids and social behavior. *Neuroscience and Biobehavioral Reviews*, *4*, 473–487.

Panksepp, J., Knutson, B., & Burgdorf, J. (2002). The role of brain emotional systems in addictions: A neuro-evolutionary perspective and new "self-report" animal model. *Addiction*, *97*, 459–469.

Panksepp, J. B., & Huber, R. (2004). Ethological analyses of crayfish behavior: A new invertebrate system for measuring the rewarding properties of psychostimulants. *Behavioural Brain Research*, in press.

Papez, J. W. (1937). A proposed mechanism of emotion. *Archives of Neurology and Psychiatry*, *38*, 725–743.

Pecina, S., & Berridge, K. C. (1995). Central enhancement of taste pleasure by intraventricular morphine. *Neurobiology*, *3*, 269–280.

Peroutka, S. J., & Howell, T. A. (1994). The molecular evolution of G protein–coupled receptors: Focus on 5-hydroxytryptamine receptors. *Neuropharmacology*, *33*, 319–324.

Pert, C. B., & Snyder, S. H. (1973). Opiate receptor: Demonstration in nervous tissue. *Science*, *179*, 1011–1014.

Petrovich, G. D., Canteras, N. S., & Swanson, L. W. (2001). Combinatorial amygdalar inputs to hippocampal domains and hypothalamic behavior systems. *Brain Research. Brain Research Reviews*, *38*, 247–289.

Petrovich, G. D., Setlow, B., Holland, P. C., & Gallagher, M. (2002). Amygdalo-hypothalamic circuit allows learned cues to override satiety and promote eating. *Journal of Neuroscience*, *22*, 8748–8753.

Peyron, C., Tighe, D. K., van den Pol, A. N., de Lecea, L., Heller, H. C., Sutcliffe, J. G., & Kilduff, T. S. (1998). Neurons containing hypocretin (orexin) project to multiple neuronal systems. *Journal of Neuroscience*, *18*, 9996–10015.

Pfaff, D. W. (1980). *Estrogens and brain function: Neural analysis of a hormone-controlled mammalian reproductive behavior*. Berlin: Springer-Verlag.

Pfaus, J. G., Damsma, G., Nomikos, G. G., Wenkstern, D. G., Blaha, C. D., Phillips, A. G., & Fibiger, H. C. (1990). Sexual behavior enhances central dopamine transmission in the male rat. *Brain Research, 530,* 345–348.

Pfaus, J. G., & Phillips, A. G. (1991). Role of dopamine in anticipatory and consummatory aspects of sexual behavior in the male rat. *Behavioral Neuroscience, 105,* 727–743.

Porzgen, P., Park, S. K., Hirsh, J., Sonders, M. S., & Amara, S. G. (2001). The antidepressant-sensitive dopamine transporter in *Drosophila melanogaster*: A primordial carrier for catecholamines. *Molecular Pharmacology, 59,* 83–95.

Pratt, W. E., & Mizumori, S. J. (2001). Neurons in rat medial prefrontal cortex show anticipatory rate changes to predictable differential rewards in a spatial memory task. *Behavioural Brain Research, 123,* 165–183.

Qi, Y. L., & Adler, J. (1989). Salt taxis in *Escherichia coli* bacteria and its lack in mutants. *Proceedings of the National Academy of Sciences of the USA, 86,* 8358–8362.

Ragozzino, K. E., Leutgeb, S., & Mizumori, S. J. (2001). Dorsal striatal head direction and hippocampal place representations during spatial navigation. *Experimental Brain Research, 139,* 372–376.

Raleigh, M. J., & McGuire, M. T. (1991). Bidirectional relationships between tryptophan and social behavior in vervet monkeys. *Advances in Experimental Medicine and Biology, 294,* 289–298.

Risold, P. Y., Thompson, R. H., & Swanson, L. W. (1997). The structural organization of connections between hypothalamus and cerebral cortex. *Brain Research. Brain Research Reviews, 24,* 197–254.

Robinson, T. E., & Berridge, K. C. (1993). The neural basis of drug craving: An incentive-sensitization theory of addiction. *Brain Research Reviews, 18,* 247–291.

Russell, B. (1921). *The analysis of mind.* London: Allen & Unwin.

Salamone, J. D., Cousins, M. S., & Snyder, B. J. (1997). Behavioral functions of nucleus accumbens dopamine: Empirical and conceptual problems with the anhedonia hypothesis. *Neuroscience and Biobehavioral Reviews, 21,* 341–359.

Saper, C. B. (2000). Hypothalamic connections with the cerebral cortex. *Progress in Brain Research, 126,* 39–48.

Saper, C. B. (2002). The central autonomic nervous system: Conscious visceral perception and autonomic pattern generation. *Annual Review of Neuroscience, 25,* 433–469.

Saudou, F., Amara, D. A., Dierich, A., LeMeur, M., Ramboz, S., Segu, L., Buhot, M. C., & Hen, R. (1994). Enhanced aggressive behavior in mice lacking 5-HT1B receptor. *Science, 265,* 1875–1878.

Saudou, F., & Hen, R. (1994). 5-Hydroxytryptamine receptor subtypes: Molecular and functional diversity. *Advances in Pharmacology, 30,* 327–380.

Schoenbaum, G., Chiba, A. A., & Gallagher, M. (1998). Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nature Neuroscience, 1,* 155–159.

Schulkin, J. (1999). Hormonal regulation of sodium and water ingestion. In *The neuroendocrine regulation of behavior* (pp. 53–84). New York: Cambridge University Press.

Schultz, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience*, *1*, 199–207.

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, *13*, 900–913.

Schultz, W., Tremblay, L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex*, *10*, 272–284.

Seamans, J. K., Floresco, S. B., & Phillips, A. G. (1998). D_1 receptor modulation of hippocampal–prefrontal cortical circuits integrating spatial memory with executive functions in the rat. *Journal of Neuroscience*, *18*, 1613–1621.

Sell, L. A., Morris, J., Bearn, J., Frackowiak, R. S., Friston, K. J., & Dolan, R. J. (1999). Activation of reward circuitry in human opiate addicts. *European Journal of Neuroscience*, *11*, 1042–1048.

Siegfried, B., Netto, C. A., & Izquierdo, I. (1987). Exposure to novelty induces naltrexone-reversible analgesia in rats. *Behavioral Neuroscience*, *101*, 436–438.

Small, D. M., Zatorre, R. J., Dagher, A., Evans, A. C., & Jones-Gotman, M. (2001). Changes in brain activity related to eating chocolate: From pleasure to aversion. *Brain*, *124*, 1720–1733.

Smeets, W. J., Lopez, J. M., & Gonzalez, A. (2001). Immunohistochemical localization of DARPP-32 in the brain of the lizard, *Gekko gecko*: Co-occurrence with tyrosine hydroxylase. *Journal of Comparative Neurology*, *435*, 194–210.

Smeets, W. J., Lopez, J. M., & Gonzalez, A. (2003). Immunohistochemical localization of DARPP-32 in the brain of the turtle, *Pseudemys scripta elegans*: Further assessment of its relationship with dopaminergic systems in reptiles. *Journal of Chemical Neuroanatomy*, *25*, 83–95.

Smith-Roe, S. L., & Kelley, A. E. (2000). Coincident activation of NMDA and dopamine D_1 receptors within the nucleus accumbens core is required for appetitive instrumental learning. *Journal of Neuroscience*, *20*, 7737–7742.

Steiner, J. E. (1973). The gustofacial response: Observation on normal and anencephalic newborn infants. *Symposia on Oral Sensory Perception*, *4*, 254–278.

Steiner, J. E., Glaser, D., Hawilo, M. E., & Berridge, K. C. (2001). Comparative expression of hedonic impact: Affective reactions to taste by human infants and other primates. *Neuroscience and Biobehavioral Reviews*, *25*, 53–74.

Sullivan, R. J., & Hagen, E. H. (2002). Psychotropic substance-seeking: Evolutionary pathology or adaptation? *Addiction*, *97*, 389–400.

Sutton, M. A., & Beninger, R. J. (1999). Psychopharmacology of conditioned reward: Evidence for a rewarding signal at D_1–like dopamine receptors. *Psychopharmacology* (*Berlin*), *144*, 95–110.

Swanson, L. W. (2000). Cerebral hemisphere regulation of motivated behavior. *Brain Research*, *886*, 113–164.

Tempel, B. L., Livingstone, M. S., & Quinn, W. G. (1984). Mutations in the dopa decarboxylase gene affect learning in *Drosophila*. *Proceedings of the National Academy of Sciences of the USA*, *81*, 3577–3581.

Tidey, J. W., & Bergman, J. (1998). Drug discrimination in methamphetamine-

trained monkeys: Agonist and antagonist effects of dopaminergic drugs. *Journal of Pharmacology and Experimental Therapeutics*, *285*, 1163–1174.

Tinbergen, N. (1963). On the aims and methods of ethology. *Zeitschrift für Tierpsychologie*, *20*, 410–433.

Tisa, L. S., & Adler, J. (1992). Calcium ions are involved in *Escherichia coli* chemotaxis. *Proceedings of the National Academy of Sciences of the USA*, *89*, 11804–11808.

Tomkins, S. S. (1982). Affect theory. In P. Ekman (Ed.), *Emotion in the human face* (pp. 353–395). Cambridge: Cambridge University Press.

Torres, G., & Horowitz, J. M. (1998). Activating properties of cocaine and cocaethylene in a behavioral preparation of *Drosophila melanogaster*. *Synapse*, *29*, 148–161.

Tso, W. W., & Adler, J. (1974). Negative chemotaxis in *Escherichia coli*. *Journal of Bacteriology*, *118*, 560–576.

Ungerstedt, U. (1971). Adipsia and aphagia after 6-hydroxydopamine induced degeneration of the nigro-striatal dopamine system. *Acta Physiologica Scandinavica. Supplementum*, *367*, 95–122.

Vallone, D., Picetti, R., & Borrelli, E. (2000). Structure and function of dopamine receptors. *Neuroscience and Biobehavioral Reviews*, *24*, 125–132.

Van Kesteren, R. E., Smit, A. B., De Lange, R. P., Kits, K. S., Van Golen, F. A., Van Der Schors, R. C., De With, N. D., Burke, J. F., & Geraerts, W. P. (1995). Structural and functional evolution of the vasopressin/oxytocin superfamily: Vasopressin-related conopressin is the only member present in *Lymnaea*, and is involved in the control of sexual behavior. *Journal of Neuroscience*, *15*, 5989–5998.

Van Ree, J. M., Niesink, R. J., Van Wolfswinkel, L., Ramsey, N. F., Kornet, M. M., Van Furth, W. R., Vanderschuren, L. J., Gerrits, M. A., & Van den Berg, C. L. (2000). Endogenous opioids and reward. *European Journal of Pharmacology*, *405*, 89–101.

Vergnes, M., Depaulis, A., & Boehrer, A. (1986). Parachlorophenylalanine-induced serotonin depletion increases offensive but not defensive aggression in male rats. *Physiology and Behavior*, *36*, 653–658.

Virkkunen, M., & Linnoila, M. (1992). Psychobiology of violent behavior. *Clinical Neuropharmacology*, *15*(*Suppl. 1, Part A*), 233A–234A.

Volkow, N. D., Fowler, J. S., & Wang, G. J. (2002). Role of dopamine in drug reinforcement and addiction in humans: Results from imaging studies. *Behavioural Pharmacology*, *13*, 355–366.

Walker, R. J., Brooks, H. L., & Holden-Dye, L. (1996). Evolution and overview of classical transmitter molecules and their receptors. *Parasitology*, *113*(*Suppl*), S3–S33.

Wang, J., & O'Donnell, P. (2001). D_1 dopamine receptors potentiate nmda-mediated excitability increase in layer V prefrontal cortical pyramidal neurons. *Cerebral Cortex*, *11*, 452–462.

Wang, J. Q., & McGinty, J. F. (1996). Acute methamphetamine-induced zif/268, preprodynorphin, and proenkephalin mRNA expression in rat striatum depends

on activation of NMDA and kainate/AMPA receptors. *Brain Research Bulletin*, *39*, 349–357.

Watson, J. B. (1924). *Behaviorism*. New York: Norton.

Will, M. J., Franzblau, E. B., & Kelley A. E. (2004). The amygdala is critical for opioid-mediated "binge" eating of fat. *Neuroreport*, in press.

Williams, G. V., & Goldman-Rakic, P. S. (1995). Modulation of memory fields by dopamine D_1 receptors in prefrontal cortex. *Nature*, *376*, 572–575.

Wilson, C., Nomikos, G. G., Collu, M., & Fibiger, H. C. (1995). Dopaminergic correlates of motivated behavior: Importance of drive. *Journal of Neuroscience*, *15*, 5169–5178.

Wise, R. A., & Rompré, P. P. (1989). Brain dopamine and reward. *Annual Review of Psychology*, *40*, 191–225.

Yeh, S. R., Fricke, R. A., & Edwards, D. H. (1996). The effect of social experience on serotonergic modulation of the escape circuit of crayfish. *Science*, *271*, 366–369.

Yellman, C., Tao, H., He, B., & Hirsh, J. (1997). Conserved and sexually dimorphic behavioral responses to biogenic amines in decapitated *Drosophila*. *Proceedings of the National Academy of Sciences of the USA*, *94*, 4131–4136.

Young, P. T. (1943). *Emotion in man and animal*. New York: Wiley.

Zhang, M., Balmadrid, C., & Kelley, A. E. (2003). Nucleus accumbens opioid, GABAergic, and dopaminergic modulation of palatable food motivation: Contrasting effects revealed by a progressive ratio schedule in the rat. *Behavioral Neuroscience*, *117*, 202–211.

This page intentionally left blank

4 Toward Basic Principles for Emotional Processing

What the Fearful Brain Tells the Robot

JEAN-MARC FELLOUS
AND JOSEPH E. LEDOUX

The field of neuroscience has, after a long period of looking the other way, embraced "emotion" as an important research area. Important progress has come from animal studies of fear, and especially fear conditioning in rats. This work has contributed to a re-evaluation of the concept of the "limbic system," and has identified the amygdala as a crucial component of the system involved in the acquisition, storage, and expression of fear memory. Researchers now understand how fearful stimuli enter, travel through, and exit the amygdala. Mechanistically, the amygdala acts as a species-specific danger detector that can be quickly activated by threatening stimuli, and that can be modulated by higher cognitive systems. In turn, the amygdala influences the cognitive system by way of projections to "arousal" centers that control the way actions and perceptions are performed.

Further research has shown that such findings from experimental animals also apply to the human brain and has directed attention to another important component of the emotional brain: the prefrontal cortex. Together, the amygdala and the prefrontal cortex can account for higher forms of fear that involve consciousness.

We conclude by discussing some recent results on positive emotions such as attachment, and by listing a set of rules that have emerged from the neuroscience of fear. These rules can inform future attempts at implementing fear and other emotions in artifacts such as robots.

The approach presented here is a straightforward experimental approach to emotion, which avoided vague concepts such as "affect," "hedonic tone," and "emotional feelings." It is important that the mistakes of the past not be made again, and that we expand from this foundation into broader aspects of mind and behavior.

The current wave of interest in the neural bases of emotion raises the question of why emotion was overlooked for so long. We will consider this question before examining what has been learned about emotional circuits because interest in emotion has now reached other research domains, such as computer science and robotics. These new areas of investigation are now faced with the same challenges that faced neuroscience a few decades ago.

WHY DID INTEREST IN EMOTION WANE?

As soon as pioneering brain researchers in the late 19th century identified regions of the brain involved in sensory perception and movement control (the neocortex), William James (1890) asked whether emotions might be explained in terms of these functions or whether emotion was the business of a separate, yet undiscovered brain system. Being a pragmatist, he proposed a theory of emotion based solely on functions of sensory and motor systems. Specifically, he argued that emotionally arousing stimuli are perceived by the sensory cortex, which activates the motor cortex to produce bodily responses appropriate to the emotionally arousing stimulus. Emotional feelings then result when the sensory cortex perceives the sensations that accompany bodily responses. Since different emotions involve different bodily responses, they have different sensory signatures and thus feel different. The essence of James' theory is captured by his conclusion that we do not run from a bear because we feel afraid but, instead, we feel afraid because we run. James' theory was quickly refuted by research showing that complete removal of the neocortex failed to disrupt the expression of emotional responses elicited by sensory stimuli; sensory and motor cortex could therefore not be the key.

During the first half of the 20th century, brain researchers were immensely interested in the brain mechanisms of emotional behavior. Some

of the early pioneers in neuroscience worked in this area, including Cannon (1987), Papez (1937), and Hebb (1949), to name but a few. Responses that occur when we defend against danger, interact with sexual partners, fight with an enemy, or have a tasty bite to eat promote the survival of individuals and their species. Emotional responses are thus inherently interesting and important. Why then did research on the brain mechanisms of emotion slow down after mid-century and become supplanted by studies of seemingly more elementary questions such as the neural bases of drives and reinforcement (Olds, 1977)?

For one thing, emotion research was a victim of the cognitive revolution. The emergence of cognitive science shifted the interest of those concerned with the relation between psychological functions and neural mechanisms toward processes (perception and memory, e.g.) that were readily thought of in terms of computer-like operations and that eventually contributed to the creation of new fields of investigation, such as artificial intelligence, that in turn reinforced and sustained the cognitive revolution. From the start, many cognitive scientists claimed that their field was not about emotion and other such topics (see Neisser, 1967; Gardner, 1987; but, contra this, see Miller, Galanter, & Pribam, 1960; Simon, 1967).

Another factor was that the limbic system concept (MacLean, 1949, 1952) provided an appealing and convincing theory that was the culmination of research on the brain mechanisms of emotion by many researchers extending back to the late 19th century (see LeDoux, 1987, 1991). Studies of how the brain mediates cognitive processes seemingly had a long way to go to catch up with the deep understanding that had been achieved about emotions, and researchers flocked to the new and exciting topic of cognition and the brain to begin filling the gap.

Cognitive questions also seemed more tractable than emotional ones, due in part to the dark cloud of subjectivity that hung over the topic of emotion and that created a "credibility problem" (LeDoux, 2002). While subjective experience and its relation to neural mechanisms is potentially a difficulty for any area of psychology, cognitive scientists figured out how to study mental processes involved in, for example, memory and perception without having to involve subjectivity. They showed, for example, that it is possible to study how the brain processes (computes and represents) external stimuli without first resolving how the conscious perceptual experiences come about. In fact, it is widely recognized that most cognitive processes occur unconsciously, with only the end products reaching awareness and only sometimes (Kihlstrom, 1987). Emotion researchers, though, remained focused on subjective emotional experience. In spite of the fact that most research on emotions and the brain was, and still is, conducted in experimental animals, creatures in which subjective states are difficult, if not impossible,

to prove, theoretical discussions of emotions and the brain typically reverted back to the age-old question of feelings.

However, even if we can account for important aspects of emotion in nonhuman animals without having to resort to subjective states, this should not be taken to mean that subjective states exist only in humans. Nonhuman animals might have domain-specific forms of consciousness, and in the case of nonhuman primates domain-independent forms of nonverbal consciousness, but only humans have verbal working memory (see below) and, thus, language-based consciousness and the mental frills that language makes possible (see Chapter 12, Arbib). The problem is that as soon as we rely on subjective states to explain behavior, we confront our inability to know whether such states really exist in creatures other than humans (LeDoux, 2002). If animals experience some subjective states of emotion, then why not robots as well? We come back to this issue in the context of feelings.

The main lesson to be learned from this brief excursion into history is that emotion researchers, whether in neuroscience or in other fields, need to figure out how to escape from the shackles of subjectivity if emotion research is to thrive. Ironically, cognitive science and artificial intelligence, which led to the neglect of emotion research, may also be able to help in its resurrection by providing a strategy that allows the study of emotion independently of subjective emotional experiences. Contrary to the intuitions that many people have about emotion, then, we shall argue that it is possible to ask how the brain processes emotional information (e.g., detects and responds to danger) without necessarily first solving the question of where conscious emotional feelings come from. Indeed, emotional responses, like cognitive processes, involve unconscious processing mechanisms (Ohman, 1992; LeDoux, 1996; Glascher & Adolphs, 2003). If we want to understand feelings, it is very likely going to be necessary to figure out how the more basic systems work. Failure to come to terms theoretically with the importance of processing systems that operate essentially unconsciously has been a major impediment to progress in understanding the neural basis of emotion. To overcome this, brain researchers and designers of complex artificial artifacts, such as autonomous robots, need to be more savvy about the often unconscious nature of emotions, rather than simply relying on common-sense beliefs about emotions as subjective feeling states. We shall speak of the *processing approach* to emotion as the approach we espouse here, which grounds emotion in possibly unconscious processes.

Any approach that omits emotions, motivations, and the like paints a highly unrealistic view of real minds. Minds are neither purely cognitive nor purely emotional but both, and more. Inclusion of work on emotion within the cognitive science and artificial intelligence frameworks can help rescue these fields from their often sterile approach to the mind as an information-

processing device that may pursue abstract goals but lacks motivation, strivings, desires, fears, and hopes.

In this connection, we should mention the so-called cognitive approach to emotions, which treats emotions as appraisals, i.e., thoughts about situations (Arnold, 1960; Schacter & Singer, 1962; Mandler, 1984; Frijda, 1986, 1993; Ellsworth, 1991; Lazarus, 1991; Scherer, 1993; see also Chapter 7, Ortony et al.). While some appraisal theorists allow for unconscious appraisals (which is consistent with a processing approach), most emphasize appraisals as conscious thoughts and use verbal self-report to understand the nature of the appraisal process. Conscious appraisals may indeed occur during an emotional state, but there are other, more fundamental processes at work as well. An understanding of these more fundamental processes is what the processing approach is all about.

The processing approach allows us to study unconscious emotional functions similarly in humans and other animals and at the same time offers an approach to understanding emotional consciousness (feelings) as well (since feelings themselves result from processes that occur unconsciously). In addition, the processing approach offers another advantage. It allows emotion and cognition to be treated similarly (as unconscious processes that can, but do not necessarily, lead to conscious experiences), and it opens the door for the much-needed integration of cognition, emotion, and motivation—the mental trilogy (LeDoux, 2002). Whether emotion, motivation, and cognition are three distinct but tightly interacting systems or whether emotion is an integral architectural feature of the cognitive and motivational systems (or vice-versa) remains to be established.

SHOULD WE INTEGRATE THE COGNITIVE BRAIN WITH THE LIMBIC SYSTEM?

The rise of cognitive science led to important advances in understanding the brain mechanisms of perception, attention, memory, and other cognitive processes. One might be tempted to say that the way to foster the synthesis of cognition and emotion into a new science of mind would be to put all this new information about the cognitive brain together with the view of the emotional brain provided by the limbic system concept put forth in the context of an evolutionary explanation of mind and behavior (MacLean, 1949, 1952; Isaacson, 1982). However, this would be a mistake. In spite of the fact that the limbic system concept remains the predominant view about how the brain makes emotions, it is a flawed and inadequate theory of the emotional brain.

The limbic system concept was built upon the view, promoted by comparative anatomists in the first half of the 20th century, that the neocortex

is a mammalian specialization: other vertebrates have a primordial cortex, but only mammals were believed to have a neocortex. Since thinking, reasoning, memory, and problem solving are especially well developed in mammals, particularly in humans and other primates that have relatively more neocortical tissue, it was argued that these cognitive processes must be mediated by the neocortex and not by the old cortex or other brain areas. In contrast, the old cortex and related subcortical ganglia form the limbic system, which was said to mediate the evolutionarily older aspects of mental life and behavior, our emotions. In this way, cognition came to be thought of as the business of the neocortex and emotions of the limbic system.

The limbic system theory ran into trouble when it was discovered, in the mid-1950s, that damage to the hippocampus, the centerpiece of the limbic system, led to severe deficits in a distinctly cognitive function, episodic long-term memory (Scoville & Milner, 1957). This was incompatible with the original idea that the primitive architecture of the limbic system, especially of the hippocampus, was poorly suited to participate in cognitive functions (MacLean, 1949, 1952). Subsequently, in the late 1960s, it was discovered that the equivalent of mammalian neocortex is present, though rudimentary, in nonmammalian vertebrates (Nauta & Karten, 1970). As a result, the old/new cortex distinction broke down, challenging the evolutionary basis of the assignment of emotion to the limbic system and cognition to the neocortex (Swanson, 1983).

The limbic system itself has been a moving target. Within a few years after the inception of the theory, it expanded from the original notion of "old cortex" and related subcortical forebrain nuclei to include some areas of the midbrain and even some regions of the neocortex. Several attempts have been made to salvage the limbic system by defining it more precisely (Livingston & Escobar, 1971; Isaacson, 1982; Swanson, 1983). Nevertheless, after half a century of debate and discussion, there are still no agreed-upon criteria that can be used to decide which areas of the brain belong to the limbic system. Some have suggested that the concept be abandoned (LeDoux, 1987, 1991; Kotter & Meyer, 1992).

In spite of these difficulties, the limbic system continues to survive, both as an anatomical concept and as an explanation of emotions, in textbooks, research articles, and scientific lectures. This is in part attributable to the fact that both the anatomical concept and the emotional function it was supposed to mediate were defined so vaguely as to be irrefutable. For example, in most discussions of how the limbic system mediates emotion, the meaning of the term *emotion* is presumed to be something akin to the common English-language use of the term, which is to say *feelings*. However, the common English use of *emotion* is at best a poor theoretical notion, for emotion is a rich and complex theoretical concept with many subtle aspects, some

of which are nonintuitive and thus inconsistent with the common use of the term (Ekman & Davidson, 1994; LeDoux, 1996). On the neural side, the criteria for inclusion of brain areas in the limbic system remain undefined, and evidence that any limbic area (e.g., the amygdala, which we will discuss below), however defined, contributes to any aspect of any emotion has been claimed to validate the whole concept. Mountains of data on the role of limbic areas in emotion exist, but there is still very little understanding of how our emotions might be the product of the limbic system.

Particularly troubling is the fact that one cannot predict, on the basis of the original limbic theory of emotion or any of its descendants, how specific aspects of emotion work in the brain. The explanations are all post hoc. Nowhere is this more apparent than in recent work using functional imaging to study emotions in the human brain. Whenever a so-called emotional task is used and a limbic area activated, the activation is explained by reference to the fact that limbic areas mediate emotion. When a limbic area is activated in a cognitive task, it is often assumed that there must have been some emotional undertone to the task. We are, in other words, at a point where the limbic theory has become a self-contained circularity. Deference to the concept is inhibiting creative thought about how mental life is mediated by the brain.

Although the limbic system theory is inadequate as an explanation of the specific brain circuits of emotion, MacLean's original ideas are quite interesting in the context of a general evolutionary explanation of emotion and the brain. In particular, the notion that emotions involve relatively primitive circuits that are conserved throughout mammalian evolution seems right on target. Further, the idea that cognitive processes might involve other circuits and might function relatively independently of emotional circuits, at least in some circumstances, also seems correct. These general functional ideas are worth retaining, even if we abandon the limbic system as a structural theory of the emotional brain. They also may be key in other areas of investigation of emotion, such as artificial intelligence.

ESCAPING THE LIMBIC SYSTEM LEGACY: FEAR CIRCUITS

The limbic system theory failed in part because it attempted to account for all emotions at once and, in so doing, did not adequately account for any one emotion. A more fruitful strategy is to take the opposite approach and study one emotion in detail. Our own approach has focused on the study of fear, but the basic principles that have been uncovered about the fear system are likely to be applicable to other systems. Different brain circuits may be involved in different emotion functions, but the relation of specific emotional

processing circuits to sensory, cognitive, motor, and other systems is likely to be similar across emotion categories. Some progress has also been made in understanding emotions other than fear, as will be discussed below.

The neural system underlying fear has been studied especially in the context of the behavioral paradigm called "fear conditioning" (Blanchard, Blanchard, & Fial, 1970; Davis, 1992; Kapp, Whalen, Supple, & Pascoe, 1992; LeDoux, 1996, 2000; Fanselow & LeDoux, 1999). In this work, the fear system has been treated as a set of processing circuits that detect and respond to danger, rather than as a mechanism through which subjective states of fear are experienced. Measurable correlates of fear include blood pressure changes, freezing responses, and release of pituitary–adrenal stress hormones. Through such measurements, fear is operationalized, or made experimentally tractable. Some limbic areas turn out to be involved in the fear system, but the exact brain areas and the nature of their involvement would never have been predicted by the limbic system theory. This operationalization of emotion may also lead to interesting work in robotics. The understanding of the processing circuits that detect and respond to danger can be used to design new types of sensor, effector, and controlling device that together would make up an "operationally fearful" autonomous robot. The general question of the role of fear and its complex interactions with cognition and with other emotional circuits can then be addressed explicitly in the fully controlled and measurable environment of the robot and can potentially give insight into the role of fear in humans and other animals.

Before describing research on fear in detail, several other approaches to the study of emotion and the brain that will not be discussed further should be mentioned. One involves stimulus–reward association learning (Aggleton & Mishkin, 1986; Everitt & Robbins, 1992; Gaffan, 1992; Ono & Nishijo, 1992; Rolls, 1998), another involves the role of septo–hippocampal circuits in anxiety (Gray, 1982), and still another involves distinct hypothalamic and brain-stem circuits for several different emotions (Panksepp, 1998; Siegel, Roeling, Gregg, & Kruk 1999).

What Is Fear Conditioning?

Since Pavlov (1927), it has been known that an initially neutral stimulus (a conditioned stimulus, or CS) can acquire affective properties upon repeated temporal pairings with a biologically significant event (the unconditioned stimulus, or US). As the CS–US relation is learned, innate physiological and behavioral responses come under the control of the CS (Fig. 4.1). For example, if a rat is given a tone CS followed by an electric shock US, after a few tone–shock pairings (one is often sufficient), *defensive responses* (responses that typi-

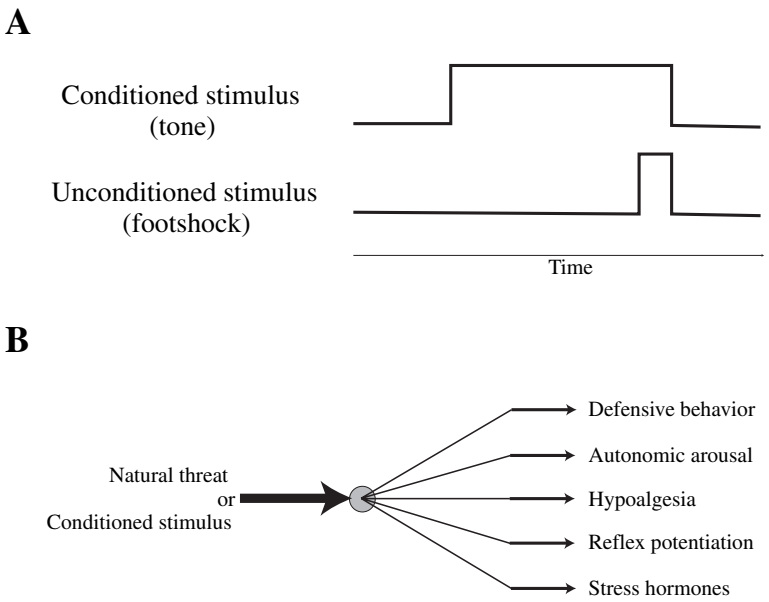


Figure 4.1. (*A*) Fear conditioning involves the presentation of a noxious unconditioned stimulus (US), such as footshock, at the end of the occurrence of a neutral conditioned stimulus (CS), such as a tone. (*B*) After conditioning, the CS elicits a wide range of behavioral and physiological responses that characteristically occur when an animal encounters a threatening or fear-arousing stimulus. Thus, a rat that has been fear-conditioned to a tone will express the same responses to a CS as to a natural threat (e.g., a cat). (Adapted

cally occur in the presence of danger) will be elicited by the tone alone. Examples of species-typical defensive responses that are brought under the control of the CS include behaviors such as freezing in rodents and autonomic (e.g., heart rate, blood pressure) and endocrine (e.g., hormone release) responses, as well as alterations in pain sensitivity (hypoanalgesia) and reflex expression (fear-potentiated startle and eye blink responses). This form of conditioning works throughout the phyla, having been observed in flies, worms, snails, fish, pigeons, rabbits, rats, cats, dogs, monkeys, and humans.

Research from several laboratories combined in the 1980s to paint a relatively simple and remarkably clear picture of the neuroanatomy of fear conditioning (Davis, 1992; Kapp, Whalen, Supple, & Pascoe 1992; LeDoux, 1992; Fanselow & Gale, 2003). In such studies, the CS and US are typically an audible tone and a foot shock, and the responses measured include freezing. It was shown that fear conditioning is mediated by the transmission of information about the CS and US to a small almond-shaped area (the

amygdala) and the control of fear reactions by way of output projections from the amygdala to behavioral, autonomic, and endocrine response control systems located in a collection of nuclei, altogether referred to as the "brain stem." We briefly describe below the input and output pathways, as well as the connections within the amygdala. The focus will be on findings from rodents and other small mammals as most of the work on fear conditioning has involved these species.

The amygdala consists of approximately 12 different regions, each of which can be further divided into several subregions. Although a number of different schemes have been used to label amygdala areas (Krettek & Price, 1978; Amaral, Price, Pitkanen, & Carmichael, 1992), the scheme adopted by Amaral et al. (1992) for the primate brain and applied to the rat brain by Pitkanen et al. (1997) will be followed here. The areas of most relevance to fear conditioning include the following nuclei: lateral (LA), basal (B), accessory basal (AB), central (CE), and intercalated (IC), as well as connections between them. Studies in several species, including rats, cats, and primates, are in close agreement about the connections of LA, B, AB, and CE (Amaral, Price, Pitkanen, & Carmichael, 1992; Paré, Smith, & Paré, 1995; Pitkanen, Savander, & LeDoux, 1997; Paré, Royer, Smith, & Lang, 2003). In brief, LA projects to B, AB, and CE and both B and AB also project to CE; IC is also an intermediate step between LA/B and CE. However, it is important to recognize that the connections of these areas are organized at the level of subnuclei within each region rather than at the level of the nuclei themselves (Pitkanen, Savander, & LeDoux, 1997). For simplicity, though, we will for the most part focus on nuclei rather than subnuclei.

The pathways through which CS inputs reach the amygdala have been studied extensively in recent years. Much of the work has involved the auditory modality, which is focused on here. Auditory and other sensory inputs to the amygdala terminate mainly in LA (Amaral, Price, Pitkanen, & Carmichael, 1992; Mascagni, McDonald, & Coleman, 1993; Romanski & LeDoux, 1993; McDonald, 1998), and damage to LA interferes with fear conditioning to an acoustic CS (LeDoux, Cicchetti, Xagoraris, & Romanski, 1990). Auditory inputs to LA come from both the auditory portion of the thalamus (a brain center considered to be a point of convergence of the perceptual senses en route to the rest of the brain) and auditory cortex, where complex sound interpretation is achieved (LeDoux, Cicchetti, Xagoraris, & Romanski, 1990; Mascagni, McDonald, & Coleman, 1993; Romanski & LeDoux, 1993). Fear conditioning to a simple auditory CS can be mediated by either of these pathways (Romanski & LeDoux, 1992) (Fig. 4.2). It appears that the projection to LA from the auditory cortex is involved with a more complex auditory stimulus pattern (Jarrell et al., 1987), but the exact conditions that require the cortex are poorly understood (Armony & LeDoux,

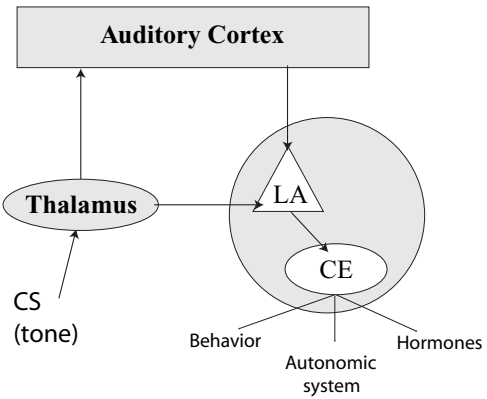


Figure 4.2. The neural pathways involved in fear conditioning are well characterized. When the conditioned stimulus (CS) is acoustic, the pathways involve transmission to the lateral nucleus of the amygdala (LA) from auditory processing areas in the thalamus and auditory cortex. LA, in turn, projects to the central nucleus of the amygdala (CE), which controls the expression of fear responses by way of projections to brain-stem areas controlling the autonomic nervous system, the production of hormones, and the appropriate behavior.

1997). Although some lesion studies have questioned the ability of the thalamic pathway to mediate conditioning (Shi & Davis, 1999), recordings from single neurons show that the cortical pathway conditions more slowly over trials than the thalamic pathway (Quirk, Armony, & LeDoux, 1997), thus indicating that the association between CS and US in the amygdala occurs initially through the thalamic pathway. Recent functional magnetic resonance imaging (fMRI) studies in humans have found that the amygdala shows activity changes during conditioning that correlate with activity in the thalamus but not the cortex (Morris, Ohman, & Dolan, 1999), further emphasizing the importance of the direct thalamo-amygdala pathway.

In addition to expressing fear responses to the CS, rats exhibit these when returned to the chamber in which the tone and shock were paired or a chamber in which shocks occur alone. This is called "contextual fear conditioning," where context refers to the various visual and olfactory aspects of the chamber, and requires both the amygdala and hippocampus, a brain structure know to enable long-term memories (Kim & Fanselow, 1992; Phillips & LeDoux, 1992; Maren, Aharonov, & Fanselow, 1997; Frankland et al., 1998). Areas of the hippocampus project to B and AB in the amygdala (Canteras & Swanson, 1992), and damage to these areas interferes with contextual conditioning (Maren & Holt, 2000). Hippocampal projections to B and AB thus seem to be involved in contextual conditioning (for a

comparison of the amygdala pathways involved in conditioning to a tone CS and to a context, see Fig. 4.3).

Given that LA is the site of termination of pathways carrying acoustic CS inputs, it is important to ask whether US inputs might also reach this area and potentially lead to CS–US association. Thalamic areas that receive afferents from the spinothalamic tract (LeDoux et al., 1987) project to LA (LeDoux, Farb, & Ruggiero, 1990) (Fig. 4.3). Further, cells in LA are responsive to nociceptive stimulation, and some of the same cells respond to auditory inputs as well (Romanski & LeDoux, 1993). Thus, the substrate for conditioning exists in LA.

Cortical areas that process somatosensory stimuli, including nociceptive stimuli, also project to LA and some other amygdala nuclei (Turner & Zimmer, 1984; McDonald, 1998). Recent behavioral studies show that conditioning can be mediated by US inputs to the amygdala from either thalamic or corti-

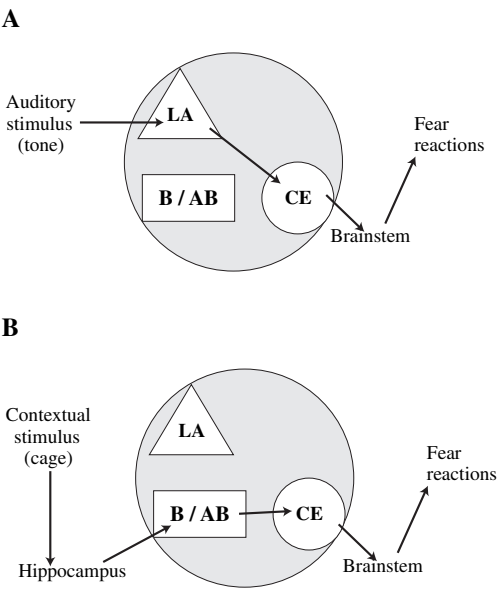


Figure 4.3. (*A*) Conditioning to a tone involves projections from the auditory system to the lateral nucleus of the amygdala (LA) and from LA to the central nucleus of the amygdala (CE). (*B*) In contrast, conditioning to the apparatus and other contextual cues present when the conditioned stimulus and unconditioned stimulus are paired involves the representation of the context by the hippocampus and the communication between the hippocampus and the basal (B) and accessory basal (AB) nuclei of the amygdala, which in turn project to CE. As for tone conditioning, CE controls the expression of the responses.

cal areas (Shi & Davis, 1999), a finding that parallels the conclusions above concerning CS inputs.

The AB amygdala receives inputs from the posterior thalamic area (LeDoux, Farb, & Ruggiero, 1990), which is a terminal region of the spinothalamic tract (LeDoux et al., 1987). While AB does not receive CS inputs from auditory systems, it does receive inputs from the hippocampus (Canteras & Swanson, 1992). The hippocampus, as described above, is necessary for forming a representation of the context, and these contextual representations, transmitted from the hippocampus to AB, may be modified by the US inputs to the AB.

The CE receives nociceptive inputs from the parabrachial area (Bernard & Besson, 1990) and directly from the spinal cord (Burstein & Potrebic, 1993). Although CE does not receive inputs from sensory areas processing acoustic CS, it is a direct recipient of inputs from LA, B, and AB. Also, US inputs to CE could be involved in higher-order integration. For example, representations created by CS–US convergence in LA or context–US convergence in AB, after transfer to CE, might converge with and be further modified by nociceptive inputs to CE.

Information about a simple CS (e.g., as a tone paired with shock) is directed toward CE (where response execution is initiated) by way of pathways that originate in LA. While LA projects to CE directly, and by way of B and AB, the direct projection from LA to CE seems to be sufficient since lesions of B and AB have no effect on simple fear conditioning to a tone (Killcross, Robbins, & Everitt, 1997). LA and B also project to CE via IC (Paré & Smith, 1993).

The CE projects to brain-stem areas that control the expression of fear responses (LeDoux, Iwata, Cicchetti, & Reis, 1988; Davis, 1992; Kapp, Whalen, Supple, & Pascoe, 1992). It is thus not surprising that damage to CE interferes with the expression of conditioned fear responses (Hitchcock & Davis, 1986; Iwata et al., 1986; Van de Kar, Piechowski, Rittenhouse, & Gray, 1991; Gentile et al., 1986). In contrast, damage to areas that CE projects to selectively interrupts the expression of individual responses. For example, damage to the lateral hypothalamus affects blood pressure but not freezing responses, and damage to the periaqueductal gray interferes with freezing but not blood pressure responses (LeDoux, Iwata, Cicchetti, & Reis, 1988). Similarly, damage to the bed nucleus of the stria terminalis has no effect on either blood pressure or freezing responses (LeDoux, Iwata, Cicchetti, & Reis, 1988) but disrupts the conditioned release of pituitary–adrenal stress hormones (Van de Kar, Piechowski, Rittenhouse, & Gray, 1991). Because CE receives inputs from LA, B, and AB (Pitkanen, Savander, & LeDoux, 1997), it is in a position to mediate the expression of conditioned fear responses elicited by both acoustic and contextual CSs (Fig. 4.3).

Is the Amygdala Necessary?

In spite of a wealth of data implicating the amygdala in fear conditioning, some authors have suggested that the amygdala is not a site of US–CS association or storage during fear conditioning (Cahill & McGaugh, 1998; McGaugh, 2000; McGaugh & Izquierdo, 2000; McGaugh, McIntyre, & Power, 2002; McIntyre, Power, Roozendaal, & McGaugh, 2003). They argue instead that the amygdala modulates memories that are formed elsewhere. It is clear that there are multiple memory systems in the brain (McDonald & White, 1993; Squire, Knowlton, & Musen, 1993; Suzuki & Eichenbaum, 2000; Eichenbaum, 2001) and that the amygdala does indeed modulate memories formed in other systems, such as declarative or explicit memories formed through hippocampal circuits or habit memories formed through striatal circuits (Packard, Cahill, & McGaugh, 1994). However, evidence for a role of the amygdala in modulation should not be confused with evidence against a role in US–CS association. That the amygdala is indeed important for learning is suggested by studies showing that inactivation of the amygdala during learning prevents learning from taking place (Muller, Corodimas, Fridel, & LeDoux, 1997). Further, if the inactivation occurs immediately after training, then there is no effect on subsequent memory (Wilensky, Schafe, & LeDoux, 1999), showing that the effects of pretraining treatment are on learning and not on processes that occur after learning. Thus, in addition to *storing implicit* memories about dangerous situations in its own circuits, the amygdala *modulates* the formation of explicit memories in circuits of the hippocampus and related areas.

THE HUMAN AMYGDALA AND COGNITIVE–EMOTIONAL INTERACTIONS

We now turn to studies on the roles of the human amygdala. Deficits in the perception of the emotional meaning of faces, especially fearful faces, have been found in humans with amygdala damage (Adolphs et al., 1996; Stone et al., 2003). Similar results were reported for detection of the emotional tone of voices (Scott et al., 1997). Further, damage to the amygdala (Bechara et al., 1995) or areas of the temporal lobe including the amygdala (LaBar et al., 1998) produced deficits in fear conditioning in humans. Also, damage to the hippocampus in humans, as in rats, disrupts fear conditioning to contextual cues (Anagnostaras, Gale, & Fanselow, 2001). Functional imaging studies have shown that the amygdala is activated more strongly in the presence of fearful and angry faces than happy ones (Breiter et al., 1996) and that subliminal presentations of such stimuli lead to stronger activations than freely seen stimuli (Whalen et al., 1998). Fear conditioning also leads

to increases in amygdala activity, as measured by fMRI (LaBar et al., 1998; Buchel & Dolan, 2000), and these effects also occur to subliminal stimuli (Morris et al., 1999). Additionally, when the activity of the amygdala during fear conditioning is cross-correlated with the activity in other regions of the brain, the strongest relations are seen with subcortical (thalamic and collicular) rather than cortical areas, further emphasizing the importance of the direct thalamic–amygdala pathway in the human brain (Morris, Ohman, & Dolan, 1999). Work in humans has further implicated the amygdala in social interactions (Hart et al., 2000; Phelps et al., 2000). Other aspects of emotion and the human brain are reviewed elsewhere (Davidson & Irwin, 1999; Critchley, Mathias, & Dolan, 2002; Dolan & Vuilleumier, 2003).

There is growing enthusiasm for the notion that fear-learning processes similar to those occurring in fear-conditioning experiments might indeed be an important factor in certain human anxiety disorders. For example, fear-conditioning models of posttraumatic stress disorder (PTSD) and panic disorder (Goddard & Charney, 1997; Rauch et al., 2000) have been proposed by researchers in these fields.

Earlier in the 20th century, the notion that conditioned fear contributes to phobias and related fear disorders was fairly popular. However, this idea fell out of favor because laboratory fear conditioning seemed to produce easily extinguishable fear, whereas clinical fear is difficult to treat. Fear disorders involve a special kind of learning, called "prepared learning," where the CS is biologically significant rather than neutral (de Silva, Rachman, & Seligman, 1977; Ohman, 1992). While preparedness may indeed contribute, there is another factor to consider. In studies of rats, easily extinguished fear could be converted into difficult to extinguish fear with damage to the medial prefrontal cortex (Morgan, Romanski, & LeDoux, 1993). This suggested that alterations in the organization of the medial prefrontal regions might predispose certain people in some circumstances (e.g., stressful situations) to learn in a way that is difficult to extinguish (treat) under normal circumstances. These changes could come about because of genetic or experiential factors or some combination. Recent imaging studies have shown amygdala alterations in PTSD, panic disorders, and depression (Price, 1999; Davidson, Pizzagalli, Nitschke, & Putnam, 2002; Anand & Shekhar, 2003; Drevets, 2003; Rauch, Shin, & Wright, 2003; Wright et al., 2003).

One of the key issues for the coming years is to integrate research on emotion and cognition. As a step in this direction, we consider how fear processing by the amygdala is influenced by and can influence the perceptual, attentional, and memory functions of the cortex.

The amygdala receives inputs from cortical sensory-processing regions of each sensory modality and projects back to these as well (Amaral, Price, Pitkanen, & Carmichael, 1992; McDonald, 1998). These projections allow

the amygdala to determine whether danger is present in the sensory world, but in addition to processing the significance of external stimuli, the amygdala can influence sensory processing occurring in cortical areas. The amygdala receives inputs only from the late stages of cortical sensory processing, but it projects back to the earliest stages (Amaral, Price, Pitkanen, & Carmichael, 1992). Thus, once the amygdala is activated by a sensory event from the thalamus or cortex, it can begin to regulate the cortical areas that project to it, controlling the kinds of input it receives from the cortex (Fig. 4.4). The amygdala also influences cortical sensory processes indirectly, by way of projections to various "arousal" networks, including the basal forebrain cholinergic system, the brain-stem cholinergic system, and the locus coeruleus noradrenergic system, each of which innervates widespread areas of the

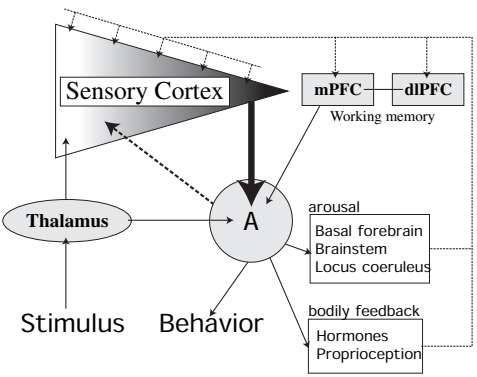


Figure 4.4. The amygdala (A) receives inputs only from the late stages of cortical sensory processing (thick arrow) but projects back to the earliest stages. Once the amygdala is activated by a sensory event from the thalamus or cortex, it can begin to regulate the cortical areas that project to it. The amygdala also influences cortical sensory processes indirectly by way of projections to various "arousal" networks, including the basal forebrain cholinergic system, the brain-stem cholinergic system, and the locus coeruleus noradrenergic system, each of which innervates widespread areas of the cortex. Thus, once the amygdala detects danger, it can activate these arousal systems, which can then influence sensory processing. The bodily responses initiated by the amygdala can also influence cortical areas by way of feedback either from proprioceptive or visceral signals or hormones. The amygdala also interacts with the medial prefrontal cortex (mPFC), which together with the dorsolateral prefrontal cortex (dlPFC) has widespread influences on cognition and behavior and sends connections to several amygdala regions, allowing cognitive functions organized in prefrontal regions, especially working memory, to regulate the amygdala and its fear reactions. Modulatory/regulatory inputs are marked with dashed lines.

cortex (Aston-Jones, Rajkowski, & Cohen, 2000; Kapp, Whalen, Supple, & Pascoe, 1992; Weinberger, 1995; Holland & Gallagher, 1999). Thus, once the amygdala detects danger, it can activate these arousal systems, which can then influence sensory processing. The bodily responses initiated by the amygdala can also influence cortical areas by way of feedback from either proprioceptive or visceral signals or hormones (McGaugh et al., 1995; Damasio, 1994). Amygdala regulation of the cortex by either direct or indirect routes could facilitate the processing of stimuli that signal danger even if such stimuli occur outside of the attentional field (Armony, Quirk, & LeDoux, 1998).

The amygdala also interacts with areas within the medial prefrontal cortex, a structure known to be involved in working memory. These areas have widespread influences on cognition and behavior, but they also send connections to several amygdala regions, including CE, as well as to brainstem outputs of CE, allowing cognitive functions organized in prefrontal regions, especially working memory, to regulate the amygdala and its fear reactions (Fig. 4.4).

The amygdala is a collection of diverse nuclei. It thus should come as no surprise that consequences of damage to this region vary, depending on where the lesion is located (Garcia, Vouimba, Baudry, & Thompson, 1999; Morgan, Schulkin, & LeDoux, 2003; Quirk & Gehlert, 2003; Rosenkranz & Grace, 2003). Some lesions led to a marked exaggeration of fear reactions, while others did not. Overall, this work suggested that the prefrontal cortex and amygdala are reciprocally related. That is, in order for the amygdala to respond to fear, the prefrontal region has to be shut down. By the same logic, when the prefrontal region is active, the amygdala would be inhibited, making it harder to express fear. Pathological fear, then, may occur when the amygdala is unchecked by the prefrontal cortex, and fear therapy may be a process by which we learn to increase activity in the prefrontal region so that the amygdala is less free to express fear. Clearly, decision-making ability in emotional situations is impaired in humans with damage to the medial prefrontal cortex (Damasio, 1994; Bechara, Damasio, & Damasio, 2003), and abnormalities in the prefrontal cortex may predispose people to develop fear and anxiety disorders. These abnormalities that bias one to develop pathological fear could be due to genetic or epigenetic organization of medial prefrontal synapses or to experiences that subtly alter medial prefrontal synaptic connections. Indeed, the behavior of nonhuman animals with abnormalities of the medial prefrontal cortex is reminiscent of humans with anxiety disorders: they develop fear reactions that are difficult to regulate. Objective information about the world may indicate that the situation is not dangerous, but because they cannot properly regulate fear circuits, they experience fear and anxiety in otherwise safe situations.

The medial prefrontal cortex may thus serve as an interface between cognitive and emotional systems, allowing cognitive information processing in the prefrontal cortex to regulate emotional processing by the amygdala. In addition, emotional processing by the amygdala may influence decision making and other cognitive functions of the prefrontal cortex. Consequently, prefrontal–amygdala interactions may be involved in the conscious feelings of fear (see section below).

In humans, damage to the amygdala interferes with implicit emotional memories but not explicit memories about emotions, whereas damage to the medial temporal lobe memory system interferes with explicit memories about emotions but not with implicit emotional memories (Bechara, Damasio, & Damasio, 2003; LaBar, Crupain, Voyvodic, & McCarthy, 2003). While explicit memories with and without emotional content are formed by way of the medial temporal lobe, those with emotional content differ from those without such content. The former tend to be longer-lasting and more vivid (Christianson, 1992; Cahill & McGaugh, 1998). Lesions of the amygdala or systemic administration of a β-adrenergic antagonist prevent this amplifying effect of emotion on declarative memory (Cahill & McGaugh, 1998), suggesting that the amygdala can modulate the storage of explicit memories in cortical areas. At the same time, the medial temporal lobe memory system projects to the amygdala (see above). Retrieval of long-term memories of traumatic events may trigger fear reactions by way of these projections to the amygdala.

WHAT ABOUT FEELINGS?

Consciousness

Our discussion above of the relation between the study of emotion in rats and humans brings us at last to the issue of how the conscious dimension of "feelings" in humans relates to the processing approach that has proved so successful in the study of rats. Subjective emotional experience, like the feeling of being afraid, results when we become consciously aware that an emotion system of the brain, like the defense system, is active. In order for this to occur, we need at least two things: a defense system and the capacity to be consciously aware of its activity. The up side of this line of thought is that once we understand consciousness, we will also understand subjective emotional experiences. Many believe that the down side is that in order to understand subjective emotional experiences, we need to understand consciousness. However, it might be argued that our "divide-and-conquer" ap-

proach to emotion may also prove relevant to the study of consciousness, revealing its manifold nature (Rorty, 1980; Churchland, 1984).

While one could hardly say that there is a general consensus on the nature of consciousness, many of the theories proposed in recent years are built around the concept of working memory. Borrowing a term from computer technology, memory researchers sometimes refer to *temporary storage* mechanisms as buffers. It is now believed that a number of specialized buffers exist. For example, each sensory system has one or more temporary buffers. These aid in perception, allowing the system to compare what it is seeing or hearing now to what it saw or heard a moment ago. There are also temporary buffers associated with aspects of language use (these help keep the first part of a sentence in mind until the last part is heard so that the whole thing can be understood). The specialized memory buffers work in parallel, independently of one another.

Working memory consists of a *workspace*, where information from the specialized buffers can be held temporarily, and a set of *executive* functions that control operations performed on this information. The executive functions take care of the overall coordination of the activities of working memory, such as determining which specialized systems should be attended to at the moment and shuffling information in and out of the workspace from these and other systems. This idea is not fundamentally different from the concept of a "blackboard" in traditional artificial intelligence (Hanson & Riseman, 1978; Erman, Hayes-Roth, Lesser, & Reddy, 1980; Jagannathan, Dodhiawala, & Baum, 1997).

A computer simulation of the weather is not the same thing as rain or sunshine (Johnson-Laird, 1988). Working memory theories, in dealing with consciousness in terms of processes rather than as content, try to explain what kinds of computational function might be responsible for and underlie conscious experiences, but they do not explain what it is like to have those experiences. These theories provide an account of the way human minds work, in a general sense, rather than an account of what a particular experience is like in a particular mind. They can suggest how a representation might be created in working memory but not what it is like to be aware of that representation. They suggest how decision processes in working memory might lead to movement but not what it is like to actually decide to move. In other words, working memory is likely to be the platform on which conscious experience stands; but consciousness, especially its phenomenal or subjective nature, is not completely explained by the computational processes that underlie working memory, at least not in a way that anyone presently comprehends.

Figuring out the exact nature of consciousness and the mechanisms by which it emerges out of collections of neurons is truly an important problem.

Many questions remain to be answered about how the brain mediates working memory and how consciousness relates to the working memory system. However, it is not necessary for emotion researchers to solve these problems, nor is it necessary to wait for the solutions before studying how emotion works. Emotion researchers need to figure out how emotional information is represented in working memory. The rest of the problem, figuring out how the contents of working memory become consciously experienced and how these subjective phenomena emerge from the brain, belongs on the shoulders of all mind scientists.

Amygdala and Consciousness

Emotional arousal influences cognitive processing. Attention, perception, memory, decision making, and the conscious concomitants of each are swayed by emotional states. The reason for this is simple. Emotional arousal organizes and coordinates brain activity (LeDoux, 1996). The emotional coordination of brain activity converts conscious experiences into emotional experiences.

If our immediate conscious content occupies working memory, then a feeling (the conscious experience of an emotion) is the representation in working memory of the various elements of an immediate emotional state. In this view, the feeling of being afraid would be a state of consciousness in which working memory integrates the following disparate kinds of information: (1) an immediately present stimulus (e.g., a snake on the path in front of you); (2) long-term memories about that stimulus (facts you know about snakes and experiences you have had with them); and (3) emotional arousal by the amygdala. The first two are components of any kind of conscious perceptual experience as the only way to identify an immediately present stimulus is by comparing its physical features (the way it looks or sounds) with memories or present knowledge about the same or similar stimuli. However, the third kind of information occurs only during an emotional experience. Amygdala activation, in other words, turns a plain perceptual experience into a fearful one.

The key question, then, is how the amygdala achieves this alteration of consciousness, this transformation of cognition into emotion, or better yet, this takeover of consciousness by emotion. The answer may be that emotion comes to monopolize consciousness, at least in the domain of fear, when the amygdala comes to dominate working memory.

The amygdala can influence working memory in a variety of ways, some of which will be described. The first is by altering sensory processing in cortical areas. Working memory finds out about the outside world from sen-

sory processing areas, so anything that alters how these areas process sensory stimuli will affect what working memory works with. By way of connections with sensory processing areas in the cortex, amygdala arousal can modify sensory processing. While only the latest stages of sensory processing in the cortex send connections to the amygdala, the amygdala sends connections to all stages, allowing the amygdala to influence even very early processing in the neocortex (Amaral, Price, Pitkanen, & Carmichael, 1992). Sensory cortex areas are influenced by activity in the amygdala, as suggested by studies showing that the rate at which cells in the auditory cortex fire to a tone is increased when that tone is paired with a shock in a fear-conditioning situation (Weinberger, 1995). Other studies show that damage to the amygdala prevents some of the cortical changes from taking place (Armony, Quirk, & LeDoux, 1998). Because the sensory cortex provides important inputs to working memory, the amygdala can influence working memory by altering processing there.

The sensory cortex is crucially involved in activation of the medial temporal lobe memory system. By influencing the sensory cortex, the amygdala can have an impact on the long-term memories that are active and available to working memory. However, the amygdala also influences the medial temporal lobe memory system (through the rhinal cortex) and, thus, the memories available to working memory.

The amygdala can also act directly on working memory circuits. Although it does not have direct connections with the lateral prefrontal cortex, it does have connections with other areas of the prefrontal cortex involved in working memory, including the medial (anterior cingulate) and ventral (orbital) prefrontal cortex (Groenewegen, Berendse, Wolters, & Lohman, 1990; McDonald, 1998; Uylings, Groenewegen, & Kolb, 2003). Damage to the medial prefrontal cortex in rats leads to a loss of fear regulation, and studies of monkeys and humans have implicated the medial orbital region in processing emotional cues (rewards and punishments) and in the temporary storage of information about such cues (Everitt & Robbins, 1992; Gaffan, 1992; Rolls, 1998; Rogers et al., 1999). The orbital region is connected with the anterior cingulate, and like the anterior cingulate, it also receives information from the amygdala and hippocampus (Fuster, 1990; Petrides & Pandya, 1999, 2002). Humans with orbital cortex damage become oblivious to social and emotional cues, have poor decision-making abilities, and may exhibit sociopathic behavior (Damasio, 1994). In addition to being connected with the amygdala, the anterior cingulate and orbital areas are intimately connected with one another, as well as with the lateral prefrontal cortex, and each of the prefrontal areas receives information from sensory processing regions and from areas involved in various aspects of implicit and explicit memory processing. The anterior cingulate and orbital areas thus provide a means through which emotional processing

by the amygdala might be related in working memory to immediate sensory information and long-term memories processed in other areas of the cortex (see also Chapter 5, Rolls).

Attention and working memory are closely related, and recent studies have shown that amygdala damage interferes with an important aspect of attention (Anderson & Phelps, 2001, 2002). Normally, if we are attending to one stimulus, we ignore others. This selective attention allows us to focus our thoughts on the task at hand. However, if the second stimulus is emotionally significant, it can override the selection process and slip into working memory. Damage to the amygdala, though, prevents this from occurring. The amygdala, in other words, makes it possible for implicitly processed (unattended) emotional stimuli to make it into working memory and consciousness.

In addition, the amygdala can influence working memory indirectly by way of projections to the various amine cell groups that participate in cortical arousal, including cholinergic, dopaminergic, noradrenergic, and serotonergic systems (see Fig. 4.4; see also Chapter 3, Kelley). These arousal pathways are relatively nonspecific since they influence many cortical areas simultaneously. Specificity comes from the fact that the effects of arousal are most significant on circuits that are active. As a result, if the cortex is focused on some threatening stimulus, the circuits involved will be facilitated by the arousal systems. This will help keep attention focused on the threatening situation.

Finally, once the outputs of the amygdala elicit alarm-related behaviors and accompanying changes in body physiology (fight/flight kinds of response), the brain begins to receive feedback from the bodily responses. Feedback can be in the form of sensory messages from internal organs (visceral sensations) or from the muscles (proprioceptive sensations) or in the form of hormones or peptides released by bodily organs that enter the brain from the bloodstream and influence neural activity. Although the exact manner in which bodily feedback influences working memory is not clear, it is likely that working memory has access to this information in one form or another. The feedback from these responses is relatively slow, on the order of seconds, when compared to the feedback that occurs by way of synaptic transmission within the brain, which transpires within a matter of milliseconds. Bodily feedback adds at least intensity and duration but may also help refine our interpretation of the emotion we are experiencing once the episode has been triggered (James, 1890; Damasio, 1999; Cacioppo, Hawkley, & Bernston, 2003). Bodily feedback in the form of stress hormones can either enhance or impair long-term memory functions of the temporal lobe memory system, which will in turn influence the content of working memory.

In the presence of fear-arousing stimuli, activation of the amygdala will lead working memory to receive a greater number of inputs and inputs of a greater variety than in the presence of emotionally neutral stimuli. These extra inputs may add affective charge to working memory representations and may be what make a particular subjective experience a fearful emotional experience.

What kinds of emotional experience do nonhuman animals without a well-developed prefrontal cortex have? It might be possible to have certain kinds of modality-specific conscious state when the activity of one system dominates the brain (LeDoux, 2002). This might happen with strong sensory stimulation (loud noise or painful stimulus) or in response to emotionally charged stimuli (sight of a predator). Modality-specific feelings can be thought of in terms of passive states of awareness, as opposed to the more flexible kind of conscious awareness, complete with on-line decision-making capacities, made possible by working memory.

Although this theory of emotional experience is based on studies of fear, it is meant as a general-purpose theory that applies to all kinds of emotional experience. The particulars will be different, but the overall scheme (whereby working memory integrates sensory information about the immediately present physical stimulus with memories from past experiences with such stimuli and with the current emotional consequences of those stimuli) will apply to all varieties of emotional experience in humans, from fear to anger to joy to dread and even love and appetitive emotions (Everitt & Robbins, 1992; Gaffan, 1992; Hatfield et al., 1996; Rolls, 1998; LeDoux, 2002; Yang et al., 2002).

WHAT ABOUT POSITIVE EMOTIONS?

Other researchers have studied the role of the amygdala in processing stimuli that predict desirable things (e.g., tasty foods and sexually receptive partners). So what about love? The key issue is whether there is some way to study the function in nonhuman animals that makes sense in terms of human behavior. For fear, we were able to use conditioning because conditioned fear responses are similar in humans and other mammals. The paradigm for behavioral love has been to focus on pair bonding, in the sense of a long-term bond between sexual partners. Application of this paradigm is based on comparison of species in which animals do and do not pair up with one another monogamously (Insel, 1997; Carter, 1998). Only about 3% of mammals are monogamous. Even in nonhuman primates, monogamy is fairly rare; but prairie voles, small rodents living in the Midwestern plains of the

United States, pair up with sexual partners. Once they mate, they stick together and raise their offspring as a family, even across generations. Given that pair bonding is so rare, the monogamous prairie vole offers a possible window into the biology of attachment.

Attachment (pair-bond formation) is a key part of love (Sternberg, 1988; Hedlund & Sternberg, 2000; Bartholomew, Kwong, & Hart, 2001). There can be attachment without love but not love without attachment (Carter, 1998). Perhaps the mechanisms that underlie attachment in voles are also at work in humans. Vole researchers used a different strategy from the one used to study fear. Rather than starting with the circuits and then trying to figure out the chemistry, they started with chemical findings and attempted to relate them to circuits (see also Chapter 3, Kelley).

Two features of prairie voles make them attractive for studying pair bonding (Insel, 1997). The first is that monogamy also occurs in voles living in laboratory settings. In the laboratory, bonding can be measured by putting a vole in the middle chamber of a box with three compartments. In one of these, it encounters its mate and, in the other, a stranger. Voles that have mated spend time with their partner, whereas unbonded ones have no particular preference. The second feature is that pair bonding is present only in prairie voles and not in closely related montane voles, which are found in the Rockies and live individually rather than in family groups. These animals do not form mate preferences after having sex, so when put in the three-chamber box, they do not spend more time with a vole they mated with than a novel one. Differences in the brains of these two kinds of vole might provide important clues about the biology of pair bonding, family organization, and perhaps love itself.

One of the main discoveries was that receptors for two hormones believed to play an important role in reproductive behavior were located in different circuits in prairie and montane voles (Insel, 1997): *vasopressin* and *oxytocin*. They are found only in mammals and are related to ancestral hormones that play a key role in behaviors like nest building in nonmammalian species. In the mammalian brain, these chemicals function not just as hormones but also as neurotransmitters and/or modulators.

The roles of these chemicals in the behavioral differences between the voles have been determined by injecting drugs that either stimulate or inhibit the action of vasopressin or oxytocin. The drugs have been injected into the ventricles, cavities that contain cerebrospinal fluid, which flows from the ventricles into the spaces surrounding neurons and, therefore, reach widespread areas of the brain. When a drug that blocks the action of naturally released oxytocin is put in the ventricles of a female prairie vole just before mating, she mates but does not bond with the sex partner. The drug disrupts attachment, not sex. This suggests that oxytocin released during

mating underlies bond formation in females. Similarly, if a drug that blocks vasopressin is put in the ventricles of a male prairie vole before mating, the male mates but does not bond. The drug blocks attachment but not sexual responses. Thus, blocking oxytocin in female and vasopressin in male prairie voles makes them act like montane voles. Oxytocin affects bonding only in female brains and vasopressin in male brains.

Oxytocin and vasopressin are also present in the brains of humans and are released during sexual behavior. These hormones have not yet been proven to underlie attachment in humans. Regardless of whether the vole findings on oxytocin and vasopressin end up being completely applicable to the human brain, this work illustrates important principles that will surely guide research for some time.

Areas of the amygdala are included in both the fear and sex circuits. However, the circuits are otherwise quite different. Even within the amygdala different areas are involved in sex (medial and posterior nuclei) and fear (lateral and central nuclei). This emphasizes the importance of mapping the circuit for different kinds of emotional system rather than assuming that there is a universal circuitry for all emotions. At the same time, different emotion circuits, like the fear and sex circuits, sometimes interact with one another. For example, the medial nucleus sends connections to the central nucleus (Canteras, Simerly, & Swanson, 1995), where oxytocin receptors are present (Veinante & Freund-Mercier, 1997). This may be related to the ability of both oxytocin and positive social interactions to reduce fear and stress.

Pair bonding in animals has given researchers a behavioral paradigm for studying a phenomenon akin to love without analyzing subjectivity, but what about the feelings of love? Although there is little research to draw upon at this point, we can use our more detailed understanding of cognitive–emotional interactions in fear to speculate about how our brain feels love. Suppose you unexpectedly see a person you care about and feel the love you have for that person. Let us follow the flow of information from the visual system through the brain to the point of the feeling of love as best we can. First, the stimulus will flow from the visual system to the prefrontal cortex (putting an image of the loved one in working memory). The stimulus also reaches the explicit memory system of the temporal lobe and activates memories about that person. Working memory then retrieves relevant memories and integrates them with the image of the person. Simultaneous with these processes, the subcortical areas presumed to be involved in attachment will be activated. Activation of attachment circuits then impacts on working memory in several ways. One involves direct connections from the attachment areas to the prefrontal cortex (as with fear, it is the medial prefrontal region that is connected with subcortical attachment areas). Activation of attachment circuits also leads to activation of brain-stem arousal

networks that participate in the focusing of attention on the loved one by working memory. Bodily responses will also be initiated as outputs of attachment circuits. These responses contrast with the alarm responses initiated by fear and stress circuits. We approach rather than try to escape from or avoid the person, and these behavioral differences are accompanied by different physiological conditions within the body (James, 1890; Damasio, 1999). This pattern of inputs to working memory from within the brain and from the body biases us more toward an open and accepting mode of processing than toward tension and vigilance (Porges, 1998). The net result in working memory is the feeling of love. This scenario is certainly incomplete, but it shows how we can build upon research on one emotion to generate hypotheses about others.

CONCLUSION

This chapter has demonstrated the ways in which a focus on the study of fear mechanisms, especially the mechanisms underlying fear conditioning, can enrich our understanding of the emotional brain (LeDoux, 1996). This work has mapped out pathways involved in fear learning in both experimental animals and humans and has begun to shed light on interactions between emotional and cognitive processes in the brain. While the focus on fear conditioning has its limits, it has proven valuable as a research strategy and provides a foundation upon which to build a broader understanding of the mind and brain.

At the same time, there is a disturbing rush to embrace the amygdala as the new center of the emotional brain. It seems unlikely that the amygdala is the answer to how all emotions work, and it may not even explain how all aspects of fear work. There is some evidence that the amygdala participates in positive emotional behaviors, but that role is still poorly understood.

Understanding fear from the neuroscience point of view is just one of many ways of understanding emotions in general. Other disciplines can undoubtedly help. The past few decades have seen the emergence of interdisciplinary work in computational modeling and neuroscience (Arbib, 2003). The use of computational modeling techniques has proved essential in understanding experimentally intractable phenomena such as complex intracellular signaling pathways involving dozen of simultaneously interacting chemical species or the way large networks of tens of thousands of neurons process information (Bialek et al., 1991, 2001; Dayan & Abbott, 2003). Conversely, neural computation has provided inspiration to many engineers and computer scientists in fields ranging from pattern recognition to machine learning (Barto & Sutton, 1997). The topic of emotion is still on the side-

lines but not for long, as this book attests (Fellous, Armony, & LeDoux, 2003). As we have discussed above, it may be fruitful for computational models to approach the problem of emotion by considering one emotion at a time and to focus on how the emotion is operationalized without losing the "big picture" of how feelings might emerge.

This approach has led to the discovery of basic principles that may apply to other emotions as well as fear:

- Emotions involve primitive circuits. These primitive circuits are basic, robust processing units that are conserved across evolution.
- In some circumstances, cognitive (i.e., nonemotional) circuits can function independently from emotions.
- Emotional memories are somewhat different from other kinds of memory. They may last longer and be more vivid (reassociate rigidly and effectively with other memory items). Some types of nonemotional memory (e.g., working memory) help extinguish emotional memory (e.g., fear).
- There are two parallel routes of emotional processing of a stimulus. One is fast (thalamic–amygdala pathway); the other is slower (cortical–amygdala pathway) and presumably modulates the fast route. (Compare the dual routes analyzed in Chapter 5, Rolls.)
- There are two physically separate inputs to an emotional (evaluation) system. The first is reserved for simple stimuli such as a tone (LA→CE in the fear circuit); the second is reserved for more complex stimuli, such as context, and includes more processing stages (hippocampus→B/AB→CE in the fear circuit).
- Emotional expressions are triggered by a central signal (CE activation), but the specifics of the expressions are determined locally (lateral hypothalamus, blood pressure; periaqueductal gray, freezing; bed nucleus, stress hormones, etc., in the fear circuit), according to the current state of the animal (current heart rate, environmental conditions, actual levels of hormones).

These basic principles might serve as a starting point in the design of computational models of emotions.

The future of emotion research will be bright if we keep in mind the importance of focusing on a physiologically well-defined aspect of emotion, using an experimental approach that simplifies the problem in such a way as to make it tractable, circumventing vague and poorly defined aspects of emotion, and removing subjective experience as a roadblock to experimentation. This is not to suggest that the problems of feelings should not be explored, but, instead, that they should be explored in a way that builds on a firm understanding of the neural mechanisms that subserve the underlying behaviors.

Note Portions of this chapter appeared in somewhat different form in LeDoux (1996, 2000, 2002).

References

Adolphs, R., Damasio, H., Tranel, D., & Damasio, A. R. (1996). Cortical systems for the recognition of emotion in facial expressions. *Journal of Neuroscience*, *16*(23), 7678–7687.

Aggleton, J. P., & Mishkin, M. (1986). The amygdala: Sensory gateway to the emotions. In R. Plutchick & H. Kellerman (Eds.), *Biological foundations of emotion* (pp. 281–300). New York: Academic Press.

Amaral, D. G., Price, J. L., Pitkanen, A., & Carmichael, T. S. (1992). Anatomical organization of the primate amygdaloid complex. In J. P. Aggleton (Ed.), *The amygdala* (pp. 1–66). New York: Wiley-Liss.

Anagnostaras, S. G., Gale, G. D., & Fanselow, M. S. (2001). Hippocampus and contextual fear conditioning: Recent controversies and advances. *Hippocampus*, *11*, 8–17.

Anand, A., & Shekhar, A. (2003). Brain imaging studies in mood and anxiety disorders: Special emphasis on the amygdala. *Annals of the New York Academy of Sciences*, *985*, 370–388.

Anderson, A. K., & Phelps, E. A. (2001). Lesions of the human amygdala impair enhanced perception of emotionally salient events. *Nature*, *411*, 305–309.

Anderson, A. K., & Phelps, E. A. (2002). Is the human amygdala critical for the subjective experience of emotion? Evidence of intact dispositional affect in patients with amygdala lesions. *Journal of Cognitive Neuroscience*, *14*, 709–720.

Arbib, M. A. (2003). *The handbook of brain theory and neural networks* (2nd ed.). Cambridge, MA: MIT Press.

Armony, J. L., & LeDoux, J. E. (1997). How the brain processes emotional information. *Annals of the New York Academy of Sciences*, *821*, 259–270.

Armony, J. L., Quirk, G. J., & LeDoux, J. E. (1998). Differential effects of amygdala lesions on early and late plastic components of auditory cortex spike trains during fear conditioning. *Journal of Neuroscience*, *18*, 2592–2601.

Arnold, M. (1960). *Emotions and personality*. New York: Columbia University Press.

Aston-Jones, G., Rajkowski, J., & Cohen, J. (2000). Locus coeruleus and regulation of behavioral flexibility and attention. *Progress in Brain Research*, *126*, 165–182.

Bartholomew, K., Kwong, M. J., & Hart, S. D. (2001). *Attachment*. New York: Guilford.

Barto, A. G., & Sutton, R. S. (1997). *Reinforcement learning in artificial intelligence*. Amsterdam: North-Holland/Elsevier.

Bechara, A., Damasio, H., & Damasio, A. R. (2003). Role of the amygdala in decision-making. *Annals of the New York Academy of Sciences*, *985*, 356–369.

Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C., & Damasio, A. R. (1995). Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science*, *269*, 1115–1118.

Bernard, J. F., & Besson, J. M. (1990). The spino(trigemino)pontoamygdaloid pathway: Electrophysiological evidence for an involvement in pain processes. *Journal of Neurophysiology, 63*, 473–490.

Bialek, W., Nemenman, I., & Tishby, N. (2001). *Predictability, complexity, and learning*. Cambridge, MA: MIT Press.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R., & Warland, D. (1991). Reading a neural code. *Science, 252*(5014), 1854–1857.

Blanchard, R. J., Blanchard, D. C., & Fial, R. A. (1970). Hippocampal lesions in rats and their effect on activity, avoidance, and aggression. *Journal of Comparative and Physiological Psychology, 71*, 92–101.

Breiter, H. C., Etcoff, N. L., Whalen, P. J., Kennedy, W. A., Rauch, S. L., Buckner, R. L., Strauss, M. M., Hyman, S. E., & Rosen, B. R. (1996). Response and habituation of the human amygdala during visual processing of facial expression. *Neuron, 17*, 875–887.

Buchel, C., & Dolan, R. J. (2000). Classical fear conditioning in functional neuroimaging. *Current Opinion in Neurobiology, 10*, 219–223.

Burstein, R., & Potrebic, S. (1993). Retrograde labeling of neurons in the spinal cord that project directly to the amygdala or the orbital cortex in the rat. *Journal of Comparative Neurology, 335*, 469–485.

Cacioppo, J. T., Hawkley, L. C., & Bernston, G. G. (2003). The anatomy of loneliness. Oxford: Blackwell.

Cahill, L., & McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. *Trends in Neurosciences, 21*, 294–299.

Cannon, W. B. (1987). The James-Lange theory of emotions: A critical examination and an alternative theory. *American Journal of Psychology, 100*, 567–586. (Original work published 1927)

Canteras, N. S., Simerly, R .B., & Swanson, L. W. (1995). Organization of projections from the medial nucleus of the amygdala: A PHAL study in the rat. *Journal of Comparative Neurology, 360*, 213–245.

Canteras, N. S., & Swanson, L. W. (1992). Projections of the ventral subiculum to the amygdala, septum, and hypothalamus: A PHAL anterograde tract-tracing study in the rat. *Journal of Comparative Neurology, 324*, 180–194.

Carter, C. S. (1998). Neuroendocrine perspectives on social attachment and love. *Psychoneuroendocrinology, 23*, 779–818.

Christianson, S. A. (1992). Remembering emotional events: Potential mechanisms. In S. A. Christianson (Ed.), *Handbook of emotion and memory: Research and theory*. Hilldale, NJ: Erlbaum.

Churchland, P. (1984). *Matter and consciousness*. Cambridge, MA: MIT Press.

Critchley, H. D., Mathias, C. J., & Dolan, R. J. (2002). Fear conditioning in humans: The influence of awareness and autonomic arousal on functional neuroanatomy. *Neuron, 33*, 653–663.

Damasio, A. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York: Harcourt Brace.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason and the human brain*. New York: Putnam.

Davidson, R. J., & Irwin, W. (1999). The functional neuroanatomy of emotion and affective style. *Trends in Cognitive Science*, *3*, 11–21.

Davidson, R. J., Pizzagalli, D., Nitschke, J. B., & Putnam, K. (2002). Depression: Perspectives from affective neuroscience. *Annual Review of Psychology*, *53*, 545–574.

Davis, M. (1992). The role of the amygdala in fear and anxiety. *Annual Review of Neuroscience*, *15*, 353–375.

Dayan, P., & Abbott, L. F. (2003). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

de Silva, P., Rachman, S., & Seligman, M. E. (1977). Prepared phobias and obsessions: Therapeutic outcome. *Behaviour Research and Therapy*, *15*, 65–77.

Dolan, R. J., & Vuilleumier, P. (2003). Amygdala automaticity in emotional processing. *Annals of the New York Academy of Sciences*, *985*, 348–355.

Drevets, W. C. (2003). Neuroimaging abnormalities in the amygdala in mood disorders. *Annals of the New York Academy of Sciences*, *985*, 420–444.

Eichenbaum, H. (2001). The hippocampus and declarative memory: Cognitive mechanisms and neural codes. *Behavioural Brain Research*, *127*, 199–207.

Ekman, P., & Davidson, R. (1994). The nature of emotion: Fundamental questions. New York: Oxford University Press.

Ellsworth, P. (1991). Some implications of cognitive appraisal theories of emotion. In K. T. Strongman (Ed.), *International review of studies on emotions* (pp. 143–161). New York: Wiley.

Erman, L. D., Hayes-Roth, F., Lesser, V. R., & Reddy, D. R. (1980). The HEARSAY II speech understanding system: Integrating knowledge to resolve uncertainty. *Computing Surveys*, *12*, 213–253.

Everitt, B. J., & Robbins, T. W. (1992). Amygdala–ventral striatal interactions and reward-related processes. In J. P. Aggleton (Ed.), *The amygdala: Neurobiological aspects of emotion, memory, and mental dysfunction* (pp. 401–429). New York: Wiley-Liss.

Fanselow, M. S., & Gale, G. D. (2003). The amygdala, fear, and memory. *Annals of the New York Academy of Sciences*, *985*, 125–134.

Fanselow, M. S., & LeDoux, J. E. (1999). Why we think plasticity underlying pavlovian fear conditioning occurs in the basolateral amygdala. *Neuron*, *23*, 229–232.

Fellous, J.-M., Armony, J., & LeDoux, J. E. (2003). Emotional circuits. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (2nd ed., pp. 398–401). Cambridge, MA: MIT Press.

Frankland, P. W., Cestari, V., Filipkowski, R. K., McDonald, R. J., & Silva, A. J. (1998). The dorsal hippocampus is essential for context discrimination but not for contextual conditioning. *Behavioral Neuroscience*, *112*, 863–874.

Frijda, N. (1986). *The emotions*. Cambridge: Cambridge University Press.

Frijda, N. (1993). The place of appraisal in emotion. *Cognition and Emotion*, *7*, 357–387.

Fuster, J. M. (1990). Behavioral electrophysiology of the prefrontal cortex of the primate. In H. B. M. Uylings, C. G. Van Eden, J. P. C. De Bruin, M. A. Corner, & M. G. P. Feenstra (Eds.), *The Prefrontal Cortex: Its Structure, Function and Pathology* (pp. 313–324). Amsterdam: Elsevier.

Gaffan, D. (1992). Amygdala and the memory of reward. In J. P. Aggleton (Ed.), *The amygdala: Neurobiological aspects of emotion, memory, and mental dysfunction* (pp. 471–483). New York: Wiley-Liss.

Garcia, R., Vouimba, R. M., Baudry, M., & Thompson, R. F. (1999). The amygdala modulates prefrontal cortex activity relative to conditioned fear. *Nature*, *402*, 294–296.

Gardner, H. (1987). *The mind's new science: A history of the cognitive revolution*. New York: Basic Books.

Gentile, C. G., Jarrell, T. W., Teich, A., McCabe, P. M., & Schneiderman, N. (1986). The role of amygdaloid central nucleus in the retention of differential pavlovian conditioning of bradycardia in rabbits. *Behavioural Brain Research*, *20*(3), 263–273.

Glascher, J., & Adolphs, R. (2003). Processing of the arousal of subliminal and supraliminal emotional stimuli by the human amygdala. *Journal of Neuroscience*, *23*, 10274–10282.

Goddard, A. W., & Charney D. S. (1997). Toward an integrated neurobiology of panic disorder. *Journal of Clinical Psychiatry*, *58*(*Suppl. 2*), 4–12.

Gray, J. A. (1982). *The neuropsychology of anxiety*. New York: Oxford University Press.

Groenewegen, H. J., Berendse, H. W., Wolters, J. G., & Lohman, A. H. (1990). The anatomical relationship of the prefrontal cortex with the striatopallidal system, the thalamus and the amygdala: Evidence for a parallel organization. *Progress in Brain Research*, *85*, 95–118.

Hanson, A. R., & Riseman, E. M. (1978). VISIONS: A computer system for interpreting scenes. In A. R. Hanson & E. M. Riseman (Eds.), *Computer vision systems* (pp. 129–163). New York: Academic Press.

Hart, A. J., Whalen, P. J., Shin, L. M., McInerney, S. C., Fischer, H., & Rauch, S. L. (2000). Differential response in the human amygdala to racial outgroup vs ingroup face stimuli. *Neuroreport*, *11*, 2351–2355.

Hatfield, T., Han, J. S., Conley, M., Gallagher, M., & Holland, P. (1996). Neurotoxic lesions of basolateral, but not central, amygdala interfere with pavlovian second-order conditioning and reinforcer devaluation effects. *Journal of Neuroscience*, *16*, 5256–5265.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Hedlund, J., & Sternberg, R. J. (2000). *Too many intelligences? Integrating social, emotional, and practical intelligence*. San Francisco: Jossey-Bass.

Hitchcock, I., & Davis, M. (1986). Lesions of the amygdala, but not of the cerebellum or red nucleus, block conditioned fear as measured with the potentiated startle paradigm. *Behavioral Neuroscience*, *100*(1), 11–22.

Holland, P. C., & Gallagher, M. (1999). Amygdala circuitry in attentional and representational processes. *Trends in Cognitive Science*, *3*, 65–73.

Insel, T. R. (1997). A neurobiological basis of social attachment. *American Journal of Psychiatry*, *154*, 726–735.

Isaacson, R. L. (1982). *The limbic system* (2nd ed.). New York: Plenum.

Iwata, J., LeDoux, J. E., Meeley, M. P., Arneric, S., & Reis, D. J. (1986). Intrinsic neurons in the amygdaloid field projected to by the medial geniculate body

mediate emotional responses conditioned to acoustic stimuli. *Brain Research, 383*(1–2), 195–214.

Jagannathan, V., Dodhiawala, R., & Baum, L. S. (1997). *Blackboard architectures and applications. Perspectives in artificial intelligence* (Vol. 3). San Diego: Academic Press.

James, W. (1890). *Principles of psychology*. New York: Holt.

Jarrell, T. W., Gentile, C. G., Romanski, L. M., McCabe, P. M., & Schneiderman, N. (1987). Involvement of cortical and thalamic auditory regions in retention of differential bradycardiac conditioning to acoustic conditioned stimuli in rabbits. *Brain Research, 412*, 285–294.

Johnson-Laird, P. N. (1988). *The computer and the mind*. Cambridge: Harvard University Press.

Kapp, B. S., Whalen, P. J., Supple, W. F., & Pascoe, J. P. (1992). Amygdaloid contributions to conditioned arousal and sensory information processing. In J. P. Aggleton (Ed.), *The amygdala: Neurobiological aspects of emotion, memory, and mental dysfunction* (pp. 229–254). New York: Wiley-Liss.

Kihlstrom, J. F. (1987). The cognitive unconscious. *Science, 237*, 1445–1452.

Killcross, S., Robbins, T. W., & Everitt, B. J. (1997). Different types of fear-conditioned behaviour mediated by separate nuclei within amygdala. *Nature, 388*, 377–380.

Kim, J. J., & Fanselow, M. S. (1992). Modality-specific retrograde amnesia of fear. *Science, 256*, 675–677.

Kotter, R., & Meyer, N. (1992). The limbic system: A review of its empirical foundation. *Behavioural Brain Research, 52*, 105–127.

Krettek, J. E., & Price, J. L. (1978). A description of the amygdaloid complex in the rat and cat with observations on intra-amygdaloid axonal connections. *Journal of Comparative Neurology, 178*, 255–280.

LaBar, K. S., Crupain, M. J., Voyvodic, J. T., & McCarthy, G. (2003). Dynamic perception of facial affect and identity in the human brain. *Cerebral Cortex, 13*, 1023–1033.

LaBar, K. S., Gatenby, J. C., Gore, J. C., LeDoux, J. E., & Phelps, E. A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: A mixed-trial fMRI study. *Neuron, 20*, 937–945.

Lazarus, R. S. (1991). Cognition and motivation in emotion. *American Psychologist, 46*(4), 352–367.

LeDoux, J. (1996). *The emotional brain*. New York: Simon & Schuster.

LeDoux, J. E. (1987). Emotion. In F. Plum (Ed.), *The nervous system* (Vol. V, pp. 419–460). Bethesda: American Physiological Society.

LeDoux, J. E. (1991). Emotion and the limbic system concept. *Concepts in Neuroscience, 2*, 169–199.

LeDoux, J. E. (1992). Brain mechanisms of emotion and emotional learning. *Current Opinion in Neurobiology, 2*, 191–197.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience, 23*, 155–184.

LeDoux, J. E. (2002). *Synaptic self: How our brains become who we are*. Harmondsworth, UK: Penguin.

LeDoux, J. E., Cicchetti, P., Xagoraris, A., & Romanski, L. R. (1990). The lateral amygdaloid nucleus: Sensory interface of the amygdala in fear conditioning. *Journal of Neuroscience*, *10*, 1062–1069.

LeDoux, J. E., Farb, C., & Ruggiero, D. A. (1990). Topographic organization of neurons in the acoustic thalamus that project to the amygdala. *Journal of Neuroscience*, *10*, 1043–1054.

LeDoux, J. E., Iwata, J., Cicchetti, P., & Reis, D. J. (1988). Different projections of the central amygdaloid nucleus mediate autonomic and behavioral correlates of conditioned fear. *Journal of Neuroscience*, *8*, 2517–2529.

LeDoux, J. E., Ruggiero, D. A., Forest, R., Stornetta, R., & Reis, D. J. (1987). Topographic organization of convergent projections to the thalamus from the inferior colliculus and spinal cord in the rat. *Journal of Comparative Neurology*, *264*, 123–146.

Livingston, K. E., & Escobar, A. (1971). Anatomical bias of the limbic system concept. *Archives of Neurology*, *24*, 17–21.

MacLean, P. D. (1949). Psychosomatic disease and the "visceral brain" (recent development bearing on the papez theory of emotion). *Psychosomatic Medicine*, *11*, 338–353.

MacLean, P. D. (1952). Some psychiatric implications of physiological studies on frontotemporal portion of the limbic system (visceral brain). *Electroencephalography and Clinical Neurophysiology*, *4*, 407–418.

Mandler, G. (1984). *Mind and body*. New York: Wiley.

Maren, S., Aharonov, G., & Fanselow, M.S. (1997). Neurotoxic lesions of the dorsal hippocampus and pavlovian fear conditioning in rats. *Behavioural Brain Research*, *88*, 261–274.

Maren, S., & Holt, W. (2000). The hippocampus and contextual memory retrieval in pavlovian conditioning. *Behavioural Brain Research*, *110*, 97–108.

Mascagni, F., McDonald, A. J., & Coleman, J. R. (1993). Corticoamygdaloid and corticocortical projections of the rat temporal cortex: A *Phaseolus vulgaris* leucoagglutinin study. *Neuroscience*, *57*, 697–715.

McDonald, A. J. (1998). Cortical pathways to the mammalian amygdala. *Progress in Neurobiology*, *55*, 257–332.

McDonald, R. J., & White, N. M. (1993). A triple dissociation of memory systems: Hippocampus, amygdala, and dorsal striatum. *Behavioral Neuroscience*, *107*, 3–22.

McGaugh, J. L. (2000). Memory—a century of consolidation. *Science*, *287*, 248–251.

McGaugh, J. L., & Izquierdo, I. (2000). The contribution of pharmacology to research on the mechanisms of memory formation. *Trends in Pharmacological Sciences*, *21*, 208–210.

McGaugh, J. L., McIntyre, C. K., & Power, A. E. (2002). Amygdala modulation of memory consolidation: Interaction with other brain systems. *Neurobiology of Learning and Memory*, *78*, 539–552.

McGaugh, J. L., Mesches, M. H., Cahill, L., Parent, M. B., Coleman-Mesches, K., &

Salinas, J. A. (1995). Involvement of the amygdala in the regulation of memory storage. In J. L. McGaugh, F. Bermudez-Rattoni, & R. A. Prado-Alcala (Eds.), *Plascitity in the central nervous system* (pp. 18–39). Hillsdale, NJ: Erlbaum.

McIntyre, C. K., Power, A. E., Roozendaal, B., & McGaugh, J. L. (2003). Role of the basolateral amygdala in memory consolidation. *Annals of the New York Academy of Sciences, 985*, 273–293.

Miller, G. A., Galanter, E., & Pribam, K. H. (1960). *Plans and the structure of behavior*. New York: Holt.

Morgan, M. A., Romanski, L. M., & LeDoux, J. E. (1993). Extinction of emotional learning: Contribution of medial prefrontal cortex. *Neuroscience Letters, 163*, 109–113.

Morgan, M. A., Schulkin, J., & LeDoux, J. E. (2003). Ventral medial prefrontal cortex and emotional perseveration: The memory for prior extinction training. *Behavioural Brain Research, 146*, 121–130.

Morris, J. S., Ohman, A., & Dolan, R. J (1999). A subcortical pathway to the right amygdala mediating "unseen" fear. *Proceedings of the National Academy of Sciences of the USA, 96*, 1680–1685.

Muller, J., Corodimas, K. P., Fridel, Z., & LeDoux, J. E. (1997). Functional inactivation of the lateral and basal nuclei of the amygdala by muscimol infusion prevents fear conditioning to an explicit conditioned stimulus and to contextual stimuli. *Behavioral Neuroscience, 111*, 683–691.

Nauta, W. J. H., & Karten, H. J. (1970). A general profile of the vertebrate brain, with sidelights on the ancestry of cerebral cortex. In F. O. Schmitt (Ed.), *Neurosciences: Second study program* (pp. 7–26). New York: Rockefeller University Press.

Neisser, U. (1967). *Cognitive psychology*. Englewood Cliffs, NJ: Prentice Hall.

Ohman, A. (1992). Fear and anxiety as emotional phenomena: Clinical, phenomenological, evolutionary perspectives, and information-processing mechanisms. In M. Lewis & J. M. Haviland (Eds.), *Handbook of emotions* (pp. 511–536). New York: Guilford.

Olds, J. (1977). *Drives and reinforcements: Behavioral studies of hypothalamic functions*. New York: Raven.

Ono, T., & Nishijo, H. (1992). Neurophysiological basis of the Kluver-Bucy syndrome: Responses of monkey amygdaloid neurons to biologically significant objects. In J. P. Aggleton (Ed.), *The amygdala: Neurobiological aspects of emotion, memory, and mental dysfunction* (pp. 167–190). New York: Wiley-Liss.

Packard, M. G., Cahill, L., & McGaugh, J. L. (1994). Amygdala modulation of hippocampal-dependent and caudate nucleus-dependent memory processes. *Proceedings of the National Academy of the Sciences of the USA, 91*, 8477–8481.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. New York: Oxford University Press.

Papez, J. W. (1937). A proposed mechanism of emotion. *Archive of Neurology and Psychiatry, 38*, 725–744.

Paré, D., Royer, S., Smith, Y., & Lang, E. J. (2003). Contextual inhibitory gating of impulse traffic in the intra-amygdaloid network. *Annals of the New York Academy of Sciences, 985*, 78–91.

Paré, D., & Smith, Y. (1993). The intercalated cell masses project to the central and medial nuclei of the amygdala in cats. *Neuroscience*, *57*, 1077–1090.

Paré, D., Smith, Y., & Paré, J. F. (1995). Intra-amygdaloid projections of the basolateral and basomedial nuclei in the cat: *Phaseolus vulgaris*-leucoagglutinin anterograde tracing at the light and electron microscopic level. *Neuroscience*, *69*, 567–583.

Pavlov, I. P. (1927). *Conditioned reflexes*. New York: Dover.

Petrides, M., & Pandya, D. N. (1999). Dorsolateral prefrontal cortex: Comparative cytoarchitectonic analysis in the human and the macaque brain and corticocortical connection patterns. *European Journal of Neuroscience*, *11*, 1011–1036.

Petrides, M., & Pandya, D. N. (2002). Comparative cytoarchitectonic analysis of the human and the macaque ventrolateral prefrontal cortex and corticocortical connection patterns in the monkey. *European Journal of Neuroscience*, *16*, 291–310.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience*, *12*, 729–738.

Phillips, R. G., & LeDoux, J. E. (1992). Differential contribution of the amygdala and hippocampus to cued and contextual fear conditioning. *Behavioral Neuroscience*, *106*, 274–285.

Pitkanen, A., Savander, V., & LeDoux, J. E. (1997). Organization of intra-amygdaloid circuitries in the rat: An emerging framework for understanding functions of the amygdala. *Trends in Neurosciences*, *20*, 517–523.

Porges, S. W. (1998). Love: An emergent property of the mammalian autonomic nervous system. *Psychoneuroendocrinology*, 23, 837–861.

Price, J. L. (1999). Prefrontal cortical networks related to visceral function and mood. *Annals of the New York Academy of Sciences*, *877*, 383–396.

Quirk, G. J., Armony, J. L., & LeDoux, J. E. (1997). Fear conditioning enhances different temporal components of tone-evoked spike trains in auditory cortex and lateral amygdala. *Neuron*, *19*, 613–624.

Quirk, G. J., & Gehlert, D. R. (2003). Inhibition of the amygdala: Key to pathological states? *Annals of the New York Academy of Sciences*, *985*, 263–272.

Rauch, S. L., Shin, L. M., & Wright, C. I. (2003). Neuroimaging studies of amygdala function in anxiety disorders. *Annals of the New York Academy of Sciences*, *985*, 389–410.

Rauch, S. L., Whalen, P. J., Shin, L. M., McInerney, S. C., Macklin, M. L., Lasko, N. B., Orr, S. P., & Pitman, R. K. (2000). Exaggerated amygdala response to masked facial stimuli in posttraumatic stress disorder: A functional MRI study. *Biological Psychiatry*, *47*, 769–776.

Rogers, R. D., Owen, A. M., Middleton, H. C., Williams, E. J., Pickard, J. D., Sahakian, B. J., & Robbins, T. W. (1999). Choosing between small, likely rewards and large, unlikely rewards activates inferior and orbital prefrontal cortex. *Journal of Neuroscience*, *19*, 9029–9038.

Rolls, E. T. (1998). *The brain and emotion*. Oxford: Oxford University Press.

Romanski, L. M., & LeDoux, J. E. (1992). Equipotentiality of thalamo-amygdala

and thalamo-cortico-amygdala circuits in auditory fear conditioning. *Journal of Neuroscience*, *12*, 4501–4509.

Romanski, L. M., & LeDoux, J. E. (1993). Information cascade from primary auditory cortex to the amygdala: Corticocortical and corticoamygdaloid projections of temporal cortex in the rat. *Cerebral Cortex*, *3*, 515–532.

Rorty, A. O. (1980). *Explaining emotions*. Berkeley: Unversity of California Press.

Rosenkranz, J. A., & Grace, A. A. (2003). Affective conditioning in the basolateral amygdala of anesthetized rats is modulated by dopamine and prefrontal cortical inputs. *Annals of the New York Academy of Sciences*, *985*, 488–491.

Schacter, D. L., & Singer E. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Reviews*, *69*, 379–399.

Scherer, K. R. (1993). Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition and Emotion*, *7*, 325–355.

Scott, S. K., Young, A. W., Calder, A. J., Hellawell, D. J., Aggleton, J. P., & Johnson, M. (1997). Impaired auditory recognition of fear and anger following bilateral amygdala lesions. *Nature*, *385*, 254–257.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurochemistry*, *20*, 11–21.

Shi, C., & Davis, M. (1999). Pain pathways involved in fear conditioning measured with fear-potentiated startle: Lesion studies. *Journal of Neuroscience*, *19*, 420–430.

Siegel, A., Roeling, T. A., Gregg, T. R., & Kruk, M. R. (1999). Neuropharmacology of brain-stimulation-evoked aggression. *Neuroscience and Biobehavioral Reviews*, *23*, 359–389.

Simon, H. A. (1967). Motivational and emotional controls of cognition. *Psychological Review*, *74*(1), 29–39.

Squire, L. R., Knowlton, B., & Musen, G. (1993). The structure and organization of memory. *Annual Review of Psychology*, *44*, 453–495.

Sternberg, R. J. (Ed.) (1988). *The psychology of love*. New Haven, CT: Yale University Press.

Stone, V. E., Baron-Cohen, S., Calder, A., Keane, J., & Young, A. (2003). Acquired theory of mind impairments in individuals with bilateral amygdala lesions. *Neuropsychologia*, *41*(2), 209–220.

Suzuki, W. A., & Eichenbaum, H. (2000). The neurophysiology of memory. *Annals of the New York Academy of Sciences*, *911*, 175–191.

Swanson, L. W. (1983). The hippocampus and the concept of the limbic system. In W. Seifert (Ed.), *Neurobiology of the hippocampus* (pp. 3–19). New York: Academic Press.

Turner, B. H., & Zimmer, J. (1984). The architecture and some of the interconnections of the rat's amygdala and lateral periallocortex. *Journal of Comparative Neurology*, *227*, 540–557.

Uylings, H. B., Groenewegen, H. J., & Kolb, B. (2003). Do rats have a prefrontal cortex? *Behavioural Brain Research*, *146*, 3–17.

Van de Kar, L. D., Piechowski, R. A., Rittenhouse, P. A., & Gray, T. S. (1991). Amygdaloid lesions: Differential effect on conditioned stress and immobilization-

induced increases in corticosterone and renin secretion. *Neuroendocrinology*, *54*, 89–95.

Veinante, P., & Freund-Mercier, M.J. (1997). Distribution of oxytocin- and vasopressin-binding sites in the rat extended amygdala: A histoautoradiographic study. *Journal of Comparative Neurology*, *383*, 305–325.

Weinberger, N. M. (1995). Retuning the brain by fear conditioning. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp 1071–1090). Cambridge, MA: MIT Press.

Whalen, P. J., Rauch, S. L., Etcoff, N. L., McInerney, S. C., Lee, M. B., & Jenike, M. A. (1998). Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *Journal of Neuroscience*, *18*, 411–418.

Wilensky, A. E., Schafe, G. E., & LeDoux, J. E. (1999). Functional inactivation of the amygdala before but not after auditory fear conditioning prevents memory formation. *Journal of Neuroscience*, *19*, RC48.

Wright, C. I., Martis, B., McMullin, K., Shin, L. M., & Rauch, S. L. (2003). Amygdala and insular responses to emotionally valenced human faces in small animal specific phobia. *Biological Psychiatry*, *54*, 1067–1076.

Yang, T. T., Menon, V., Eliez, S., Blasey, C., White, C. D., Reid, A. J., Gotlib, I. H., & Reiss, A. L. (2002). Amygdalar activation associated with positive and negative facial expressions. *Neuroreport*, *13*, 1737–1741.

This page intentionally left blank

5 What Are Emotions, Why Do We Have Emotions, and What Is Their Computational Basis in the Brain?

EDMUND T. ROLLS

Emotions may be defined as states elicited by reinforcers (rewards and punishers). This approach helps with understanding the functions of emotion, and with classifying different emotions; and in understanding what information processing systems in the brain are involved in emotion, and how they are involved. The hypothesis is developed that brains are designed around reward and punishment evaluation systems, because this is the way genes can build a complex system that will produce appropriate but flexible behavior to increase their fitness. By specifying goals rather than particular behavioral patterns of responses, genes are open to a much wider range of behavioral strategies, including strategies that increase their fitness.

The primate brain represents the identity of a primary (unlearned) reinforcer first (e.g., for taste in the primary taste cortex) before it decodes the reward or punishment value of the innate reinforcers (in the orbitofrontal cortex, which includes the secondary taste cortex, and the amygdala). Brain regions that represent the identity of objects independently of their reward or punishment value (in the case of vision, the inferior temporal visual cortex) project into the orbitofrontal cortex and amygdala, where neurons learn associations between previously neutral (e.g., visual) stimuli and primary reinforcers (such as taste). This process of stimulus-reinforcement association learning can be very rapid

and flexible in the orbitofrontal cortex, and allows appropriate behavioral responses, such as approach to rewarded stimuli or withdrawal from aversive stimuli, to be generated. It is suggested that there are two types of route to action performed in relation to reward or punishment in humans. Examples of such actions include emotional and motivational behavior. The first route is by way of the brain systems that control behavior in relation to previous associations of stimuli with reinforcement, and include the amygdala and, particularly well-developed in primates, the orbitofrontal cortex. The second route in humans involves a computation with many "if . . . then" statements, to implement a plan to obtain a reward. In this case, syntax is required, because the many symbols that are part of the plan must be correctly linked or bound. The issue of emotional feelings is part of the much larger problem of consciousness and I suggest that it is the second route that is related to consciousness.

What are emotions? Why do we have emotions? What are the rules by which emotion operates? What are the brain mechanisms of emotion, and how can disorders of emotion be understood? Why does it feel like something to have an emotion?

What motivates us to work for particular rewards, such as food when we are hungry or water when we are thirsty? How do these motivational control systems operate to ensure that we eat approximately the correct amount of food to maintain our body weight or to replenish our thirst? What factors account for the overeating and obesity that some humans show?

Why is the brain built to have reward and punishment systems, rather than in some other way? Raising these issues of brain design produces a fascinating answer based on how genes can direct our behavior to increase their fitness. How does the brain produce behavior using reward and punishment mechanisms? These are some of the questions considered in the book *The Brain and Emotion* (Rolls, 1999a) as well as here.

A THEORY OF EMOTION AND SOME DEFINITIONS

Emotions can usefully be defined as states elicited by rewards and punishments, including changes in rewards and punishments (Rolls, 1999a; see also Rolls, 1986a,b, 1990, 2000a). A *reward* is anything for which an animal will work. A *punishment* is anything that an animal will work to escape or avoid. An example of an emotion might thus be happiness produced by being given a reward, such as a pleasant touch, praise, or a large sum of money. Another

example of an emotion might be fear produced by the sound of a rapidly approaching bus or the sight of an angry expression on someone's face. We will work to avoid such stimuli, which are punishing. Another example would be frustration, anger, or sadness produced by the omission of an expected reward, such as a prize, or the termination of a reward, such as the death of a loved one. (O*mission* refers to omitting a reward on an individual trial. *Termination* refers to the end reward presentations.) Another example would be relief produced by the omission or termination of a punishing stimulus, such as occurs with the removal of a painful stimulus or sailing out of danger. These examples indicate how emotions can be produced by the delivery, omission, or termination of rewarding or punishing stimuli and indicate how different emotions could be produced and classified in terms of the rewards and punishments received, omitted, or terminated. A diagram summarizing some of the emotions associated with the delivery of reward or punishment or a stimulus associated with them or with the omission of a reward or punishment is shown in Figure 5.1.

Before accepting this approach, we should consider whether there are any exceptions to the proposed rule. Are any emotions caused by stimuli, events, or remembered events that are not rewarding or punishing? Do any rewarding or punishing stimuli not cause emotions? We will consider these questions in more detail below. The point is that if there are no major exceptions, or if any exceptions can be clearly encapsulated, then we may have a good working definition at least of what causes emotions. Moreover, many approaches to, or theories of, emotion (see Strongman, 1996) have in common that part of the process involves "appraisal" (e.g., Frijda, 1986; Lazarus, 1991; Oatley & Jenkins, 1996). In all these theories, the concept of appraisal presumably involves assessing whether something is rewarding or punishing. The description in terms of reward or punishment adopted here seems more tightly and operationally specified. I next consider a slightly more formal definition than rewards or punishments, in which the concept of reinforcers is introduced, and show how there has been a considerable history in the development of ideas along this line.

Instrumental *reinforcers* are stimuli which, if their occurrence, termination, or omission is made contingent upon the making of an action, alter the probability of the future emission of that action. Rewards and punishers are instrumental reinforcing stimuli. The notion of an action here is that an arbitrary action, for example, turning right versus turning left, will be performed in order to obtain the reward or avoid the punisher, so that there is no prewired connection between the response and the reinforcer. Machines that refuel are not performing instrumental actions unless they are learning arbitrary types of behavior to obtain the fuel. The proposal that emotions can be usefully seen as states produced by instrumental reinforcing stimuli

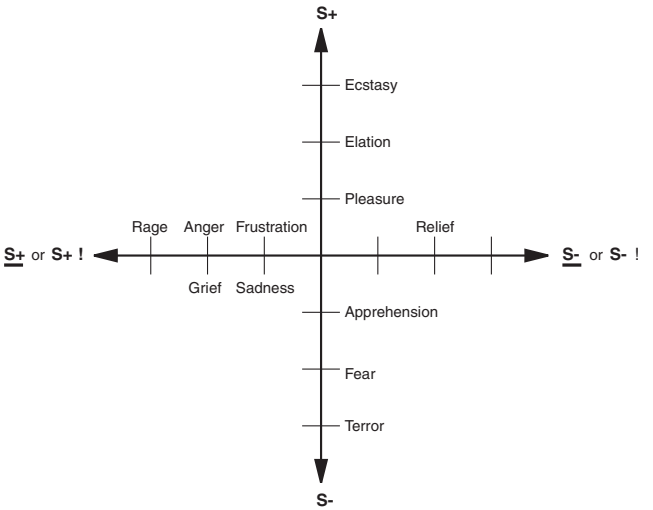


Figure 5.1. Some of the emotions associated with different reinforcement contingencies are indicated. Intensity increases away from the center of the diagram on a continuous scale. The classification scheme created by the different reinforcement contingencies consists of (1) the presentation of a positive reinforcer (S+), (2) the presentation of a negative reinforcer (S–), (3) the omission of a positive reinforcer (S+) or the termination of a positive reinforcer (S+!), and (4) the omission of a negative reinforcer (S–) or the termination of a negative reinforcer (S–!). (From Rolls, 1999a, Fig. 3.1.)

follows earlier work by Millenson (1967), Weiskrantz (1968), Gray (1975, 1987), and Rolls (1986a,b, 1990). Some stimuli are unlearned reinforcers (e.g., the taste of food if the animal is hungry or pain), while others may become reinforcing by learning because of their association with such primary reinforcers, thereby becoming "secondary reinforcers." This type of learning may thus be called "stimulus–reinforcement association" and occurs via an associative process like classical conditioning. If a reinforcer increases the probability of emission of a response on which it is contingent, it is said to be a "positive reinforcer." Rewards are usually positive reinforcers, although one could imagine a situation in which taking no action would produce rewards. If a reinforcer decreases the probability of a response, it is a "negative reinforcer." Punishers can be positive reinforcers (active avoidance) or negative reinforcers (passive avoidance). An example making the link to emotion clear is that fear is an emotional state which might be produced by a sound (the conditioned stimulus) that has previously been associated with an electrical shock (the primary reinforcer).

The converse reinforcement contingencies produce the opposite effects on behavior. The omission or termination of a reward (*extinction* and *time*

out, respectively, sometimes described as "punishing") decreases the probability of response. Responses followed by the omission or termination of a punisher increase in probability, this pair of negative reinforcement operations being termed *active avoidance* and *escape*, respectively (see Gray, 1975; Mackintosh, 1983).

The link between emotion and instrumental reinforcers is partly operational. Most people find that it is not easy to think of exceptions to the statements that emotions occur after rewards or punishers are given (sometimes continuing for long after the eliciting stimulus has ended, as in a mood state) and that rewards and punishers, but not other stimuli, produce emotional states. Emotions are states elicited by reinforcing stimuli. If those states continue for a long time after the eliciting stimulus has gone, or if the states occur spontaneously, we can refer to these as mood states. That is, mood states can be used to refer to states that do not take an object, i.e., when there is no clearly related eliciting stimulus. However, the link is deeper than this, as we will see as I develop the theory that genes specify primary reinforcers in order to encourage the animal to perform arbitrary actions to seek particular goals, which increase the probability of their own (the genes') survival into the next generation. The emotional states elicited by the reinforcers have a number of functions, described below, related to these processes.

This foundation has been developed (see Rolls, 1986a,b, 1990, 1999a, 2000a) to show how a very wide range of emotions can be accounted for, as a result of the operation of a number of factors, including the following:

1. The reinforcement contingency (e.g., whether reward or punishment is given or withheld) (see Fig. 5.1).
2. The intensity of the reinforcer (see Fig. 5.1).
3. Any environmental stimulus might have a number of different reinforcement associations (e.g., a stimulus might be associated with the presentation of both a reward and a punisher, allowing states such as conflict and guilt to arise).[1]
4. Emotions elicited by stimuli associated with different primary reinforcers will be different.
5. Emotions elicited by different secondary reinforcing stimuli will be different from each other (even if the primary reinforcer is similar). For example, if two different people were each associated with the same primary reinforcer, then the emotions would be different. This is in line with my hypothesis that emotions consist of states elicited by reinforcers and that these states include whatever representations are needed for the eliciting stimulus, which could be cognitive, and the resulting mood change (Rolls, 1999a). Moods then may continue in the absence

of the eliciting stimulus or can be produced, as in depression, sometimes in the absence of an eliciting stimulus, perhaps owing to dysregulation in the system that normally enables moods to be long-lasting (see Rolls, 1999a).

6. The emotion elicited can depend on whether an active or passive behavioral response is possible (e.g., if an active behavioral response can occur to the omission of a positive reinforcer, then anger—a state which tends to lead to action—might be produced, but if only passive behavior is possible, then sadness, depression, or grief might occur).

By combining these six factors, it is possible to account for a very wide range of emotions (for elaboration, see Rolls, 1990, 1999a). Emotions can be produced just as much by the recall of reinforcing events as by external reinforcing stimuli[2]; cognitive processing (whether conscious or not) is important in many emotions, for very complex cognitive processing may be required to determine whether or not environmental events are reinforcing. Indeed, emotions normally consist of cognitive processing that analyzes the stimulus and determines its reinforcing valence, then elicits a mood change according to whether the valence is positive or negative. In that an emotion is produced by a stimulus, philosophers say that emotions have an "object" in the world and that emotional states are intentional, in that they are about something. A mood or affective state may occur in the absence of an external stimulus, as in some types of depression; but normally the mood or affective state is produced by an external stimulus, with the whole process of stimulus representation, evaluation in terms of reward or punishment, and the resulting mood or affect being referred to as "emotion." The external stimulus may be perceived consciously, but stimuli that are not perceived consciously may also produce emotion. Indeed, there may be separate routes to action for conscious and unconscious stimuli (Rolls, 1999a).

Three issues are discussed here (see Rolls, 1999a, 2000a). One is that rewarding stimuli, such as the taste of food, are not usually described as producing emotional states (though there are cultural differences here). It is useful here to separate rewards related to internal homeostatic need states associated with regulation of the internal milieu, for example, hunger and thirst, and to note that these rewards are not generally described as producing emotional states. In contrast, the great majority of rewards and punishment are external stimuli not related to internal need states such as hunger and thirst, and these stimuli do produce emotional responses. An example is fear produced by the sight of a stimulus that is about to produce pain. A second issue is that philosophers usually categorize fear in the example as an emotion but not pain. The distinction they make may be that primary

(unlearned) reinforcers do not produce emotions, whereas secondary reinforcers (stimuli associated by stimulus–reinforcement learning with primary reinforcers) do. They describe the pain as a sensation, but neutral stimuli (e.g., a table) can produce sensations when touched. It accordingly seems to be much more useful to categorize stimuli according to whether they are reinforcing (in which case they produce emotions) or not (in which case they do not produce emotions). Clearly, there is a difference between primary reinforcers and learned reinforcers; and operationally, it is whether a stimulus is reinforcing that determines whether it is related to emotion. A third issue is that, as we are about to see, emotional states (i.e., those elicited by reinforcers) have many functions, and the implementations of only some of these functions by the brain are associated with emotional feelings, that is, with conscious emotional states (Rolls, 1999a). Indeed, there is evidence for interesting dissociations in some patients with brain damage between actions performed to reinforcing stimuli and what is subjectively reported. In this sense, it is biologically and psychologically useful to consider that emotional states include more than those states associated with conscious feelings of emotion (Rolls, 1999a).

THE FUNCTIONS OF EMOTION

The functions of emotion also provide insight into the nature of emotion. These functions, described more fully elsewhere (Rolls, 1990, 1999a), can be summarized as follows:

1. *Elicitation of autonomic responses* (e.g., a change in heart rate) *and endocrine responses* (e.g., the release of adrenaline). While this is an important function of emotion, it is the next function that is crucial in my evolutionary theory of why emotion is so important.
2. *Flexibility of behavioral responses to reinforcing stimuli*. Emotional (and motivational) states allow a simple interface between sensory inputs and action systems. The essence of this idea is that goals for behavior are specified by reward and punishment evaluation and that innate goals are specified by genes. When an environmental stimulus has been decoded as a primary reward or punishment or (after previous stimulus–reinforcer association learning) a secondary one it becomes a goal for action. The animal can then perform any action (instrumental response) to obtain the reward or avoid the punishment. The instrumental action, or *operant*, is arbitrary and could consist of a left turn

or a right turn to obtain the goal. It is in this sense that by specifying goals, and not particular actions, the genes are specifying flexible routes to action. This is in contrast to specifying a reflex response and to stimulus–response, or habit, learning in which a particular response to a particular stimulus is learned. It also contrasts with the elicitation of species-typical behavioral responses by sign-releasing stimuli (e.g., pecking at a spot on the beak of the parent herring gull in order to be fed; Tinbergen, 1951), where there is inflexibility of the stimulus and the response, which can be seen as a very limited type of brain solution to the elicitation of behavior. The emotional route to action is flexible not only because any action can be performed to obtain the reward or avoid the punishment but also because the animal can learn in as little as one trial that a reward or punishment is associated with a particular stimulus, in what is termed *stimulus–reinforcer association learning*. It is because goals are specified by the genes, and not actions, that evolution has achieved a powerful way for genes to influence behavior without having to rather inflexibly specify particular responses. An example of a goal might be a sweet taste when hunger is present. We know that particular genes specify the sweet taste receptors (Buck, 2000), and other genes must specify that the sweet taste is rewarding only when there is a homeostatic need state for food (Rolls, 1999a). Different goals or rewards, including social rewards, are specified by different genes; each type of reward must only dominate the others under conditions that prove adaptive if it is to succeed in the phenotype that carries the genes.

To summarize and formalize, two processes are involved in the actions being described. The first is stimulus–reinforcer association learning, and the second is instrumental learning of an operant response made to approach and obtain the reward or to avoid or escape the punisher. Emotion is an integral part of this, for it is the state elicited in the first stage, by stimuli which are decoded as rewards or punishers, and this state is motivating. The motivation is to obtain the reward or avoid the punisher, and animals must be built to obtain certain rewards and avoid certain punishers. Indeed, primary or unlearned rewards and punishers are specified by genes which effectively specify the goals for action. This is the solution which natural selection has found for how genes can influence behavior to promote their fitness (as measured by reproductive success) and for how the brain could interface sensory systems to action systems.

Selecting between available rewards with their associated costs and avoiding punishers with their associated costs is a process which can take place both implicitly (unconsciously) and explicitly using a language system to enable long-term plans to be made (Rolls, 1999a). These many different brain systems, some involving implicit evaluation of rewards and others explicit, verbal, conscious evaluation of rewards and planned long-term goals, must all enter into the selection systems for behavior (see Fig. 5.2). These selector systems are poorly understood but might include a process of competition between all the calls on output and might involve structures such as the cingulate

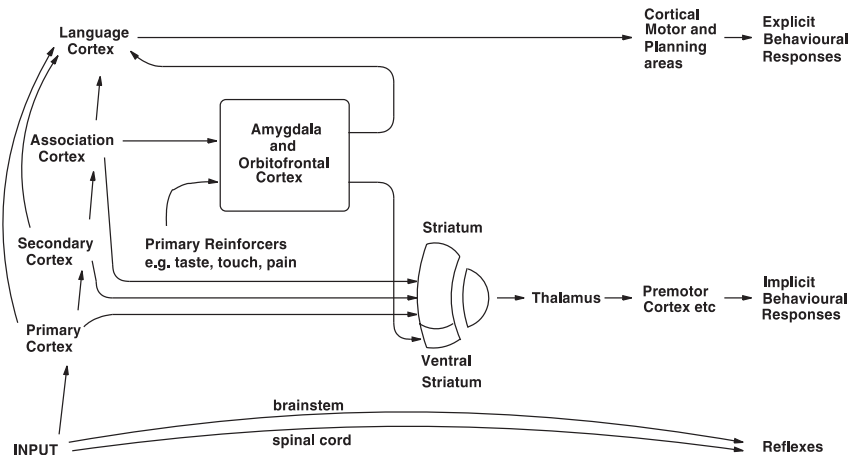


Figure 5.2. Dual routes to the initiation of action in response to rewarding and punishing stimuli. The inputs from different sensory systems to brain structures such as the orbitofrontal cortex and amygdala allow these brain structures to evaluate the reward- or punishment-related value of incoming stimuli or of remembered stimuli. The different sensory inputs enable evaluations within the orbitofrontal cortex and amygdala based mainly on the primary (unlearned) reinforcement value for taste, touch, and olfactory stimuli and on the secondary (learned) reinforcement value for visual and auditory stimuli. In the case of vision, the "association cortex," which outputs representations of objects to the amygdala and orbitofrontal cortex, is the inferior temporal visual cortex. One route for the outputs from these evaluative brain structures is via projections directly to structures such as the basal ganglia (including the striatum and ventral striatum) to enable implicit, direct behavioral responses based on the reward- or punishment-related evaluation of the stimuli to be made. The second route is via the language systems of the brain, which allow explicit (verbalizable) decisions involving multistep syntactic planning to be implemented. (From Rolls, 1999a, Fig. 9.4.)

cortex and basal ganglia in the brain, which receive input from structures such as the orbitofrontal cortex and amygdala that compute the rewards (see Fig. 5.2; Rolls, 1999a).

3. *Motivation*. Emotion is motivating, as just described. For example, fear learned by stimulus–reinforcement association provides the motivation for actions performed to avoid noxious stimuli. Genes that specify goals for action, such as rewards, must as an intrinsic property make the animal motivated to obtain the reward; otherwise, it would not be a reward. Thus, no separate explanation of motivation is required.
4. *Communication*. Monkeys, for example, may communicate their emotional state to others by making an open-mouth threat to indicate the extent to which they are willing to compete for resources, and this may influence the behavior of other animals. This aspect of emotion was emphasized by Darwin (1872/1998) and has been studied more recently by Ekman (1982, 1993). Ekman reviews evidence that humans can categorize facial expressions as happy, sad, fearful, angry, surprised, and disgusted and that this categorization may operate similarly in different cultures. He also describes how the facial muscles produce different expressions. Further investigations of the degree of cross-cultural universality of facial expression, its development in infancy, and its role in social behavior are described by Izard (1991) and Fridlund (1994). As shown below, there are neural systems in the amygdala and overlying temporal cortical visual areas which are specialized for the face-related aspects of this processing. Many different types of gene-specified reward have been suggested (see Table 10.1 in Rolls, 1999a) and include not only genes for kin altruism but also genes to facilitate social interactions that may be to the advantage of those competent to cooperate, as in reciprocal altruism.
5. *Social bonding*. Examples of this are the emotions associated with the attachment of parents to their young and the attachment of young to their parents. The attachment of parents to each other is also beneficial in species, such as many birds and humans, where the offspring are more likely to survive if both parents are involved in the care (see Chapter 8 in Rolls, 1999a).
6. The current mood state can affect the *cognitive evaluation of events or memories* (see Oatley & Jenkins, 1996). This may facilitate continuity in the interpretation of the reinforcing value of events in the environment. The hypothesis that backprojections from parts of the brain involved in emotion, such as the orbito-

frontal cortex and amygdala, to higher perceptual and cognitive cortical areas is described in *The Brain and Emotion*, and developed in a formal model of interacting attractor networks by Rolls and Stringer (2001). In this model, the weak backprojections from the "mood" attractor can, because of associative connections formed when the perceptual and mood states were originally present, influence the states into which the perceptual attractor falls.

7. Emotion may facilitate the *storage of memories*. One way this occurs is that *episodic memory* (i.e., one's memory of particular episodes) is facilitated by emotional states. This may be advantageous in that storing many details of the prevailing situation when a strong reinforcer is delivered may be useful in generating appropriate behavior in situations with some similarities in the future. This function may be implemented by the relatively nonspecific projecting systems to the cerebral cortex and hippocampus, including the cholinergic pathways in the basal forebrain and medial septum and the ascending noradrenergic pathways (see Rolls, 1999a; Rolls & Treves, 1998). A second way in which emotion may affect the storage of memories is that the current emotional state may be stored with episodic memories, providing a mechanism for the current emotional state to affect which memories are recalled. A third way that emotion may affect the storage of memories is by guiding the cerebral cortex in the representations of the world which are established. For example, in the visual system, it may be useful for perceptual representations or analyzers to be built which are different from each other if they are associated with different reinforcers and for these to be less likely to be built if they have no association with reinforcement. Ways in which backprojections from parts of the brain important in emotion (e.g., the amygdala) to parts of the cerebral cortex could perform this function are discussed by Rolls and Treves (1998) and Rolls and Stringer (2001).
8. Another function of emotion is that by enduring for minutes or longer after a reinforcing stimulus has occurred, it may help to produce *persistent and continuing motivation and direction of behavior*, to help achieve a goal or goals.
9. Emotion may trigger the *recall of memories* stored in neocortical representations. Amygdala backprojections to the cortex could perform this for emotion in a way analogous to that in which the hippocampus could implement the retrieval in the neocor-

tex of recent (episodic) memories (Rolls & Treves, 1998; Rolls & Stringer, 2001). This is one way in which the recall of memories can be biased by mood states.

REWARD, PUNISHMENT, AND EMOTION IN BRAIN DESIGN: AN EVOLUTIONARY APPROACH

The theory of the functions of emotion is further developed in Chapter 10 of *The Brain and Emotion* (Rolls, 1999a). Some of the points made help to elaborate greatly on the second function in the list above. In that chapter, the fundamental question of why we and other animals are built to use rewards and punishments to guide or determine our behavior is considered. Why are we built to have emotions as well as motivational states? Is there any reasonable alternative around which evolution could have built complex animals? In this section, I outline several types of brain design, with differing degrees of complexity, and suggest that evolution can operate to influence action with only some of these types of design.

Taxes

A simple design principle is to incorporate mechanisms for taxes into the design of organisms. *Taxes* consist at their simplest of orientation toward stimuli in the environment, for example, *phototaxis* can take the form of the bending of a plant toward light, which results in maximum light collection by its photosynthetic surfaces. (When just turning rather than locomotion is possible, such responses are called *tropisms*.) With locomotion possible, as in animals, taxes include movements toward sources of nutrient and away from hazards, such as very high temperatures. The design principle here is that animals have, through natural selection, built receptors for certain dimensions of the wide range of stimuli in the environment and have linked these receptors to mechanisms for particular responses in such a way that the stimuli are approached or avoided.

Reward and Punishment

As soon as we have "approach toward stimuli" at one end of a dimension (e.g., a source of nutrient) and "move away from stimuli" at the other end (in this case, lack of nutrient), we can start to wonder when it is appropriate to introduce the terms *reward* and *punishers* for the different stimuli. By

convention, if the response consists of a fixed reaction to obtain the stimulus (e.g., locomotion up a chemical gradient), we shall call this a "taxis," not a "reward." If an arbitrary operant response can be performed by the animal in order to approach the stimulus, then we will call this "rewarded behavior" and the stimulus the animal works to obtain is a "reward." (The operant response can be thought of as any arbitrary action the animal will perform to obtain the stimulus.) This criterion, of an arbitrary operant response, is often tested by bidirectionality. For example, if a rat can be trained to either raise or lower its tail in order to obtain a piece of food, then we can be sure that there is no fixed relationship between the stimulus (e.g., the sight of food) and the response, as there is in a taxis. Similarly, reflexes are arbitrary operant actions performed to obtain a goal.

The role of natural selection in this process is to guide animals to build sensory systems that will respond to dimensions of stimuli in the natural environment along which actions can lead to better ability to pass genes on to the next generation, that is, to increased fitness. Animals must be built by such natural selection to make responses that will enable them to obtain more rewards, that is, to work to obtain stimuli that will increase their fitness. Correspondingly, animals must be built to make responses that will enable them to escape from, or learn to avoid, stimuli that will reduce their fitness. There are likely to be many dimensions of environmental stimuli along which responses can alter fitness. Each of these may be a separate reward–punishment dimension. An example of one of these dimensions might be food reward. It increases fitness to be able to sense nutrient need, to have sensors that respond to the taste of food, and to perform behavioral responses to obtain such reward stimuli when in that need or motivational state. Similarly, another dimension is water reward, in which the taste of water becomes rewarding when there is body fluid depletion (see Chapter 7 of Rolls, 1999a). Another dimension might be quite subtly specified rewards to promote, for example, kin altruism and reciprocal altruism (e.g., a "cheat" or "defection" detector).

With many primary (genetically encoded) reward–punishment dimensions for which actions may be performed (see Table 10.1 of Rolls, 1999a, for a nonexhaustive list!), a selection mechanism for actions performed is needed. In this sense, rewards and punishers provide a *common currency* for inputs to response selection mechanisms. Evolution must set the magnitudes of the different reward systems so that each will be chosen for action in such a way as to maximize overall fitness (see the next section). Food reward must be chosen as the aim for action if a nutrient is depleted, but water reward as a target for action must be selected if current water depletion poses a greater threat to fitness than the current food depletion. This indicates that each genetically specified reward must be carefully calibrated by evolution to have

the right value in the common currency for the competitive selection process. Other types of behavior, such as sexual behavior, must be selected sometimes, but probably less frequently, in order to maximize fitness (as measured by gene transmission to the next generation). Many processes contribute to increasing the chances that a wide set of different environmental rewards will be chosen over a period of time, including not only need-related satiety mechanisms, which decrease the rewards within a dimension, but also sensory-specific satiety mechanisms, which facilitate switching to another reward stimulus (sometimes within and sometimes outside the same main dimension), and attraction to novel stimuli. Finding novel stimuli rewarding is one way that organisms are encouraged to explore the multidimensional space in which their genes operate.

The above mechanisms can be contrasted with typical engineering design. In the latter, the engineer defines the requisite function and then produces special-purpose design features that enable the task to be performed. In the case of the animal, there is a multidimensional space within which many optimizations to increase fitness must be performed, but the fitness function is just how successfully genes survive into the next generation. The solution is to evolve reward–punishment systems tuned to each dimension in the environment which can increase fitness if the animal performs the appropriate actions. Natural selection guides evolution to find these dimensions. That is, the design "goal" of evolution is to maximize the survival of a gene into the next generation, and emotion is a useful adaptive feature of this design. In contrast, in the engineering design of a robot arm, the robot does not need to tune itself to find the goal to be performed. The contrast is between design by evolution which is "blind" to the purpose of the animal and "seeks" to have individual genes survive into future generations and design by a designer or engineer who specifies the job to be performed (cf. Dawkins, 1986; Rolls & Stringer, 2000). A major distinction here is between the system designed by an engineer to perform a particular purpose, for example a robot arm, and animals designed by evolution where the "goal" of each gene is to replicate copies of itself into the next generation. Emotion is useful in an animal because it is part of the mechanism by which some genes seek to promote their own survival, by specifying goals for actions. This is not usually the design brief for machines designed by humans. Another contrast is that for the animal the space will be high-dimensional, so that the most appropriate reward to be sought by current behavior (taking into account the costs of obtaining each reward) needs to be selected and the behavior (the operant response) most appropriate to obtain that reward must consequently be selected, whereas the movement to be made by the robot arm is usually specified by the design engineer.

The implication of this comparison is that operation by animals using reward and punishment systems tuned to dimensions of the environment

that increase fitness provides a mode of operation that can work in organisms that evolve by natural selection. It is clearly a natural outcome of Darwinian evolution to operate using reward and punishment systems tuned to fitness-related dimensions of the environment if arbitrary responses are to be made by animals, rather than just preprogrammed movements, such as taxes and reflexes. Is there any alternative to such a reward–punishment-based system in this evolution by natural selection situation? It is not clear that there is, if the genes are efficiently to control behavior by specifying the goals for actions. The argument is that genes can specify actions that will increase their fitness if they specify the goals for action. It would be very difficult for them in general to specify in advance the particular responses to be made to each of a myriad different stimuli. This may be why we are built to work for rewards, to avoid punishers, and to have emotions and needs (motivational states). This view of brain design in terms of reward and punishment systems built by genes that gain their adaptive value by being tuned to a goal for action (Rolls, 1999a) offers, I believe, a deep insight into how natural selection has shaped many brain systems and is a fascinating outcome of Darwinian thought.

DUAL ROUTES TO ACTION

It is suggested (Rolls, 1999a) that there are two types of route to action performed in relation to reward or punishment in humans. Examples of such actions include emotional and motivational behavior.

The First Route

The first route is via the brain systems that have been present in nonhuman primates, and, to some extent, in other mammals for millions of years. These systems include the amygdala and, particularly well developed in primates, the orbitofrontal cortex. (More will be said about these brain regions in the following section.) These systems control behavior in relation to previous associations of stimuli with reinforcement. The computation which controls the action thus involves assessment of the reinforcement-related value of a stimulus. This assessment may be based on a number of different factors. One is the previous reinforcement history, which involves stimulus–reinforcement association learning using the amygdala and its rapid updating, especially in primates, using the orbitofrontal cortex. This stimulus–reinforcement association learning may involve quite specific information about a stimulus, for example, the energy associated with each type of food

by the process of conditioned appetite and satiety (Booth, 1985). A second is the current motivational state, for example, whether hunger is present, whether other needs are satisfied, etc. A third factor which affects the computed reward value of the stimulus is whether that reward has been received recently. If it has been received recently but in small quantity, this may increase the reward value of the stimulus. This is known as *incentive motivation* or the *salted peanut phenomenon*. The adaptive value of such a process is that this positive feedback of reward value in the early stages of working for a particular reward tends to lock the organism onto behavior being performed for that reward. This means that animals that are, for example, almost equally hungry and thirsty will show hysteresis in their choice of action, rather than continually switching from eating to drinking and back with each mouthful of water or food. This introduction of hysteresis into the reward evaluation system makes action selection a much more efficient process in a natural environment, for constantly switching between different types of behavior would be very costly if all the different rewards were not available in the same place at the same time. (For example, walking half a mile between a site where water was available and a site where food was available after every mouthful would be very inefficient.) The amygdala is one structure that may be involved in this increase in the reward value of stimuli early in a series of presentations; lesions of the amygdala (in rats) abolish the expression of this reward incrementing process, which is normally evident in the increasing rate of working for a food reward early in a meal and impair the hysteresis normally built into the food–water switching mechanism (Rolls & Rolls, 1973). A fourth factor is the computed absolute value of the reward or punishment expected or being obtained from a stimulus, for example, the sweetness of the stimulus (set by evolution so that sweet stimuli will tend to be rewarding because they are generally associated with energy sources) or the pleasantness of touch (set by evolution to be pleasant according to the extent to which it brings animals together, e.g., for sexual reproduction, maternal behavior, and grooming, and depending on the investment in time that the partner is willing to put into making the touch pleasurable, a sign which indicates the commitment and value for the partner of the relationship).

After the reward value of the stimulus has been assessed in these ways, behavior is initiated based on approach toward or withdrawal from the stimulus. A critical aspect of the behavior produced by this type of system is that it is aimed directly at obtaining a sensed or expected reward, by virtue of connections to brain systems such as the basal ganglia which are concerned with the initiation of actions (see Fig. 5.2). The expectation may, of course, involve behavior to obtain stimuli associated with reward, which might even be present in a linked sequence. This expectation is built by stimulus–reinforcement association learning in the amygdala and orbitofrontal cortex,

reversed by learning in the orbitofrontal cortex, from where signals may be sent to the dopamine system (Rolls, 1999a).

Part of the way in which the behavior is controlled with this first route is according to the reward value of the outcome. At the same time, the animal may work for the reward only if the cost is not too high. Indeed, in the field of behavioral ecology, animals are often thought of as performing optimally on some cost–benefit curve (see, e.g., Krebs & Kacelnik, 1991). This does not at all mean that the animal thinks about the rewards and performs a cost–benefit analysis using thoughts about the costs, other rewards available and their costs, etc. Instead, it should be taken to mean that in evolution the system has so evolved that the way in which the reward varies with the different energy densities or amounts of food and the delay before it is received can be used as part of the input to a mechanism which has also been built to track the costs of obtaining the food (e.g., energy loss in obtaining it, risk of predation, etc.) and to then select, given many such types of reward and associated costs, the behavior that provides the most "net reward." Part of the value of having the computation expressed in this reward-minus-cost form is that there is then a suitable "currency," or net reward value, to enable the animal to select the behavior with currently the most net reward gain (or minimal aversive outcome).

The Second Route

The second route in humans involves a computation with many "if . . . then" statements, to implement a plan to obtain a reward. In this case, the reward may actually be deferred as part of the plan, which might involve working first to obtain one reward and only then for a second, more highly valued reward, if this was thought to be overall an optimal strategy in terms of resource usage (e.g., time). In this case, syntax is required because the many symbols (e.g., names of people) that are part of the plan must be correctly linked or bound. Such linking might be of the following form: "if A does this, then B is likely to do this, and this will cause C to do this." This implies that an output to a language system that at least can implement syntax in the brain is required for this type of planning (see Fig. 5.2; Rolls, 2004). Thus, the explicit language system in humans may allow working for deferred rewards by enabling use of a one-off, individual plan appropriate for each situation. Another building block for such planning operations in the brain may be the type of short-term memory in which the prefrontal cortex is involved. For example, this short-term memory in nonhuman primates may be of where in space a response has just been made. Development of this type of short-term response memory system in humans enables multiple short-term

memories to be held in place correctly, preferably with the temporal order of the different items coded correctly. This may be another building block for the multiple-step "if . . . then" type of computation in order to form a multiple-step plan. Such short-term memories are implemented in the (dorsolateral and inferior convexity) prefrontal cortex of nonhuman primates and humans (see Goldman-Rakic, 1996; Petrides, 1996; Rolls & Deco, 2002) and may be part of the reason why prefrontal cortex damage impairs planning (see Shallice & Burgess, 1996; Rolls & Deco, 2002).

Of these two routes (see Fig. 5.2), it is the second, involving syntax, which I have suggested above is related to consciousness. The hypothesis is that consciousness is the state that arises by virtue of having the ability to think about one's own thoughts, which has the adaptive value of enabling one to correct long, multistep syntactic plans. This latter system is thus the one in which explicit, declarative processing occurs. Processing in this system is frequently associated with reason and rationality in that many of the consequences of possible actions can be taken into account. The actual computation of how rewarding a particular stimulus or situation is or will be probably still depends on activity in the orbitofrontal cortex and amygdala as the reward value of stimuli is computed and represented in these regions and verbalized expressions of the reward (or punishment) value of stimuli are dampened by damage to these systems. (For example, damage to the orbitofrontal cortex renders painful input still identifiable as pain but without the strong affective "unpleasant" reaction to it; see Rolls, 1999a.) This language system that enables long-term planning may be contrasted with the first system in which behavior is directed at obtaining the stimulus (including the remembered stimulus) that is currently the most rewarding, as computed by brain structures that include the orbitofrontal cortex and amygdala. There are outputs from this system, perhaps those directed at the basal ganglia, which do not pass through the language system; behavior produced in this way is described as "implicit," and verbal declarations cannot be made directly about the reasons for the choice made. When verbal declarations are made about decisions made in this first system, they may be confabulations, reasonable explanations, or fabrications of reasons why the choice was made. Reasonable explanations would be generated to be consistent with the sense of continuity and self that is a characteristic of reasoning in the language system.

The question then arises of how decisions are made in animals such as humans that have both the implicit, direct, reward-based and the explicit, rational, planning systems (see Fig. 5.2). One particular situation in which the first, implicit, system may be especially important is when rapid reactions to stimuli with reward or punishment value must be made, for then the direct connections from structures such as the orbitofrontal cortex to

the basal ganglia may allow rapid actions. Another is when there may be too many factors to be taken into account easily by the explicit, rational, planning system when the implicit system may be used to guide action. In contrast, when the implicit system continually makes errors, it would be beneficial for the organism to switch from automatic, direct action based on obtaining what the orbitofrontal cortex system decodes as being the most positively reinforcing choice currently available to the explicit, conscious control system, which can evaluate with its long-term planning algorithms what action should be performed next. Indeed, it would be adaptive for the explicit system to regularly assess performance by the more automatic system and to switch itself to control behavior quite frequently as otherwise the adaptive value of having the explicit system would be less than optimal. Another factor which may influence the balance between control by the implicit and explicit systems is the presence of pharmacological agents such as alcohol, which may alter the balance toward control by the implicit system, may allow the implicit system to influence more the explanations made by the explicit system, and may within the explicit system alter the relative value it places on caution and restraint versus commitment to a risky action or plan.

There may also be a flow of influence from the explicit, verbal system to the implicit system such that the explicit system may decide on a plan of action or strategy and exert an influence that will alter the reinforcement evaluations made by and the signals produced by the implicit system. An example of this might be that if a pregnant woman feels that she would like to escape a cruel mate but is aware that she may not survive in the jungle, then it would be adaptive if the explicit system could suppress some aspects of her implicit behavior toward her mate so that she does not give signals that she is displeased with her situation. (In the literature on self-deception, it has been suggested that unconscious desires may not be made explicit in consciousness [or actually repressed] so as not to compromise the explicit system in what it produces; see Alexander, 1975, 1979; Trivers, 1976, 1985; and the review by Nesse & Lloyd, 1992). Another example is that the explicit system might, because of its long-term plans, influence the implicit system to increase its response to a positive reinforcer. One way in which the explicit system might influence the implicit system is by setting up the conditions in which, when a given stimulus (e.g., a person) is present, positive reinforcers are given to facilitate stimulus–reinforcement association learning by the implicit system of the person receiving the positive reinforcers. Conversely, the implicit system may influence the explicit system, for example, by highlighting certain stimuli in the environment that are currently associated with reward, to guide the attention of the explicit system to such stimuli.

However, it may be expected that there is often a conflict between these systems in that the first, implicit, system is able to guide behavior particularly to obtain the greatest immediate reinforcement, whereas the explicit system can potentially enable immediate rewards to be deferred and longer-term, multistep plans to be formed. This type of conflict will occur in animals with a syntactic planning ability (as described above), that is, in humans and any other animals that have the ability to process a series of "if . . . then" stages of planning. This is a property of the human language system, and the extent to which it is a property of nonhuman primates is not yet fully clear. In any case, such conflict may be an important aspect of the operation of at least the human mind because it is so essential for humans to correctly decide, at every moment, whether to invest in a relationship or a group that may offer long-term benefits or whether to directly pursue immediate benefits (Nesse & Lloyd, 1992). As Nesse and Lloyd (1992) describe, psychoanalysts have come to a somewhat similar position, for they hold that intrapsychic conflicts usually seem to have two sides, with impulses on one side and inhibitions on the other. Analysts describe the source of the impulses as the *id* and the modules that inhibit the expression of impulses, because of external and internal constraints, as the *ego* and *superego*, respectively (Leak & Christopher, 1982; Trivers, 1985; see Nesse & Lloyd, 1992, p. 613). The superego can be thought of as the conscience, while the ego is the locus of executive functions that balance satisfaction of impulses with anticipated internal and external costs. A difference of the present position is that it is based on identification of dual routes to action implemented by different systems in the brain, each with its own selective advantage.

BRAIN SYSTEMS UNDERLYING EMOTION

Overview

Animals are built with neural systems that enable them to evaluate which environmental stimuli, whether learned or not, are rewarding and punishing, that is, will produce emotions and will be worked for or avoided. Sensory stimuli are normally processed through several stages of cortical processing to produce a sensory representation of the object before emotional valence is decoded, and subcortical inputs to, e.g., the amygdala (LeDoux, 2000) will be of little use when most emotions are to stimuli that require processing to the object level (Rolls, 1999a). For example, in the taste system, taste is analyzed in primates to provide a representation of what the taste is in the primary taste cortex, and this representation is independent of the reward

value of the taste in that it is not affected by hunger. In the secondary taste cortex, in the orbitofrontal region (see Figs. 5.3 and 5.4), the reward value of the taste is represented in that neurons respond to the taste only if the primate is hungry. In another example, in the visual system, representations of objects which are view-, position- and size-invariant are produced in the inferior temporal visual cortex after many stages of cortical processing (see Rolls & Deco, 2002); and these representations are independent of the emotional valence of the object. Then, in structures such as the orbitofrontal cortex and amygdala, which receive input from the inferior temporal visual

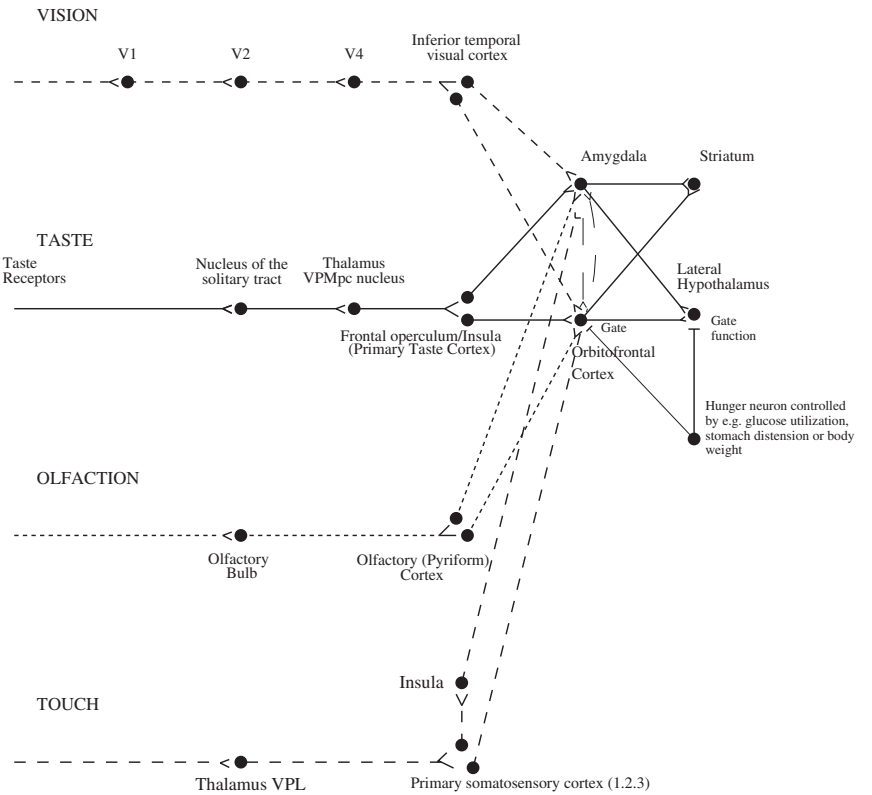


Figure 5.3. Schematic diagram showing some of the gustatory, olfactory, visual, and somatosensory pathways to the orbitofrontal cortex and amygdala and some of the outputs of the orbitofrontal cortex and amygdala. The secondary taste cortex and the secondary olfactory cortex are within the orbitofrontal cortex. V1, primary visual cortex; V2 and V4, visual cortical areas; VP1, ventral posterolateral; VPM, ventral posterior medial.

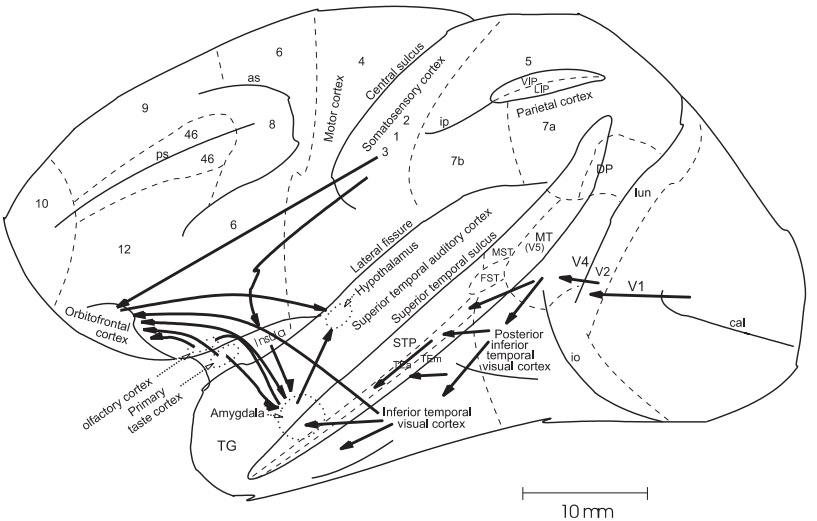


Figure 5.4. Some of the pathways involved in emotion described in the text are shown on this lateral view of the brain of the macaque monkey. Connections from the primary taste and olfactory cortices to the orbitofrontal cortex and amygdala are shown. Connections are also shown in the "ventral visual system" from V1 to V2, V4, the inferior temporal visual cortex (TEO and TE), etc., with some connections reaching the amygdala and orbitofrontal cortex. In addition, connections from somatosensory cortical areas 1, 2, and 3 that reach the orbitofrontal cortex directly and via the insular cortex and that reach the amygdala via the insular cortex are shown. *Abbreviations*: as, arcuate sulcus; cal, calcarine sulcus; cs, central sulcus; lf, lateral (or sylvian) fissure; lun, lunate sulcus; ps, principal sulcus; io, inferior occipital sulcus; ip, intraparietal sulcus (which has been opened to reveal some of the areas it contains); sts, superior temporal sulcus (which has been opened to reveal some of the areas it contains); AIT, anterior inferior temporal cortex; FST (fundus superior temporal) visual motion processing area; LIP, lateral intraparietal area; MST, and MT (also called VS), are visual motion processing areas; PIT, posterior inferior temporal cortex; STP, superior temporal plane; TA, architectonic area including auditory association cortex; TE, architectonic area including high-order visual association cortex and some of its subareas (TEa and Tem); TG, architectonic area in the temporal pole; V1–V4, visual areas 1–4; VIP, ventral intraparietal area; TEO, architectonic area including posterior visual association cortex. The numerals refer to architectonic areas and have the following approximate functional equivalence: 1, 2, 3, somatosensory cortex (posterior to the central sulcus); 4, motor cortex; 5, superior parietal lobule; 7a, inferior parietal lobule, visual part; 7b, inferior parietal lobule, somatosensory part; 6, lateral premotor cortex; 8, frontal eye field; 12, part of orbitofrontal cortex; 46, dorsolateral prefrontal cortex.

cortex, associations are learned between the objects and the primary reinforcers associated with them by the process of stimulus–reinforcement association learning. This is implemented by pattern association neural networks (Rolls & Deco, 2002). In the orbitofrontal cortex and amygdala, emotional states are thus represented. Consistent with this, electrical stimulation of the orbitofrontal cortex and amygdala is rewarding, and damage to these structures affects emotional behavior by affecting stimulus–reinforcement association learning. These brain regions influence the selection of behavioral actions through brain systems such as the ventral striatum and other parts of the basal ganglia (see Fig. 5.2).

The Amygdala

The amygdala receives information about primary reinforcers (e.g., taste and touch) and about visual and auditory stimuli from higher cortical areas (e.g., the inferior temporal cortex) that can be associated by learning with primary reinforcers (Figs. 5.3 and 5.4). Bilateral removal of the amygdala in monkeys produces tameness; a lack of emotional responsiveness; excessive examination of objects, often with the mouth; and eating of previously rejected items, such as meat (the Klüver-Bucy syndrome). In analyses of the bases of these behavioral changes, it has been observed that there are deficits in learning to associate stimuli with primary reinforcement, including both punishments and rewards (see Rolls, 2000c). The association learning deficit is present when the associations must be learned from a previously neutral stimulus (e.g., the sight of an object) to a primary reinforcing stimulus (e.g., the taste of food). Further evidence linking the amygdala to reinforcement mechanisms is that monkeys will work in order to obtain electrical stimulation of the amygdala, that single neurons in the amygdala are activated by brain-stimulation reward of a number of different sites, and that some amygdala neurons respond mainly to rewarding stimuli and others to punishing stimuli (see Rolls, 1999a, 2000c). The association learning in the amygdala may be implemented by associatively modifiable synapses from visual and auditory neurons onto neurons receiving inputs from taste, olfactory, or somatosensory primary reinforcers (LeDoux, 1996; and Fellous & LeDoux in this volume). Consistent with this, Davis (2000) found that at least one type of associative learning in the amygdala can be blocked by local application to the amygdala of an *N*-methyl-D-aspartate receptor blocker, which blocks long-term potentiation and is a model of the synaptic changes that underlie learning (see Rolls & Treves, 1998). Consistent with the hypothesis that the learned incentive (conditioned reinforcing) effects of previously neutral stimuli paired with rewards are mediated by the amygdala

acting through the ventral striatum, amphetamine injections into the ventral striatum enhanced the effects of a conditioned reinforcing stimulus only if the amygdala was intact (see Everitt et al., 2000).

An interesting group of neurons in the amygdala (e.g., in the basal accessory nucleus) responds primarily to faces. They are probably part of a system which has evolved for the rapid and reliable identification of individuals from their faces and of facial expressions because of the importance of this in primate social behavior. Consistent with this, activation of the human amygdala can be produced in neuroimaging studies by some facial expressions, and lesions of the human amygdala may cause difficulty in the identification of some facial expressions (see Rolls, 1999a, 2000c).

The Orbitofrontal Cortex

The orbitofrontal cortex receives inputs from the inferior temporal visual cortex, superior temporal auditory cortex, primary taste cortex, primary olfactory (pyriform) cortex (see Figs. 5.3 and 5.4), amygdala, and midbrain dopamine neurons. Damage to the caudal orbitofrontal cortex in the monkey produces emotional changes. These include decreased aggression to humans and to stimuli such as a snake and a doll and a reduced tendency to reject foods such as meat. These changes may be related to a failure to react normally to and learn from nonrewards in a number of different situations. This failure is evident as a tendency to respond when responses are inappropriate, for example, no longer rewarded. For example, monkeys with orbitofrontal damage are impaired on Go/NoGo task performance (in which they should make a response to one stimulus to obtain a reward and should not make a response to another stimulus in order to avoid a punishment), in that they Go on the NoGo trials. They are also impaired in an object reversal task in that they respond to the object which was formerly rewarded with food. They are also impaired in extinction in that they continue to respond to an object which is no longer rewarded. Further, the visual discrimination learning deficit shown by monkeys with orbitofrontal cortex damage may be due to their tendency not to withhold responses to nonrewarded stimuli (see Rolls, 1999a, 2002).

The primate orbitofrontal cortex contains neurons which respond to the reward value of taste (a primary reinforcer) in that they respond to the taste of food only when hunger is present (which is when food is rewarding). It also contains neurons which learn to respond to visual stimuli associated with a primary reward, such as taste, and which reverse their responses to another visual stimulus in one trial when the rewards and punishers available from those visual stimuli reverse. Further, these visual responses reflect reward in that feeding the monkey to satiety reduces the responses of these neurons to zero.

Moreover, in part of this orbitofrontal region, some neurons combine taste and olfactory inputs in that they are bimodal and, in 40% of cases, affected by olfactory-to-taste association learning and by feeding the monkey to satiety, which reduces the reward value (see Rolls, 1999a, 2000b, 2002). In addition, some neurons in the primate orbitofrontal cortex respond to the sight of faces. These neurons are likely to be involved in learning which emotional responses are currently appropriate to particular individuals and in making appropriate emotional responses given the facial expression.

Another class of neurons in the orbitofrontal cortex of the monkey responds in certain nonreward situations. For example, some neurons responded in extinction immediately after a lick action was not rewarded when it was made after a visual stimulus was shown which had previously been associated with fruit juice reward. Other neurons responded in a reversal task immediately after the monkey had responded to the previously rewarded visual stimulus but had obtained punishment rather than reward. Another class of orbitofrontal neuron responded to particular visual stimuli only if they were associated with reward, and these neurons showed one trial stimulus–reinforcement association reversal (Thorpe, Rolls, & Maddison, 1983; Rolls, 1999a, 2000b, 2002). Another class of neuron conveyed information about whether a reward had been given, responding, for example, to the taste of sucrose or of saline.

These types of information may be represented in the responses of orbitofrontal neurons because they are part of a mechanism which evaluates whether a reward is expected and generate a mismatch (evident as a firing of the nonreward neurons) if reward is not obtained when it is expected (see Rolls, 1999a, 2000a,b, 2002; Kringelbach & Rolls, 2003). These neuronal responses provide further evidence that the orbitofrontal cortex is involved in emotional responses, particularly when these involve correcting previously learned reinforcement contingencies, in situations which include those usually described as involving frustration.

It is of interest and potential clinical importance that a number of the symptoms of frontal lobe damage in humans appear to be related to this type of function, of altering behavior when stimulus–reinforcement associations alter, as described next. Thus, humans with frontal lobe damage can show impairments in a number of tasks in which an alteration of behavioral strategy is required in response to a change in environmental reinforcement contingencies (Rolls, Hornak, Wade, & McGrath, 1994; Damasio, 1994; Rolls, 1999b). Some of the personality changes that can follow frontal lobe damage may be related to a similar type of dysfunction. For example, the euphoria, irresponsibility, lack of affect, and lack of concern for the present or future which can follow frontal lobe damage may also be related to a dysfunction in altering behavior appropriately in response to a change in reinforcement contingencies. At one time, following a report by Moniz (1936), prefrontal lobotomies or leucotomies

(cutting white matter) were performed in humans to attempt to alleviate a variety of problems; and although irrational anxiety or emotional outbursts were sometimes controlled, intellectual deficits and other side effects were often apparent (see Valenstein, 1974). Thus, these operations have been essentially discontinued. To investigate the possible significance of face-related inputs to orbitofrontal visual neurons described above, the responses to faces that were made by patients with orbitofrontal damage produced by pathology or trauma were tested. Impairments in the identification of facial and vocal emotional expression were demonstrated in a group of patients with ventral frontal lobe damage who had socially inappropriate behavior (Hornak, Rolls, & Wade, 1996; Rolls, 1999b; Hornak et al., 2003a,b). The expression identification impairments could occur independently of perceptual impairments in facial recognition, voice discrimination, or environmental sound recognition. Thus, the orbitofrontal cortex in humans appears to be important not only in the rapid relearning of stimulus–reinforcement associations but also in representing some of the stimuli, such as facial expression, which provide reinforcing information. Consistent with this, neuroimaging studies in humans show representations which reflect the pleasantness of the taste and smell of food and of touch, as well as quite abstract rewards and punishers such as winning or losing money (O'Doherty et al., 2001).

The behavioral selection system must deal with many competing rewards, goals, and priorities. This selection process must be capable of responding to many different types of reward decoded in different brain systems that have evolved at different times, even including the use in humans of a language system to enable long-term plans to be made to obtain goals. These many different brain systems, some involving implicit (unconscious) evaluation of rewards and others explicit, verbal, conscious evaluation of rewards and planned long-term goals, must all enter into the selection of behavior. Although poorly understood, emotional feelings are part of the much larger problem of consciousness and may involve the capacity to have thoughts about thoughts, that is, higher-order thoughts (see Rolls, 1999a, 2000a).

CONCLUSION

This approach leads to an appreciation that in order to understand brain mechanisms of emotion and motivation, it is necessary to understand how the brain decodes the reinforcement value of primary reinforcers, how it performs stimulus–reinforcement association learning to evaluate whether a previously neutral stimulus is associated with reward or punishment and

is therefore a goal for action, and how the representations of these neutral sensory stimuli are appropriate as input to such stimulus–reinforcement learning mechanisms. (Some of these issues are considered in *The Brain and Emotion*: emotion in Chapter 4, feeding in Chapter 2, drinking in Chapter 7, and sexual behavior in Chapter 8.)

This approach also does not deny that it would be possible to implement emotions in computers and specifies what may need to be implemented for both implicit and explicit emotions, that is, emotions with conscious feelings. It could even be useful to implement some aspects of emotion in computers as humans may find it more natural to then deal with computers. However, I have summarized a theory of the evolutionary utility of emotion, which is that emotion arises from the gene-based design of organisms by which individual genes maximize their own survival into the next generation by specifying the goals for flexible (arbitrary) actions. As such, emotion arises as part of a blind search by genes to maximize their own survival, which is the "goal" of evolution. In contrast, the goal of human-designed computers and robots is not to provide for survival of competing genes but, instead, to achieve particular design goals specified by the engineer, such as exploring new terrain and sending back pictures to earth, lifting a heavy weight, or translating from one language to another.

Notes

The author has worked on some of the experiments described here with G. C. Baylis, L. L. Baylis, M. J. Burton, H. C. Critchley, M. E. Hasselmo, J. Hornak, M. Kringelbach, C. M. Leonard, F. Mora, J. O'Doherty, D. I. Perrett, M. K. Sanghera, T. R. Scott, S. J. Thorpe, and F. A. W. Wilson; and their collaboration and helpful discussions with or communications from M. Davies and M. S. Dawkins are sincerely acknowledged. Some of the research described was supported by the Medical Research Council.

1. Rewards and punishers are generally external, that is, exteroceptive, stimuli, such as the sight, smell, and taste of food when hungry. Interoceptive stimuli, even when produced by rewards and punishers after ingesting foods and including digestive processes and the reduction of the drive (hunger) state, are not good reinforcers. Some of the evidence for this is that the taste of food is an excellent reinforcer, but placing food into the stomach is not. This important distinction is described by Rolls (1999a).

2. Part of the basis for this is that when memories are recalled, top-down connections into the higher perceptual and cognitive cortical areas lead to reinstatement of activity in those areas (Treves & Rolls, 1994; Rolls & Deco, 2002), which in turn can produce emotional states via onward connections to the orbitofrontal cortex and amygdala (Rolls, 1999a).

References

Alexander, R. D. (1975). The search for a general theory of behavior. *Behavioral Sciences*, *20*, 77–100.

Alexander, R. D. (1979). *Darwinism and human affairs*. Seattle: University of Washington Press.

Booth, D. A. (1985). Food-conditioned eating preferences and aversions with interoceptive elements: Learned appetites and satieties. *Annals of the New York Academy of Sciences*, *443*, 22–37.

Buck, L. (2000). Smell and taste: The chemical senses. In E. R. Kandel, J. H. Schwartz, & T. H. Jessel (Eds.), *Principles of neural science* (4th ed.) New York: McGraw-Hill.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York : Putnam.

Darwin, C. (1998). *The expression of the emotions in man and animals* (3rd ed.). Chicago: University of Chicago Press. (Original work published 1872)

Davis, M. (2000). The role of the amygdala in conditioned and unconditioned fear and anxiety. In J. P. Aggleton (Ed.), *The amygdala: A functional analysis* (2nd ed., pp. 213–228). Oxford: Oxford University Press.

Dawkins, R. (1986). *The blind watchmaker*. Harlow: Longman.

Ekman, P. (1982). *Emotion in the human face* (2nd ed.). Cambridge: Cambridge University Press.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, *48*, 384–392.

Everitt, B. J., Cardinal, R. N., Hall, J., Parkinson, J. A. & Robbins, T. W. (2000). Differential involvement of amygdala subsystems in appetitive conditioning and drug addiction. In J. P. Aggleton (Ed.), *The amygdala: A functional analysis* (2nd ed., pp. 353–390). Oxford: Oxford University Press.

Fridlund, A. J. (1994). *Human facial expression: An evolutionary view*. San Diego: Academic Press.

Frijda, N. H. (1986). *The Emotions*. Cambridge: Cambridge University Press.

Goldman-Rakic, P. S. (1996). The prefrontal landscape: Implications of functional architecture for understanding human mentation and the central executive. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*, 1445–1453.

Gray, J. A. (1975). *Elements of a two-process theory of learning*. London: Academic Press.

Gray, J. A. (1987). *The psychology of fear and stress* (2nd ed.). Cambridge: Cambridge University Press.

Hornak, J., Bramham, J., Rolls, E. T., Morris, R. G., O'Doherty, J., Bullock, P. R., & Polkey, C.E. (2003a). Changes in emotion after circumscribed surgical lesions of the orbitofrontal and cingulate cortices. *Brain*, *126*, 1691–1712.

Hornak, J., O'Doherty, J., Bramham, J., Rolls, E. T., Morris, R. G., Bullock, P. R., & Polkey, C. E. (2004). Reward-related reversal learning after surgical exci-

sions in orbitofrontal and dorsolateral prefrontal cortex in humans. *Journal of Cognitive Neuroscience*, *16*, 463–478.

Hornak, J., Rolls, E. T., & Wade, D. (1996). Face and voice expression identification in patients with emotional and behavioural changes following ventral frontal lobe damage. *Neuropsychologia*, *34*, 247–261.

Izard, C. E. (1991). *The psychology of emotions*. New York: Plenum.

Krebs, J. R., & Kacelnik, A. (1991). Decision making. In J. R. Krebs & N. B. Davies (Eds.), *Behavioural ecology* (3rd ed., pp. 105–136). Oxford: Blackwell.

Kringelbach, M. L., & Rolls, E. T. (2003). Neural correlates of rapid reversal learning in a simple model of human social interaction. *Neuroimage*, *20*, 1371–1383.

Lazarus, R. S. (1991). *Emotion and adaptation*. New York: Oxford University Press.

Leak, G. K., & Christopher, S. B. (1982). Freudian psychoanalysis and sociobiology: A synthesis. *American Psychologist*, *37*, 313–322.

LeDoux, J. (2000). The amygdala and emotion: A view through fear. In J. P. Aggleton (Ed.), *The amygdala: A functional analysis* (2nd ed., pp. 289–310). Oxford: Oxford University Press.

LeDoux, J. E. (1996). *The emotional brain*. New York: Simon & Schuster.

Mackintosh, N. J. (1983). *Conditioning and associative learning*. Oxford: Oxford University Press.

Millenson, J. R. (1967). *Principles of behavioral analysis*. New York: MacMillan.

Moniz, E. (1936). *Tentatives opératoires dans le traitement de certaines psychoses*. Paris: Masson.

Nesse, R. M., & Lloyd, A. T. (1992). The evolution of psychodynamic mechanisms. In J. H. Barlow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind* (pp. 601–624). New York: Oxford University Press.

Oatley, K., & Jenkins, J. M. (1996). *Understanding emotions*. Oxford: Blackwell.

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience*, *4*, 95–102.

Petrides, M. (1996). Specialized systems for the processing of mnemonic information within the primate frontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*, 1455–1462.

Rolls, B. J., & Rolls, E. T. (1973). Effects of lesions in the basolateral amygdala on fluid intake in the rat. *Journal of Comparative and Physiological Psychology*, *83*, 240–247.

Rolls, E. T. (1986a), Neural systems involved in emotion in primates. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research, and experience. Biological foundations of emotion* (Vol. 3, pp. 125–143). New York: Academic Press.

Rolls, E. T. (1986b). A theory of emotion, and its application to understanding the neural basis of emotion. In Y. Oomura (Ed.), *Emotions. Neural and chemical control* (pp. 325–344). Tokyo: Japan Scientific Societies Press.

Rolls, E. T. (1990). A theory of emotion, and its application to understanding the neural basis of emotion. *Cognition and Emotion*, *4*, 161–190.

Rolls, E. T. (1999a). *The brain and emotion*. Oxford: Oxford University Press.
Rolls, E. T. (1999b). The functions of the orbitofrontal cortex. *Neurocase*, *5*, 301–312.
Rolls, E. T. (2000a). Précis of the brain and emotion. *Behavioral and Brain Sciences*, *23*, 177–234.
Rolls, E. T. (2000b). The orbitofrontal cortex and reward. *Cerebral Cortex*, *10*, 284–294.
Rolls, E. T. (2000c). Neurophysiology and functions of the primate amygdala, and the neural basis of emotion. In J. P. Aggleton (Ed.), *The amygdala: A functional analysis* (2nd ed., pp. 447–478). Oxford: Oxford University Press.
Rolls, E. T. (2002). The functions of the orbitofrontal cortex. In D. T. Stuss & R. T. Knight (Eds.), *Principles of frontal lobe function* (pp. 354–375). New York: Oxford University Press.
Rolls, E. T. (2004). A higher order syntactic thought (HOST) theory of consciousness. In R. J. Gennaro (Ed.), *Higher order theories of consciousness* (chap. 7, pp. 137–172). Amsterdam: John Benjamins.
Rolls, E. T., & Deco, G. (2002). *Computational neuroscience of vision*. Oxford: Oxford University Press.
Rolls, E. T., Hornak, J., Wade, D., & McGrath, J. (1994). Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *Journal of Neurology, Neurosurgery and Psychiatry*, *57*, 1518–1524.
Rolls, E. T., & Stringer, S. M. (2000). On the design of neural networks in the brain by genetic evolution. *Progress in Neurobiology*, *61*, 557–579.
Rolls, E. T., & Stringer, S. M. (2001). A model of the interaction between mood and memory. *Network*, *12*, 89–109.
Rolls, E. T., & Treves, A. (1998). *Neural networks and brain function*. Oxford: Oxford University Press.
Shallice, T., & Burgess, P. (1996). The domain of supervisory processes and temporal organization of behaviour. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*, 1405–1411.
Strongman, K. T. (1996). *The psychology of emotion* (4th ed.). Chichester: Wiley.
Thorpe, S. J., Rolls, E. T., & Maddison, S. (1983). The orbitofrontal cortex: Neuronal activity in the behaving monkey. *Experimental Brain Research*, *49*, 93–115.
Tinbergen, N. (1951). *The study of instinct*. Oxford: Clarendon.
Treves, A., & Rolls, E. T. (1994). A computational analysis of the role of the hippocampus in memory. *Hippocampus*, *4*, 374–391.
Trivers, R. L. (1976). Foreword. In R. Dawkins (Author), *The selfish gene*. Oxford: Oxford University Press.
Trivers, R. L. (1985). *Social evolution*. Menlo Park, CA: Benjamin/Cummings.
Valenstein, E. S. (1974). *Brain control. A critical examination of brain stimulation and psychosurgery*. New York: Wiley.
Weiskrantz, L. (1968). Emotion. In L. Weiskrantz (Ed.), *Analysis of behavioural change* (pp. 50–90). New York: Harper & Row.

6 How Do We Decipher Others' Minds?

MARC JEANNEROD

The central issue of how we access the mental contents of other individuals can be grounded in the concept of "self," both the narrative self who knows who we are, where we are, what we are presently doing, and what we were doing before, and the embodied self which is bound to particular bodily events, like actions. This chapter emphasizes communication between embodied selves, operating at a subpersonal level outside the awareness and conscious strategies of the two selves. We will show how mental states of others can be accessed through mind reading, a classical account of which is the simulation theory which holds that we exploit our own psychological responses in order to simulate others' minds. We first describe experiments that provide support to the notion of simulation from outside the realm of communication, stressing how the self's representation of its own actions are reflected in terms of changes in brain activity. We then extend the notion of simulation to the observation of others—and then show that this mechanism is not immune to misattribution of mental states in either direction, i.e., self attributing mental states of others as well as attributing to others one's own mental states.

The aim of this chapter is to understand how we access the mental contents of other individuals. People generate intentions, have goals, and feel emotions and affects. It is essential for each of us to penetrate the internal world of others, particularly when their intentions or goals are

directed to us or when their emotions relate to us. This very fact of knowing that one is the subject of others' mental states (that one is what other people think about) is a critical condition for fully human communication between individuals.

There are a few preliminary queries to answer before discussing the problem of communication between individuals. The first query is about ourselves: "What makes us self-conscious?" or "What makes us such that we can consciously refer to ourselves as that particular self, different from other selves?" There are several ways to answer this question, according to the level at which one considers the idea of a self. One of these levels is that of the narrative self. As a narrator, we obviously know who we are, where we are, what we are presently doing, and what we were doing before. Unless we become demented or amnesic, we have a strong feeling of continuity in our conscious experience. We rely on declarative memory systems where souvenirs (albeit distorted) can be retrieved and used as material for verbalization or imagination. Another level is that of the embodied self. We recognize ourselves as the owner of a body and the author of actions. At variance with the narrative self, the type of self-consciousness that is linked to the experience of the embodied self is discontinuous: it operates on a moment-to-moment basis as it is bound to particular bodily events, like actions. Instead of explicitly answering questions like "Who am I?" (something that the narrative self needs to know permanently), the embodied self will answer questions like "Is this mine?" or "Did I do this?"—questions to which we rarely care to give an explicit response. In other words, the embodied self mostly carries an implicit mode of self-consciousness, whereby self-consciousness is around but becomes manifest only when required by the situation. The related information has a short life span and usually does not survive the bodily event for very long.

The second question that has to be answered as a preliminary to the discussion about communication is actually related to the first one: "Which level of conscious experience are we considering for discussing communication with other individuals?" By keeping a parallel with the above distinction between a narrative level and an embodied level of the self, one could propose that communication between individuals can be established at either level. An act of communication between narrative selves commonly uses a verbal approach, for example, "What are you going to do?" or "What do you think?" or "Do you love me?" In other words, a narrative self aims at establishing communication with a narrative other. He or she uses a rational way of putting together available information and building a narrative structure about the other person's experience. By contrast, an act of communication between embodied selves operates at a subpersonal level outside the awareness and conscious strategies of the two selves. In this mode of communication, the two selves establish contact to the extent that their mental

states are embodied (i.e., transcribed into bodily states) and to the extent that their intentions, feelings, emotions, and attitudes can be read by an external observer.

In this chapter, emphasis will be clearly put on communication between selves at the embodied level. We will show how mental states of others can be accessed through *mind reading*, a general human ability for understanding other minds with the purpose of establishing communication with them. From a philosophical point of view, a classical account of mind reading is the simulation theory. Accordingly, it is thought that we exploit our own psychological responses in order to simulate others' minds or, in other words, that we internally simulate others mental states in our own mind. The outcome of this simulation process provides us with information about how others think or feel by reading our own mind (Goldie, 1999; for a full account of the philosophical issues raised by the simulation theory, see Davies & Stone, 1995).

We will first describe experiments that support the notion of simulation from a *solipsist* point of view, i.e., outside the realm of communication with others. The reason for this choice is that most of the empirical arguments for the simulation theory have been developed on the basis of how a subject represents his or her own actions to him- or herself and, more specifically, how the representation of actions reflects changes in brain activity. We will extend the notion of simulation to the observation of others on the basis of more recent experimental data which suggest that actions and emotions of others can be represented by an observer to the same extent as he or she represents his or her own actions. Finally, we will see that this mechanism is not immune to errors of identification: simulation of one's own mind or of the minds of other individuals can yield to misattribution of mental states in either direction, i.e., self-attribution of the mental states of others as well as attribution to others of one's own mental states.

THE SIMULATION THEORY IN THE SOLIPSIST CONTEXT

The simulation theory postulates that covert actions are in fact actions in their own right, except for the fact that they are not executed. Covert and overt stages represent a continuum such that every overtly executed action implies the existence of a covert stage, whereas a covert action does not necessarily turn into an overt action. As will be argued below, most of the neural events which lead to an overt action already seem to be present in the covert stages of that action. The theory therefore predicts a close similarity, in neural terms, of the state where an action is internally simulated and the state which precedes execution of that action (Jeannerod, 1994).

Specific methods, partly based on introspection but also relying on changes of physiological variables, have been designed to experimentally access these mental states characterized by absence or paucity of overt behavior. One of the most extensively studied of these representational aspects of action is mental motor imagery. Behavioral studies of motor imagery have revealed that motor images retain the same temporal characteristics as the corresponding real action when it comes to execution. For example, it takes the same time to mentally "walk" to a prespecified target as it takes to actually walk to the same place (Decety, Jeannerod, & Prablanc, 1989). Similarly, temporal regularities which are observed in executed actions, such as the classical speed–accuracy tradeoff, are retained in their covert counterparts (Sirigu et al., 1996). Along the same line, other situations have been described where the subject uses a motor imagery strategy in spite of the fact that no conscious image is formed. Those are situations where the subject is requested to make a perceptually based "motor" decision. Consider, for example, the situation where a subject is simply requested to make an estimate about the feasibility of an action, like determining the feasibility of grasping an object placed at different orientations: the time to give the response will be a function of the object's orientation, suggesting that the arm has to be mentally moved to an appropriate position before the response can be given. Indeed, the time to make this estimate is closely similar to the time it takes to actually reach and grasp an object placed at the same orientation (Frak, Paulignan, & Jeannerod, 2001; see also Parsons, 1994). One may speculate whether the same isochrony would also exist for performing an action with a disembodied artifact (e.g., a car) and mentally estimating its consequences. The question would be whether one can simulate an action performed, not by a human body, but with a mechanical device. A tentative answer will be given below.

This indication of a similar temporal structure for executed and nonexecuted actions by a biological system is reinforced by a similarity at the level of physiological indicators. Examining autonomic activity in subjects imagining an action at different effort rates reveals changes in heart rate and respiration frequency proportional to the imagined effort in the absence of any metabolic need. These results (Decety, Jeannerod, Durozard, & Baverel, 1993, see review in Jeannerod, 1995) reveal the existence of a central patterning of vegetative commands during covert actions, which would parallel the preparation of muscular commands. Autonomic changes occurring during motor imagery are closely related to those observed during central preparation of an effortful action (Krogh & Lindhard, 1913). Those are mechanisms that anticipate forthcoming metabolic needs, with the function of shortening the intrinsic delay required for heart and respiration to adapt to effort (e.g., Adams, Guz, Innes, & Murphy, 1987).

Interestingly, a similar involvement of autonomic mechanisms has been observed in the context of emotions. Lang (1979) proposed that emotional imagery can be analyzed objectively as a product of information processed by the brain and that this processing can be defined by measurable outputs. Indeed, experimental findings similar to those described for motor imagery have been reported with emotional imagery. Levenson, Ekman, & Friesen (1990), for example, showed that imagining or mimicking an emotional state induces in the subject the appearance of physiological reactions specific for the imagined or mimicked emotion (Chapter 2 [Adolph] for a review).

SIMULATING OTHERS' MINDS

Mental imagery is only one of the forms an action or an emotion representation can take. In this section, another form of representation is described, which relates to social interaction between people. Following the simulation hypothesis laid down in the first section, we will develop the idea that the mechanism for understanding the actions and emotions of other selves can be conceived as an extension of the mechanism of oneself having intentions and feeling emotions. We will first describe the conditions for bodily movements and expressions to be recognized as actions and emotions, respectively. Then, we will discuss the advantages and limitations of the simulation theory in explaining how we understand others.

Conditions for Action and Emotion Recognition

What makes an action performed by a living being (a *biological action*) so attractive for a human observer? What are the conditions that have to be fulfilled for a visual stimulus to be treated as a biologically significant action or emotional expression? Consider, for example, the classical experiments of Johansson in the early 1970s. He equipped a human actor with small lights placed at the level of his trunk and limb joints. The actor was moving in complete darkness, except for the small lights. The actor's movements (e.g., walking or dancing) are immediately recognizable by an observer, even though the actor's body cannot be seen. Visual information reduced to the trajectories and kinematics of the actor's movements is sufficient to provide cues not only to the activity portrayed by the actor but also to his age and sex (Johansson, 1973). A display of the same, but stationary, lights will not provide any recognizable information. Very young infants also easily distinguish biological movements from motions produced by mechanical devices, (Dasser, Ulbaek, & Premack, 1989).

Movements performed by living organisms owe their specificity to the fact that they usually have a goal. As a consequence, they display a number of kinematic properties that reveal their "intentional" origin. One of these properties is that goal-directed movements have an asymmetrical kinematic profile—a fast acceleration followed by a much longer deceleration—as opposed to the symmetrical profile of the ballistic motion of a projectile, for example. Another property is that the tangential velocity of the moving limb varies with the radius of curvature of the movement (Lacquaniti, Terzuolo, & Viviani, 1983). A further characteristic of biological movements is that they follow biomechanically compatible trajectories. Consider the perceptual effect produced by fast sequential presentation of pictures of an actor with an arm at two different postures. This alternated presentation is perceived as a continuous apparent movement between the two arm postures. If, however, the presentation of the two postures is such that the arm should go across an obstacle (e.g., another body part), then the apparent movement is perceived as going around and not across the obstacle. This striking effect (Shiffrar & Freyd, 1990) reflects the implicit representation built from visual perception of motion when it refers to a biological (or intentional) origin. Obviously, this is not to say that a robot could not be programmed for accurately reaching a goal with a different strategy (e.g., using movements with a symmetrical velocity profile or violating biomechanical constraints): these movements would simply look "unnatural" and would not match the internal representation that a human subject has of an intentional movement.

As a matter of fact, a normal subject cannot depart from the relation between geometry and kinematics which characterizes biological action: he or she cannot track a target moving with a spatiotemporal pattern different from the biological one (e.g., accelerating rather than decelerating in the curves). According to Viviani (1990), the subject's movements during the attempts to track the target "continue to bear the imprint of the general principle of organization for spontaneous movements, even though this is in contrast with the specifications of the target." Interestingly, the same relation between velocity and curvature is also present in a subject's perceptual estimation of the shape of the trajectory of a luminous target. A target moving at a uniform velocity is paradoxically seen as moving in a nonuniform way and, conversely, a kinematic structure which respects the above velocity–curvature relation is the condition for perceiving a movement at a uniform velocity. According to Viviani and Stucchi (1992), perception is constrained by the implicit knowledge that the central nervous system has concerning the movements that it is capable of producing. In other words, there is a central representation of what a uniform movement should be, and this representation influences visual perception. Whether this implicit knowledge is a result of learning (e.g., by imitation) or an effect of some innate

property of visual perception is a matter of speculation. The fact that young infants are more interested by movements that look biological than by those that look mechanical (e.g., Dasser, Ulbaek, & Premack, 1989) is an indication in favor of the latter. Another argument is the fact that intentionality of biological movements can be mimicked by the motion of objects, provided this motion obeys certain rules. As shown by Heider and Simmel (1944), seeing the self-propelled motion of geometrical stimuli can trigger judgments of protosocial goals and intentions. The main condition is that the object motion appears to be internally caused rather than caused by an external force. A preference for self-propelled motion can be demonstrated with this type of stimuli in 6-month-old infants (Gergely, Nadasdy, Czibra, & Biro, 1995; Czibra et al., 1999).

Another critical aspect of communication between individuals is the face-perception system. Faces, particularly in humans, carry an essential aspect of the expression of emotions. Humans have a rich repertoire of facial gestures: the eyes, the eyebrows, the forehead, the lips, the tongue, and the jaws can move relative to the rest of the face. Not only can lip, tongue, and jaw movements convey a speaker's communicative intentions, but mouth movements and lip positions can be powerful visual cues of a person's emotional states: by opening the mouth and moving the lips, a person can display a neutral face, smile, laugh, or express grief. The movements and the position of the eyes in their orbits also convey information about the person's emotional state, the likely target of attention and/or intention. To the same extent as discussed for the perception of biological actions, the perception of emotional expression on faces seems to stimulate a system tuned to extract specific features of the visual stimulus. According to the influential model of Bruce and Young (1986), a human face can give rise to two sorts of perceptual process: perception of the invariant aspects and of the changing aspects of a face. The former contributes to the recognition of the identity of a person. The latter contributes to the perception of another's social intentions and emotional states.

The visual processing of face patterns has been a topic of considerable interest for the past three decades. The neuropsychological investigation of the condition known as "prosopagnosia" has revealed that patients with damage to the inferior occipitotemporal region are selectively impaired in visual face recognition, while their perception and recognition of other objects are relatively unimpaired. As emphasized by Haxby, Hoffman, and Gobbini (2000), face processing is mediated by a distributed neural system that includes three bilateral regions in the occipitotemporal extrastriate cortex: the inferior occipital gyrus, the lateral fusiform gyrus, and the superior temporal sulcus. There is growing evidence that the lateral fusiform gyrus might be specially involved in identifying and recognizing faces, that is, in the

processing of invariant aspects of faces (e.g., Kanwisher, McDermott, & Chun, 1997). By contrast, the superior temporal sulcus might be more involved in processing variable aspects of faces, those that carry emotional expressions (Hoffman & Haxby, 2000).

Empathy Revisited

The ability to recognize biologically significant actions and expressions and to exploit this information for communication between individuals can operate at different levels. At the beginning of this chapter, arguments were presented for the choice of focusing our attention on recognition of other selves at the level of embodied selves, as opposed to narrative selves. Yet, within this limitation, there are still several possibilities for thinking of others' minds. Following Goldie (1999), we will describe two of those, contagion and empathy, which can both be interpreted in terms of the simulation of mental states but with different contents.

Actions and emotions are contagious; they can be caught like colds. Suffering from a contagious emotion transmitted by another individual, however, is not a sufficient condition for understanding this individual's mental states: it provides the information that the person one sees is enacting a certain type of behavior or experiencing a certain type of emotion, but it does not tell what the action is or what the emotion is about. In the realm of action recognition, a concept developed during the 19th century, ideomotor action (Lotze, 1852), seems close to the concept of contagion of emotions. *Ideomotor action* accounts for the familiar observation that people tend to perform the movements they see. This phenomenon is particularly observed in situations with an emotional content, where the observer feels strongly involved (e.g., watching sports actions). It has been argued that ideomotor action is a direct mapping of perceived movements onto the corresponding motor output. This possibility may sound familiar to those who adhere to the so-called direct perception–action transformation heralded by the Gibsonian school (see Jeannerod, 1993, for review). It has been used to explain not only interactions with the visual environment (e.g., steering locomotion, avoiding obstacles, maintaining posture) but also relationships between selves (Neisser, 1993). In this conception, ideomotor phenomena could represent a form of compulsive imitation where the subject cannot refrain from reproducing the perceived performance. Examples of this sort are contagious yawning or laughing. This could also be the case for resonant behavior observed in other species (e.g., wing flapping in bird flocks). True imitation, by contrast, would have the additional property of not being bound

to the observed action and, thus, to having the possibility of being delayed with respect to the observed action.

In the context of emotions, arguments for dissociating the recognition of an emotion (as in contagion) from the recognition of the cause of that emotion (and the appropriate response to be given to it) are twofold. First, very young babies (at around 5 months of age) are able to recognize that a person is expressing an emotion, although they do not seem to experience this emotion themselves. It is only later (at 2 years of age) that they begin to respond to the emotions expressed by other people (e.g., Nelson, 1987). In other words, it is possible to recognize an emotion without having it. The second argument is that pure contagion would not yield an appropriate response. Imagine facing somebody who expresses anger and threat. The appropriate response is not to experience anger oneself but to experience fear and perhaps to run away. Because contagion cannot tell what the emotion is about, it cannot be used as a useful means for reacting to others' emotions. Contagion of action and emotion would act as a relatively primitive form of recognition of others' behavior, with a role in activating imitation mechanisms, particularly in young infants, a behavior that becomes progressively inhibited at later stages. Pathology offers an example of disinhibition of compulsive imitation in patients with frontal lobe lesions (Lhermitte, 1983).

Another way of getting close to others' minds is empathy. The concept of *empathy* is more clearly related to the idea of simulation than contagion in that it implies that individuals involved in a given interaction share a similar mental state. In contradistinction to contagion, empathy requires that one has information on the experience and intentions of the person who is observed and whose mental content one is attempting to understand. Observing a person about whose intentions, thoughts, or experience one has no information would provide only a limited access to what the person is experiencing or doing, insufficient for empathizing with that person (this is what Goldie, 1999, calls "in his shoes imagining"). Consider, for example, the two characters described in the experiment of Kahneman and Tversky (1982). These two persons arrive independently at the airport 30 minutes late. To one of them, one explains that his plane left 30 minutes ago; one tells the other that his plane was delayed and left only 5 minutes ago, too late for him to catch it. The experiment consists in asking subjects which of the two characters would be more frustrated. The response of most subjects is that the second one will be more frustrated. As there is no rationale for that answer, the way subjects proceed to give it is to place themselves in the shoes of the person and to feel what it would be like to miss a plane by only 5 minutes. Obviously, subjects would give more circumstantial answers if they could empathize more fully with the characters, that is, if they knew

more about them, their reason for taking the plane (e.g., going on a business trip or on vacation), their mood, etc.

Originally, the term *empathy* was translated by Titchener (1908) from the German term *Einfühlung*. In the texts of 19th-century German philosophers and psychologists, like Theodor Lipps (1903), *Einfühlung* was used to designate an implicit process of knowledge, different from the rational mode of knowledge, which gave access to the esthetic or the emotional content of a situation. The viewer of a painting or the listener to a piece of music, for example, resented its beauty not through an act of perception but, rather, through a modification of his or her own emotional state. The same concept of empathy eventually became used in the context of clinical psychology (see Pigman, 1995). Lipps considered empathy to be the source of knowledge about other individuals, to the same extent as sensory perception is the source of knowledge about objects. His idea was that we understand, for example, facial expressions displayed by other persons not because we compare them with our own expressions, which we cannot see, but because the vision of an expression on the face of someone else "awakens [in the observer] impulses to such movements that are suited to call just this expression into existence" (Lipps, 1903, p. 193; see Pigman, 1995). Note that empathy, as defined here by Lipps, fits the concept of communication between individuals at the level of embodied selves rather than of narrative selves, a distinction which was the starting point of this paper.

The concept of empathy, and its consequences for behavior, is taken by the simulation theorists as equivalent to *mind reading*, the ability for normal people to understand and predict the behavior of their conspecifics, which we mentioned earlier. Gallese and Goldman (1998) proceeded one step further in proposing an explanation of mind reading in terms of brain mechanisms: they proposed that the observed behavior would activate, in the observer's brain, the same mechanisms that would be activated were that behavior intended or imagined by the observer. They state that "when one is observing the action of another, one undergoes a neural event that is qualitatively the same as [the] event that triggers actual movement in the observed agent" and, thus, "a mind-reader represents an actor's behavior by recreating in himself the plans or movement intentions of the actor." Gallese and Goldman's view is close to that heralded in the field of linguistics to account for perception of speech by a listener: the general idea is that the listener would implicitly repeat the auditory message and access the spoken message via a subliminal activation of his neural and muscular speech mechanisms (the so-called motor theory of speech perception; Liberman & Mattingly, 1985). Thus, the simulation theory would encompass a number of propositions coming from different fields, with the common aim of explaining communication between people.

THE CONCEPT OF SHARED REPRESENTATIONS: A NEURAL BASIS FOR THE SIMULATION THEORY

Little is known of the biological effects of observing someone's actions and emotions. However, extending the simulation theory to understanding the representations underlying actions and emotions of other people requires that a continuity be established between the embodiment of representations of the observer and those of the agent being observed.

Consider a simple experiment with normal subjects observing the action of an actor. The subjects, equipped for recording their respiration rate, sit in front of a large screen on which they see an actor performing an effortful action. The actor stands on a treadmill that either is motionless, moves at a constant velocity (2.5, 7, or 10 km/h), or progressively accelerates from 0 to 10 km/h over 1 minute. The main result of this experiment (Paccalin & Jeannerod, 2000) is that the respiration rate increased during the observation of the actor walking or running at an increasing speed. Typically, the average increase during observation of the actor running at 10 km/h is about 25% above the resting level. Further, the increase in respiration frequency correlates with running velocity. Watching an action is thus different from watching a visual scene with moving objects. While watching an action, the observer is not only seeing visual motion but also internally (and nonconsciously) simulating (or rehearsing) the action. Simulating accelerated running implies an increase in the breathing rate because, if the running movements were actually executed, they would require an anticipatory increase in metabolic needs. This finding thus substantiates the hypothesis that perceiving an action triggers a neural state where the neural structures potentially involved in executing that action are facilitated (see details below).

Now, consider a similar experiment using the same paradigm of observation of an actor. However, instead of performing a neutral action like running, the actor displays an emotional state. Imagine an observer in front of a screen, watching the actor's face display different degrees of an emotion, like joy. Joy would be expressed on that face by a set of expressions ranging from a subtle smile to excitation and laughing. Also imagine that vegetative indices (e.g., heart rate, galvanic skin response, breathing) are recorded from the observer. What would be the conclusion to be drawn from that experiment if, say, the respiration rate or the heart rate of the observer increased as a function of the degree of joy expressed by the face of the actor? According to the conclusion drawn from the action observation experiment, the conclusion here should be that the observer is simulating the emotion displayed by the actor and that this simulated emotion activates his or her own vegetative system to the same extent as when actually experiencing the emotion. Although this particular experiment may not have been performed (but

see Zajonc, 1985), the imaginary results proposed here seem highly plausible. We will come back later to other experiments that strongly support the present speculation. Experiencing and watching (an action, an emotion) would thus be two faces of the same coin. The problem with such examples, however, is that they remain within the realm of nonintentional communication and explore only some basic conditions for understanding the observed agent. This limitation, which also extends to the brain-mapping experiments to be described below, has to be kept in mind for evaluating the relevance of the simulation theory as an explanation for understanding other people's minds.

A critical condition for assigning motor images and observed actions the status of covert and simulated actions is that they should activate brain areas known to be devoted to executing actions. Early work by Ingvar and Philipsson (1977), using measurement of local cerebral blood flow, had showed that "pure motor ideation" (e.g., thinking of rhythmic clenching movements) produced a marked frontal activation and a more limited activation in the area of the motor cortex. More recent brain mapping experiments, using positron emission tomography (PET) or functional magnetic resonance imaging (fMRI), have led to the conclusion that represented actions involve a subliminal activation of the motor system (Jeannerod, 1999, 2001; Jeannerod & Frak, 1999). They show the existence of a cortical and subcortical network activated during both motor imagery and action observation. This network involves structures directly concerned with motor execution, such as the motor cortex, dorsal and ventral premotor cortex, lateral cerebellum, and basal ganglia; it also involves areas concerned with action planning, such as the dorsolateral prefrontal cortex and posterior parietal cortex. Concerning the primary motor cortex itself, fMRI studies unambiguously demonstrate that voxels activated during contraction of a muscle are also activated during imagery of a movement involving the same muscle (Roth et al., 1996). During action observation, the involvement of primary motor pathways was demonstrated using direct measurement of corticospinal excitability by transcranial magnetic stimulation. Fadiga, Fogassi, Pavesi, and Rizzolatti (1995) found that subjects observing an actor executing hand-grasping movements were more responsive to stimulation in their own hand motor area. The area involved during observation of the hand movements was superimposed with that activated while the subjects themselves actually performed the movement.

In principle, a theory that postulates that both actions of the self and actions of the other can be distinguished on the basis of their central representations should predict separate representations for these two types of action. At the neural level, one should expect the existence of different networks devoted to action recognition, whether the action originates from

the self or not. One network would be involved with recognizing actions as belonging to the self, and another would correspond to attributing actions to another person. We know from the results described above that brain areas activated during self-produced actions (executed or not) and when observing actions of other people partly overlap: this is the basis for the concept of *shared representations*, introduced by Daprati et al. (1997) and Georgieff and Jeannerod (1998), according to which different mental states concerning actions (e.g., intending an action and observing it from another person) share the same neural representations yet still have distinct patterns of neural activity.

To clarify this concept, let us briefly describe experimental results obtained from monkeys. A dramatic example of a shared representation is offered by the finding of mirror neurons (Rizzolatti, Fadiga, Gallese, & Fogassi, 1995). Mirror neurons were identified in the monkey premotor cortex. They are activated in two conditions: first, they fire when the animal is involved in a specific motor action, like picking a piece of food with a precision grip; second, they also fire when the immobile animal watches the same action performed by an external agent (another monkey or an experimenter). In other words, mirror neurons represent one particular type of action, irrespective of the agent who performs it. At this point, it could be suspected that the signal produced by these neurons, and exploited by other elements downstream in the information-processing flow, would be the same for actions performed by the self and by another agent: the two modalities of that action (executed and observed) would thus have the same neural representation. The problem of actor identification, however, is solved by the fact that other premotor neurons (the canonical neurons), and presumably many other neuron populations, fire only when the monkey performs the action and not when it observes it from another agent. This is indeed another critical feature of the shared representations concept: representations overlap only partially, and the nonoverlapping part of a given representation can be the cue for attributing the action to the self or to the other. The same mechanism operates in humans. Brain activity during different conditions where subjects were simulating actions (e.g., intending actions and preparing for execution, imagining actions, or observing actions performed by other people) was compared (Decety et al., 1994, 1997; Grafton, Arbib, Fadiga, & Rizzolatti, 1996; Rizzolatti et al., 1996; Gérardin et al., 2000). The outcome of these studies is twofold: first, there exists a cortical network common to all conditions, to which the inferior parietal lobule (areas 39 and 40), the ventral premotor area (ventral area 6), and part of supplementary motor area contribute; second, motor representations for each individual condition are clearly specified by the activation of cortical zones which do not overlap between conditions (Ruby & Decéty, 2001).

The question now arises about whether there are mirror neurons for the simulation of emotions. This question could only be answered positively if the neural structures activated in the brain of the observer could be found in areas devoted to the expression of emotions and not only in areas specialized for processing the perception of emotions. At present, empirical and clinical data do not entirely support this view. These data stress the fact that the amygdala, one of the main structures involved in recognition of emotions, is mostly devoted to processing emotional stimuli and to appraising emotional situations. Monkeys with large bilateral lesions including the amygdala and the temporal neocortex were still able to respond to visual stimuli, but their behavior was emotionally inappropriate: compulsive manipulation of objects, hypersexual behavior, tameness, and lack of emotional response to seeing aversive objects like snakes (the classical syndrome described by Kluwer & Bucy, 1939). Human pathological cases have also contributed to the study of this problem: patients with amygdala damage fail to properly extract emotional expression from faces and voices and misjudge the emotional states of other people (Adolphs et al., 1994, 1998). The fact that the impairment in emotion recognition is independent of the modality in which it is expressed suggests the existence of a polysensory representation of emotional stimuli in the amygdala. Finally, in normal subjects, PET and fMRI studies show that the amygdala is activated during presentation of visual or vocal expressions of fear (see review in Stone et al., 2003).

To recognize an emotion is thus more than simply to perceive a face or a voice with an emotional expression. The amygdala seems to be involved not in perceiving emotional states but, more generally, in inferring emotion from all relevant cues. Stone et al. (2003) found an activation of the amygdala in subjects hearing stories requiring inference about affective mental states. This result, obtained using verbal (as opposed to perceptual) information, reveals that the amygdala is involved in abstracting emotions from social situations, rather than merely decoding emotional signals arising from the external world.

FAILURE OF SELF-RECOGNITION MECHANISMS IN PATHOLOGICAL STATES

In this section, we investigate the effects of pathological conditions as another potential source of information concerning the mechanisms of self-recognition and recognition of others. Pathological conditions offer many examples of misidentification of the self: a typical case is that of schizophrenia.

The pattern of self misidentification in schizophrenic patients is twofold: first, patients may attribute their own actions or thoughts to others

rather than to themselves (*underattributions*); second, patients may attribute the actions or thoughts of others to themselves (*overattributions*). According to the French psychiatrist Pierre Janet (1937), these false attributions reflect the existence in each individual of a representation of others' actions and thoughts in addition to the representation of one's own actions and thoughts: false attributions were thus due to an imbalance between these two representations. A typical example of underattribution is hallucination. Hallucinating schizophrenic patients may show a tendency to project their own experience onto external events. Accordingly, they may misattribute their own intentions or actions to external agents. During auditory hallucinations, the patient will hear voices that are typically experienced as coming from a powerful external entity but in fact correspond to subvocal speech produced by the patient (Gould, 1949; David, 1994). The voices offer comments where the patient is addressed in the third person and which include commands and directions for action (Chadwick & Birchwood, 1994). The patient may declare that he or she is being acted upon by an alien force, as if his or her thoughts or acts were controlled by an external agent. The so-called mimetic behavior (where the patient compulsively imitates other people) observed at the acute stage of psychosis also relates to this category.

The reverse pattern of misattribution can also be observed. Overattributions were early described by Janet (1937): what this author called "excess of appropriation" corresponds for the patient to the illusion that actions of others are in fact initiated or performed by him- or herself and that he or she is influencing other people (the clinical picture of megalomania). In this case, patients are convinced that their intentions or actions can affect external events, for example, that they can influence the thought and actions of other people. Accordingly, they tend to misattribute the occurrence of external events to themselves. The consequence of this misinterpretation would be that external events are seen as the result expected from their own actions. More recently, impairments in the recognition of others and the self in schizophrenia have been categorized, together with other manifestations of this disease, among the so-called first-rank symptoms. According to Schneider (1955), these symptoms, which are considered critical for the diagnosis of schizophrenia, refer to a state where patients interpret their own thoughts or actions as due to alien forces or to other people and feel controlled or influenced by others. First-rank symptoms might reflect the disruption of a mechanism which normally generates consciousness of one's own actions and thoughts and allows their correct attribution to their author.

One possible explanation for these impairments in self-recognition could be the dysfunction of a specific system for perceiving, recognizing, and attributing actions. This hypothesis is supported by the fact that schizophrenic patients with delusion of influence make frequent errors in experimental

situations where they have to attribute a movement to its author. For example, the situation can be such that the movements patients execute with their own hands differ from the movements they see when shown the movements of a hand of an uncertain origin (a hand that could equally likely belong to them or to someone else). In this situation, schizophrenic patients tend to massively attribute these movements to themselves (Daprati et al., 1997). These attribution errors might be the consequence of an impairment in detecting some aspects of biological movements, like their direction (Franck et al., 2001). Perceiving the direction of a movement is indeed useful information for an observer to understand the action of the agent of this movement: during a movement, the arm points to the goal of the action, and its direction may reveal the intention of the agent. It is thus not surprising that a patient deprived of this information will misinterpret the intention displayed by others in their movements and that this will have consequences for attributing actions to their agent and ultimately understanding interactions between people.

Misrecognition of one's own movements is also a highly plausible explanation for another typical symptom that is of the first rank in Schneider's sense, verbal hallucinations. As mentioned above, auditory verbal hallucinations in schizophrenic patients are related to the production of their own speech: they perceive their inner speech as voices arising from an external source. Experiments using neuroimaging techniques have greatly contributed to the study of this problem by examining brain activation during hallucinations or during inner speech in patients predisposed to hallucinations and subjects experiencing no hallucinations. The results show that, during hallucinations (as signaled by the patients), brain metabolism is increased in the primary auditory cortex (Heschl's gyrus) on the left side (Dierks et al., 1999), as well as in the basal ganglia (Silbersweig et al., 1995). Thus, whereas self-generated inner speech is normally accompanied by a mechanism that decreases the responsiveness of the primary auditory cortex, during verbal hallucinations, the auditory temporal areas remain active, which suggests that the nervous system in these patients behaves as if it were actually processing the speech of an external speaker.

The obvious question at this point is whether patients of this sort would also fail to detect (and therefore to attribute) emotions expressed by other individuals. This seems to be confirmed by the abundant literature on the processing of emotion in schizophrenia. Schizophrenic patients typically fail to detect facial expressions of emotions (e.g., Feinberg, Rifkin, Schaffer, & Walker, 1986; see review in Baudouin et al., 2002). In addition, these patients also have decreased responsiveness to emotional stimuli: their observable facial expressiveness in response to emotional stimuli is decreased (e.g., Berenbaum & Oltmanns, 1992).

CONCLUSION

In this chapter, we have laid down a framework for integrating social cognition to the neural substrate. This conception of action and emotion recognition is based on the existence of neural networks subserving the various forms of representation of an act of communication. Accordingly, each representation entails a cortical/subcortical network of interconnected neural structures. According to the simulation theory, these networks become activated as a consequence of the simulation of the represented actions and emotions by the selves who are engaged in the act of communication. As we have argued, this theory relies on the fact that simulation of one's own actions (e.g., in motor images) or of others' actions produces a subliminal activation of some of the areas normally devoted to action execution (Jeannerod, 2001). Another feature of the theory is that, although these networks are clearly distinct from one form of representation to another (e.g., the representation of a self-generated action versus the representation of an action observed or predicted from another agent), they partly overlap. When two agents socially interact with one another, this overlap creates *shared representations* (i.e., activation of neural structures that are common to several modalities of representation). In normal conditions, however, the existence of nonoverlapping parts as well as possible differences in intensity of activation between the activated zones allow each agent to discriminate between representations activated from within (for a self-generated intention or emotional state) and those activated from without (by an action or an emotion displayed by another agent) and to disentangle which belongs to the self from which belongs to the other. This process would thus be the basis for correctly attributing a representation to the proper agent or, in other words, for answering the question of "who" is the author of the act of communication (Georgieff & Jeannerod, 1998; Jeannerod, 1999).

The flow chart of Figure 6.1 is a tentative illustration of the many interactions between two agents. Each agent builds in the brain a representation of both his or her own intended actions, using internal cues like beliefs and desires, and the potential actions of the other agent with whom he or she interacts. These partly overlapping representations are used by each agent to build a set of predictions and estimates about the social consequences of the represented actions, if and when they are executed. Indeed, when an action comes to execution, it is perceived by the other agent as a set of social signals which confirm (or not) predictions and possibly modify beliefs and desires.

This conception allows hypotheses about the nature of the dysfunction responsible for misattribution of actions by schizophrenic patients. Changes

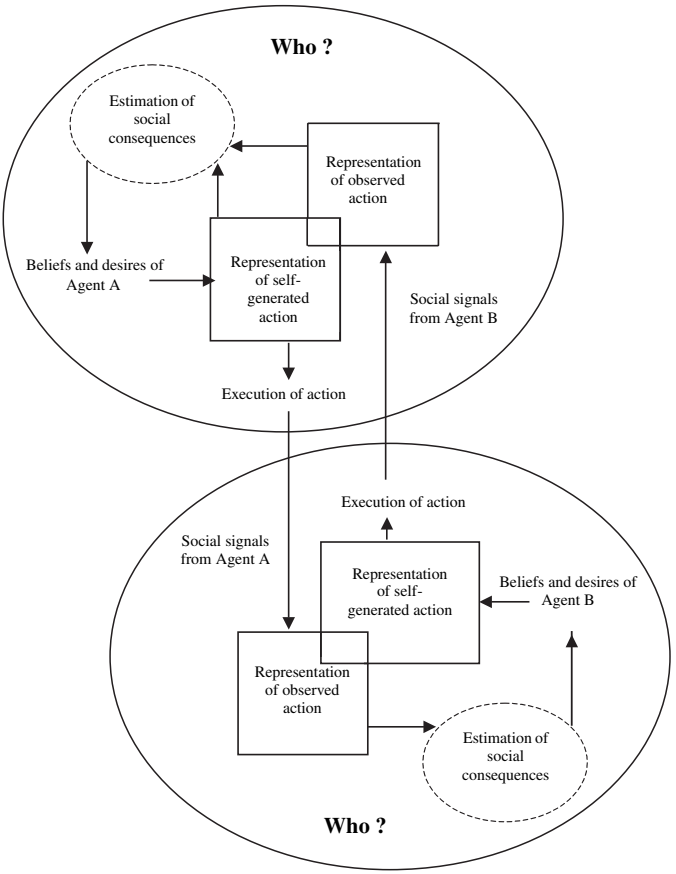


Figure 6.1. A tentative illustration of the many interactions between two agents. Each agent builds in the brain a representation of both his or her own intended actions, using internal cues like beliefs and desires, and the potential actions of the other agent. These partly overlapping representations are used by each agent to build a set of predictions and estimates about the social consequences of the represented actions, if and when they would be executed. When an action comes to execution, it is perceived by the other agent as a set of social signals which do or do not confirm predictions and possibly modify beliefs and desires.

in the pattern of cortical connectivity could alter the shape of the networks corresponding to different representations or the relative intensity of activation in the areas composing these networks. Although little is known on the functional aspects of cortical connectivity underlying the formation of these networks and, a fortiori, their dysfunction in schizophrenia, several studies have pointed to the prefrontal cortex as one of the possible sites for

perturbed activation (e.g., Weinberger & Berman, 1996). Because prefrontal areas normally exert an inhibitory control on other areas involved in various aspects of motor and sensorimotor processing, alteration of this control in schizophrenic patients might result in aberrant representations of actions and emotions. Referring to the diagram in Figure 6.1, one of the two agents would become "schizophrenic" if, due to an alteration in the pattern of connectivity of the corresponding networks, the degree of overlap between the representations in the brain increased in such a way that the representations would become indistinguishable from each other. The pattern of misattribution in this agent would be a direct consequence of this alteration: for example, decreased self attribution if frontal inhibition were too strong or increased if it were too weak.

References

Adams, L., Guz, A., Innes, J. A., & Murphy, K. (1987). The early circulatory and ventilatory response to voluntary and electrically induced exercise in man. *Journal of Physiology*, *383*, 19–30.

Adolphs, R., Tranel, D., & Damasio, A. R. (1998). The human amygdala in social judgment. *Nature*, *393*, 470–474.

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expression following bilateral damage to the human amygdala. *Nature*, *372*, 669–672.

Baudouin, J. Y., Martin, F., Tiberghien, G., Verlut, I., & Franck, N. (2002). Selective attention to facial emotion and identity in schizophrenia. *Neuropsychologia*, *40*, 503–511.

Berenbaum, H., & Oltmanns, T. F. (1992). Emotional experience and expression in schizophrenia and depression. *Journal of Abnormal Psychology*, *101*, 37–44.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*, 305–327.

Chadwick, P., & Birchwood, M. (1994). The omnipotence of voices. A cognitive approach to auditory hallucinations. *British Journal of Psychiatry*, *164*, 190–201.

Czibra, G., Gergely, G., Biro, S., Koos, O., & Brockbank, M. (1999). Goal attribution without agency cues: The perception of "pure reason" in infancy. *Cognition*, *72*, 237–267.

Daprati, E., Franck, N., Georgieff, N., Proust, J., Pacherie, E., Dalery, J., & Jeannerod, M. (1997). Looking for the agent. An investigation into consciousness of action and self-consciousness in schizophrenic patients. *Cognition*, *65*, 71–86.

Dasser, V., Ulbaek, I., & Premack, D. (1989). The perception of intention. *Science*, *243*, 365–367.

David, A. S. (1994). The neuropsychological origin of auditory hallucinations. In A. S. David & J. C. Cutting (Eds.), *The neuropsychology of schizophrenia* (pp. 269–313). Hove, UK: Erlbaum.

Davies, M., & Stone, T. (Eds.) (1995). *Folk psychology. The theory of mind debate*. Oxford: Blackwell.

Decety, J., Grezes, J., Costes, N., Perani, D., Jeannerod, M., Procyk, E., Grassi, F., & Fazio, F. (1997). Brain activity during observation of action. Influence of action content and subject's strategy. *Brain, 120*, 1763–1777.

Decety, J, Jeannerod, M., Durozard, D., & Baverel, G. (1993). Central activation of autonomic effectors during mental simulation of motor actions in man. *Journal of Physiology, 461*, 549–563.

Decety, J., Jeannerod, M., & Prablanc, C. (1989). The timing of mentally represented actions. *Behavioural Brain Research, 34*, 35–42.

Decety, J., Perani, D., Jeannerod, M., Bettinardi, V., Tadary, B., Woods, R., Mazziotta, J. C., & Fazio, F. (1994). Mapping motor representations with PET. *Nature, 371*, 600–602.

Dierks, T., Linden, D. E. J., Jandl, M., Formisano, E., Goebel, R., Lanferman, H., & Singer, W. (1999). Activation of the Heschl's gyrus during auditory hallucinations. *Neuron, 22*, 615–621.

Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation. A magnetic stimulation study. *Journal of Neurophysiology, 73*, 2608–2611.

Feinberg, T. E., Rifkin, A., Schaffer, C., & Walker, E. (1986). Facial discrimination and emotional recognition in schizophrenia and affective disorders. *Archives of General Psychiatry, 43*, 276–279.

Frak, V. G., Paulignan, Y., & Jeannerod, M. (2001). Orientation of the opposition axis in mentally simulated grasping. *Experimental Brain Research, 136*, 120–127.

Franck, N., Farrer, C., Georgieff, N., Marie-Cardine, M., Daléry, J. D'Amato, T., & Jeannerod, M. (2001). Defective recognition of one's own actions in schizophrenic patients. *American Journal of Psychiatry, 158*, 454–459.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Science, 12*, 493–501.

Georgieff, N., & Jeannerod, M. (1998). Beyond consciousness of external reality. A "who" system for consciousness of action and self-consciousness. *Consciousness and Cognition, 7*, 465–477.

Gérardin, E., Sirigu, A., Lehéricy, S., Poline, J.-B., Gaymard, B., Marsault, C., Agid, Y., & Le Bihan, D. (2000). Partially overlapping neural networks for real and imagined hand movements. *Cerebral Cortex, 10*, 1093–1104.

Gergely, G., Nadasdy, Z., Czibra, G., & Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*, 165–193.

Goldie, P. (1999). How we think of others' emotions. *Mind and Language, 14*, 394–423.

Gould, L. N. (1949). Auditory hallucinations in subvocal speech: Objective study in a case of schizophrenia. *Journal of Nervous and Mental Diseases, 109*, 418–427.

Grafton, S. T., Arbib, M. A., Fadiga, L., & Rizzolatti, G. (1996). Localization of grasp representations in humans by PET: 2. Observation compared with imagination. *Experimental Brain Research, 112*, 103–111.

Haxby, J. V., Hoffman, E. A., & Gobbini, I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Science*, *4*, 223–233.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*, 243–259.

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature America*, *3*, 80–84.

Ingvar, D., & Philipsson, L. (1977). Distribution of the cerebral blood flow in the dominant hemisphere during motor ideation and motor performance. *Annals of Neurology*, *2*, 230–237.

Janet, P. (1937). Les troubles de la personnalité sociale. *Annales Médico-Psychologiques*, *II*, 149–200.

Jeannerod, M. (1993). A theory of representation-driven actions. In U. Neisser (Ed.), *The perceived self: Ecological and interpersonal sources of self-knowledge* (pp. 89–101). Cambridge: Cambridge University Press.

Jeannerod, M. (1994). The representing brain. Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences*, *17*, 187–245.

Jeannerod, M. (1995). Mental imagery in the motor context. *Neuropsychologia*, 33, 1419–1432.

Jeannerod, M. (1999). To act or not to act: Perspectives on the representation of actions. *Quarterly Journal of Experimental Psychology*, *52A*, 1–29.

Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *Neuroimage*, *14*, S103–S109.

Jeannerod, M., & Frak, V. G. (1999). Mental simulation of action in human subjects. *Current Opinions in Neurobiology*, *9*, 735–739.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, *14*, 201–211.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In P. Slovic, D. Kahneman, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*, 4302–4311.

Kluwer, H., & Bucy, P. C. (1939). Preliminary analysis of the function of parietal lobes in monkeys. *Archives of Neurological Psychiatry*, *42*, 979–997.

Krogh, A., & Lindhard, J. (1913). The regulation of respiration and circulation during the initial stages of muscular work. *Journal of Physiology*, *47*, 112–136.

Lacquaniti, F., Terzuolo, C., & Viviani, P. (1983). The law relating kinematic and figural aspects of drawing movements. *Acta Psychologica*, *54*, 115–130.

Lang, P. (1979). A bioinformational theory of emotional imagery. *Psychophysiology*, *16*, 495–512.

Levenson, R. W., Ekman, P., & Friesen, W. V. (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology*, *27*, 363–384.

Lhermitte, F. (1983). Utilisation behaviour and its relation to lesions of the frontal lobes. *Brain, 106*, 237–255.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of perception of speech revisited. *Cognition, 21*, 1–36.

Lipps, T. (1903). *Aesthetik: Psychologie des schönen un der kunst*. Hamburg: Voss.

Lotze, R. H. (1852). *Medicinische Psychologie oder Physiologie der Seele*. Leipzig: Weidmann'sche Buchhandlung.

Neisser, U. (1993). The self perceived. In U. Neisser (Ed.), *The perceived self. Ecological and interpersonal sources of self-knowledge* (pp. 3–21). Cambridge: Cambridge University Press.

Nelson, C. A. (1987). The recognition of facial expression in the first two years of life. Mechanisms of development. *Child Development, 58*, 889–909.

Paccalin, C., & Jeannerod, M. (2000). Changes in breathing during observation of effortful actions. *Brain Research, 862*, 194–200.

Parsons, L. M. (1994). Temporal and kinematic properties of motor behavior reflected in mentally simulated action. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 709–730.

Pigman, G. W. (1995). Freud and the history of empathy. *International Journal of Psycho-Analysis, 76*, 237–256.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1995). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*, 131–141.

Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D., & Fazio, G. (1996). Localization of grasp representations in humans by PET: 1. Observation versus execution. *Experimental Brain Research, 111*, 246–252.

Roth, M., Decety, J., Raybaudi, M., Massarelli, R., Delon-Martin, C., Segebarth, C., Morand, S., Gemignani, A., Décorps, M., & Jeannerod, M. (1996). Possible involvement of primary motor cortex in mentally simulated movement. A functional magnetic resonance imaging study. *Neuroreport, 7*, 1280–1284.

Ruby, P., & Decéty, J. (2001). Effect of subjective perspective taking during simulation of action: A PET investigation of agency. *Nature Neurosciences, 4*, 546–550.

Schneider, K. (1955). *Klinische Psychopathologie*. Stuttgart: Thieme Verlag.

Shiffrar, M., & Freyd, J. J. (1990). Apparent motion of the human body. *Psychological Science, 1*, 257–264.

Silbersweig, D. A., Stern, E., Frith, C., Cahill, C., Holmes, A., Grootoonk, S., Seaward, J., McKenna, P., Chua, S. E., Schnorr, L., Jones, T., & Frackowiak, R. S. J. (1995). A functional neuroanatomy of hallucinations in schizophrenia. *Nature, 378*, 176–179.

Sirigu, A., Duhamel, J.-R., Cohen, L., Pillon, B., Dubois, B., & Agid, Y. (1996). The mental representation of hand movements after parietal cortex damage. *Science, 273*, 1564–1568.

Stone, V. E., Baron-Cohen, S., Calder, A., Keane, J., & Young, A. (2003). Acquired theory of mind impairments in individuals with bilateral amygdala lesions. *Neuropsychologia, 41*, 209–220.

Titchener, E. B. (1908). *Lectures on the elementary psychology of feeling and attention*. New York: MacMillan.

Viviani, P. (1990). Common factors in the control of free and constrained movements. In M. Jeannerod (Ed.), *Motor representation and control, Attention and Performance* (Vol. XIII, pp. 345–373). Hillsdale, NJ: Erlbaum.

Viviani, P., & Stucchi, N. (1992). Biological movements look uniform. Evidence of motor–perceptual interactions. *Journal of Experimental Psychology, Human Perception and Performance*, 18, 603–623.

Weinberger, D. R., & Berman, K. F. (1996). Prefrontal function in schizophrenia: Confounds and controversies. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*, 1495–1503.

Zajonc, R. B. (1985). Emotion and facial efference: A theory reclaimed. *Science*, *228*, 15–21.

This page intentionally left blank

PART III

ROBOTS

This page intentionally left blank

7 Affect and Proto-Affect in Effective Functioning

ANDREW ORTONY, DONALD A. NORMAN, AND WILLIAM REVELLE

We propose a functional model of effective functioning that depends on the interplay of four relatively independent domains, namely, affect (value), motivation (action tendencies), cognition (meaning), and behavior (the organism's actions). These domains of functioning all need to be considered at each of three levels of information processing: the reactive, the routine, and the reflective levels. The reactive level is primarily a hard-wired releaser of fixed action patterns and an interrupt generator, limited to such things as processing simple stimuli and initiating approach and avoidance behaviors. This level has only proto-affect. The routine level is the locus of unconscious, uninterpreted expectations and well-learned automatized activity, and is characterized by awareness, but not self-awareness. This level is the locus of primitive and unconscious emotions. The reflective level is the home of higher-order cognitive functions, including metacognition, consciousness, and self-reflection, and features full-fledged emotions. In this framework, we characterize personality as a self-tunable system comprised of the temporal patterning of affect, motivation, cognition, and behavior. Personality traits are a reflection of the various parameter settings that govern the functioning of these different domains at all three processing levels. Our model constitutes a good way of thinking about the design of emotions in computational artifacts of arbitrary complexity that must

perform unanticipated tasks in unpredictable environments. It stresses the need for these artifacts, if they are to function effectively, to be endowed with curiosity and expectations and to have the ability to reflect on their own actions.

What does it take for an organism to function effectively in the world? What would a comprehensive model of the fit between an organism's functioning and the environmental conditions in which the organism finds itself look like? What role does affect play in effective functioning? Our general answer to these questions is that, for organisms of any complexity, effective functioning depends on the interplay of four domains: *affect*, what the organism feels; *motivation*, what the organism needs and wants; *cognition*, what it knows, thinks, and believes; and *behavior*, what it does.

For us, *behavior* refers only to physical action,[1] both externally observable (e.g., movements of the limbs or facial muscles) and internal (e.g., contractions of the gut or changes in heart rate). Just as the cognitive areas of the cortex are largely separable from the motor areas, we believe that, from a functional perspective, cognitive activity such as thinking and problem solving needs to be treated separately from motor activity. Cognitive activity, or *cognition*, is essentially concerned with meaning. This is in contrast to affect, which has to do with value (positive or negative). We use the term *affect* as a superordinate concept that subsumes particular valenced conditions such as emotions, moods, feelings, and preferences. Emotions are that subset of affective conditions that are *about* something, rather than being vague and amorphous, as are, for example, moods (Clore & Ortony, 2000). We also distinguish emotions from feelings. We take feelings to be readouts of the brain's registration of bodily conditions and changes—muscle tension, autonomic system activity, internal musculature (e.g., the gut), as well as altered states of awareness and attentiveness. Emotions are *interpreted* feelings, which means that feelings are necessary but not sufficient for emotions. This definition is different from that of Damasio (2000), who views emotion itself as the registration of the bodily changes and the feeling (of an emotion) as a mental image of those changes. We prefer our view to Damasio's because we think that emotions proper have cognitive content, whereas feelings themselves do not; thus, we view feelings as components of emotions rather than the other way around. The last domain of functioning, *motivation*, concerns tendencies to behave in certain ways—in particular, to attain or avoid certain kinds of state, such as satiation, danger, or becoming successful.

A central organizing theme of our discussion is that affect and the other domains of functioning need to be considered at each of three levels of

information processing: the *reactive*, the *routine*, and the *reflective* (Fig. 7.1). One of our main claims is that affect manifests itself in different ways at the different levels of processing. We believe that viewing affect and its relation to information processing in this way helps to resolve some of the debates about affect and emotion.

Some of the more important differences between the three levels are presented in Table 7.1. The main function of the most elementary level, the reactive level, is to control the organism's approach and avoidance behavior and, as described by Sloman, Chrisley, and Scheutz (see Chapter 8), to interrupt and signal higher levels. At this level, there is only simple, unelaborated affect, which we refer to as "proto-affect." The realm of proto-affect is restricted to the here and now, as opposed to the future or the past.

The second, routine, level is primarily concerned with the execution of well-learned behaviors. At this level, affect begins to show some of the features of what we would ordinarily call emotions but in a rather limited and primitive manner. These "primitive emotions" can involve information relating to the future as well as to the present. For example, simple forms of hope and fear necessitate some minimal form of expectation. We consider

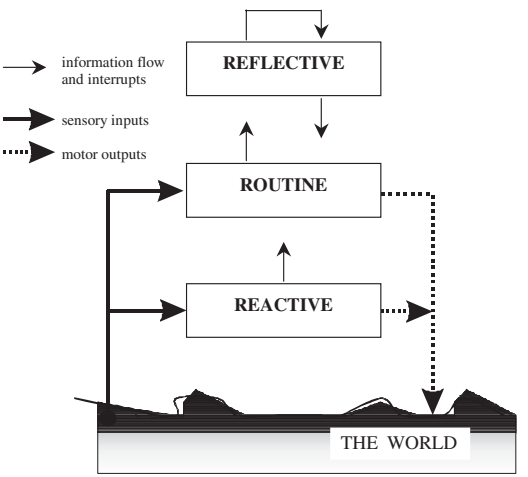


Figure 7.1. Simplified schematic of the three processing levels, *reactive*, *routine*, and *reflective*, showing their principal interconnections and relationships to one another and to the state of the world. Small solid lines represent information flow and interrupt signals that serve to indicate activity; broken lines indicate response initiation; thick lines indicate sensory signals.

Table 7.1. Principal Organism Functions at Three Levels of Information Processing

	Processing Level		
	Reactive	Routine	Reflective
Perceptual input	Yes	Yes	No
Motor system output	Yes	Yes	No
Learning	Habituation, some classical conditioning	Operant and some classical conditioning, case-based reasoning	Conceptualization, analogical, metaphorical, and counterfactual reasoning
Temporal representation	The present and primitive representation of the past	The past, present, and primitive representation of the future	The past, present, future, and hypothetical situations

the states discussed by Fellous and LeDoux (see Chapter 4) or the states related to reinforcement discussed by Rolls (see Chapter 5) to be routine-level, "primitive" emotions.

Finally, the third and most sophisticated level, reflective, is the locus of higher-level cognitive processes and consciousness. At this level, we get full-fledged emotions that are cognitively elaborated; that can implicate representations of the present, the future, or the past; and that can be named. These are the kinds of emotional state that are the focus of appraisal theories (e.g., Arnold, 1960; Lazarus, 1966; Mandler, 1984; Ortony, Clore, & Collins, 1988; Roseman, 1984; Scherer, 1984).

We are certainly not the first to propose a multilevel analysis of information processing. Many others have proposed such accounts, although often starting from quite different places. For example, Broadbent (1971), considering evidence of similarities and differences in the effects of various stress manipulations, argued for at least two levels of cognitive control; and Sanders (1986) discussed how multiple levels of energetic and cognitive control (including arousal, activation, and effort) are utilized as a function of task demands (see also Revelle, 1993). Advocates of computational approaches have also proposed models involving several levels of information processing (e.g., Sloman & Logan, 2000; also Sloman et al., Chapter 8, and Minsky, in preparation), representing impressive and highly elaborated examples. Our approach is in the same spirit as these, focusing as it does on the implications of the information-processing levels we identify for a model of affect and effective functioning.

The model that we propose is a functional one. However, we believe that many aspects of it are consistent with neuroanatomical accounts, with the three levels—the reactive, the routine, and the reflective—corresponding roughly to the assumed functions of the spine/midbrain basal ganglia, cortex, cerebellum, and prefrontal cortex. It thus bears some similarity to MacLean's (1990) early proposal of the triune brain, which although in many ways problematic is still a useful framework (see Fellous and LeDoux, Chapter 4). Proposals such as those of Fellous (1999) and of Lane (2000) are also compatible with our general approach.

As already indicated, we consider affect to be a general construct that encompasses a wide range of psychological conditions relating to value. However, even though emotions are more highly specified than other affective states, they do not comprise a discrete category with easily identifiable boundaries. Rather, they vary in their typicality, with some cases being better examples than others. Thus, we propose that the best examples of emotions, which we often refer to as "full-fledged emotions," are interpretations of lower-level feelings and occur only at the reflective level, influenced by a combination of contributions from the behavioral, motivational,

and cognitive domains. At the middle, routine, level, we propose basic feelings, "primitive emotions," which have minimal cognitive content, while at the most elementary, reactive, level of processing, we argue that there are no emotions at all. All that is possible at the reactive level is an assignment of value to stimuli, which we call "proto-affect." This in turn can be interpreted in a wide range of ways at higher levels from a vague feeling that something is right or wrong (routine level) to a specific, cognitively elaborated, full-fledged emotion (reflective level). Although these different kinds of affective state vary in the degree to which they involve components of prototypical emotions, many are still emotion-like, albeit not very good examples (Ortony, Clore, & Foss, 1987). The graded nature of the concept of emotion is readily accommodated by our account partly because the specificity of an affective state is held to depend on the information-processing level at which it appears.

Although our focus in this chapter is on affect and emotion, we generally consider the affective aspects of an organism's functioning in light of their interactions with the other domains of functioning (behavior, motivation, and cognition) because we believe that this sort of integration is important. We start with a discussion of these issues in the context of biological beings and then consider how some of these ideas might apply to the design of robots and autonomous agents. For convenience, we sometimes discuss the levels within an organism (e.g., a human or a robot) and sometimes in terms of different organisms (e.g., organisms that might be restricted in the number of levels available to them). We discuss our analysis as it relates to personality and individual differences. Since we take affect, motivation, and cognition to be the internal control mechanisms of behavior, we view differences in their steady-state parameters as comprising the strikingly consistent and strikingly different organizations that are known as personality. So, for example, an individual who exhibits strong positive affect and strong approach behaviors might be classified by a personality theorist as an *extravert*. Such a person responds very differently to environmental inputs than one—an *introvert*—who has a strong negative affect system and strong inhibitory behaviors.

Our goal here is to lay out a general framework that might help us in thinking about a number of issues relating to affect and emotion. Thus, when we discuss some particular function or behavior in the context of one or other of the levels of processing, we are not necessarily rigidly committed to the idea that the function is performed at, or exclusively at, that level. In a number of instances, our assignments of functions to levels are speculative and provisional. For example, classical conditioning encompasses a wide range of behaviors and learning which may not all involve the same brain structures (e.g., some involve the hippocampus and some do not). Accordingly, some of the phenomena of classical conditioning should probably be thought

of as originating from only the reactive level and some from the routine level, but we are undecided about exactly how to conceptualize the distribution of these phenomena across levels.[2] Readers should understand, however, that we are more concerned with articulating a way of thinking about how affect, motivation, cognition, and behavior interact to give rise to effective functioning than we are with particular details. It is our hope that future research will enable some of our proposals to be tested empirically. In the meantime, we are finding them to be a fruitful way of reconceptualizing some issues related to effective functioning in general, as well as issues in more specific domains such as personality theory, people's responses to products, and the design of autonomous, intelligent systems.

AFFECT AT THREE LEVELS OF PROCESSING

In this section, we discuss the way in which affect is manifested at each level of processing. The reactive level is primarily a releaser of fixed action patterns and an interrupt generator (Fig. 7.1). These interrupts are generally registered at the next level up, the routine level, which is the locus of well-learned automatized activity, characterized by awareness but not self-awareness. Self-awareness arises only at the highest, reflective, level, which is the home of higher-order cognitive functions, including metacognition, consciousness, and self-reflection. We interpret the existing psychological and neurological evidence to indicate that the reflective level processes are prefrontal, which means that, unlike the reactive and routine levels, the reflective level neither receives direct sensory inputs nor directly controls motor output. Reflection is limited to analyzing internal operations and to biasing and otherwise controlling routine-level activity. In fact, both reactive- and routine-level processing can modulate the operating characteristics of the reflective level, for example, by changing attentional focus, both by patterns of neural firing and chemically, through neurotransmitter changes.

Affect at the Reactive Level: Proto-Affect

Reactive-level processing comprises biologically determined responses to survival-relevant stimuli and is thus rapid and relatively unsophisticated with respect to both its detection mechanisms and its behavioral repertoire. New activity at the reactive level generally results in modification of output, but none of the activity is cognitive in nature—there is no cognition at the reactive level. Furthermore, at the reactive level, the other three domains of functioning—affect, behavior, and motivation—are so closely intertwined

that they are better thought of as different perspectives on the same phenomenon rather than different phenomena. Reactive-level behavioral responses are of two broad classes—approach and avoidance—each governed by mechanisms of activation and inhibition. These responses also serve as alerting mechanisms, interrupting and causing higher levels of processing (in organisms that have them) to attend to the interrupting event and thus sometimes permitting a better course of action than would otherwise have been possible. The sophistication of the reactive level varies with the sophistication of the organism so that amoebas, newts, dogs, and humans vary considerably in the range of stimuli to which their reactive levels are responsive as well as in the types of behavior that reactive-level processes can initiate. Most reactive-level processing is accomplished through pattern recognition, a mechanism which is fast but simple and thus limited in scope. This means that it has a high potential for error, in both false diagnoses (false alarms) and missed ones (misses). The associated behaviors—motor responses—are either very simple, such as reflexes or simple fleeing or freezing behaviors, or preparatory to more complex behaviors governed by higher information-processing levels. In rare cases, reactive-level behaviors can involve more coordinated responses such as those necessary for maintaining balance.

The reactive level is highly constrained, registering environmental conditions only in terms of immediately perceptible components. Consequently, it needs and has a very restricted, simple, representational system; in particular, its crude and limited representations of the past are restricted to those that are necessary for habituation and simple forms of classical conditioning. In particular, registration of anomalous events is highly restricted, limited to such things as local violations of temporal sequencing. Nevertheless, although processing that necessitates comparing a current event with past events (e.g., case-based reasoning) is unavailable to reactive-level processes, in complex organisms such as humans, a great deal can still be accomplished through the hard-wired, reactive-level mechanisms.

So far, our discussion of reactive-level processes has focused on behavioral responses. This is because there is much less to say about reactive-level motivation and affect. The only forms of motivation that are operative at the reactive level are simple drives (e.g., appetitive and survival drives). Given a modicum of evolutionary complexity, organisms can have multiple drives that are sometimes incompatible. For example, the newt, motivated to copulate (below the water), is also motivated to breathe (above the water); sometimes the behavior can be modified to accommodate both drives, but otherwise, the more critical will dominate and temporarily inhibit the other (Halliday, 1980).

As for affect, our proposal is that at the reactive level there is only the simplest form of affect imaginable, what we call "proto-affect." For all of the stimuli that the organism encounters, the reactive level assigns values

along two output dimensions, one of which we call "positive" and the other "negative."[3] These signals, which are the fundamental bases of affect and emotion, are interpreted by and interrupt activity at higher levels of processing. Thus, proto-affect represents nothing more than the assignment of valence to stimuli. At the same time, these reactive-level affective signals are so intimately related to behavioral (especially motor) responses and to the motivation to approach or avoid a stimulus that it makes little sense to try to distinguish them from one another. Throughout evolutionary history, the specificity of the automatic response systems has grown so much that, in the human, there are specific, prepared responses to a wide range of stimulus classes. For example, smiling faces, warm environments, rhythmic beats, and sweet tastes automatically give rise to predominantly positive valence, while frowning faces, extreme heat or cold, loud or dissonant sounds, bitter tastes, heights, and looming objects immediately induce negative valence.

Reactive-level processes enervate the motor system in preparation for one of a limited set of fixed action pattern responses. Consider, for example, how a human responds to the taste of a bitter or caustic substance. On our analysis, the reactive level assigns negative value to the substance, so the motivation is to immediately reject it. The body withdraws, the mouth puckers, the diaphragm forces air through the mouth, ejecting the food. At the same time, human observers typically attribute specific affective states based on their observations of such behavior. For example, if, as observers, we see a baby grimace, move its head away, and spit out a substance, we say that the baby "dislikes" the substance. However, in our analysis, at the reactive level, all that exists is proto-affect and the tightly coupled motivation to expel the substance and the associated behavior of spitting it out. Because as observers we see this nexus of motivation and behavior, we attribute an emotional state to the baby—we talk about the baby disliking the substance. However, our view is that an emotion of dislike or disgust involves an interpretation that can take place only at higher levels of processing.[4]

Of course, reactive-level responses to any given stimulus are not identical across different people (or even in the same person on all occasions). In other words, the parameters governing the operating characteristics of reactive-level functioning can vary across individuals and, although generally to a lesser degree, within individuals from one time to another. The kinds of parameter that we have in mind here include the strength, speed, accuracy, and sensitivity of a variety of basic functions carried out at the reactive level. For example, it is likely that the strengths of approach and avoidance behaviors vary and interact with activation and inhibition, which also vary in strength. Variations in sensitivities might be expected of, for example, perceptual acuity and anomaly detection (as in the detection of temporal

irregularities or discontinuities). In addition, we assume that the latency and intensity of signals sent to higher levels of processing vary.

Affect at the Routine Level: Primitive Emotions

The core of the routine level is the execution of well-learned routines—"automatic" as opposed to "controlled" processes (e.g., Schneider & Shiffrin, 1977). In contrast to the reactive level, the routine level is capable of a wide range of processes, from conditioning involving expectancies to quite sophisticated symbolic processing. This is the level at which much human behavior and cognition is initiated and controlled. Here, elementary units are organized into the more complex patterns that we call "skills." Behavior at the routine level can be initiated in various ways, including activity at the reflective level (e.g., deciding to do or not to do something) and, as just discussed, activity at the reactive level (Fig. 7.1). Some routine-level processes are triggered by other routine-level activity. Finally, some routines can be triggered by the output of sensory systems that monitor both internal and external signals.

As well as routinized behavior, the routine level is the home of well-learned, automatic cognitive processes, such as the cognitive aspects of perception and categorization, basic processes of language comprehension and production, and so on. Indeed, we call this level "routine" because it encompasses all non-reactive-level processing that is executed automatically, without conscious control (which is the purview of the reflective level). However, although there is no consciousness at the routine level, awareness is an important cognitive aspect of it. Earlier, we defined cognition as the domain associated with meaning and affect as the domain associated with value. One of the things of which we can be (cognitively) aware is (affective) feeling; but although there is awareness at the routine level, there is no self-awareness. This is because self-awareness is a reflexive function. Since routine-level processes cannot examine their own operations, self-awareness is possible only at the reflective level.

Expectations play an important role at the routine level. Routine-level processes are able to correct for simple deviations from expectations, although when the discrepancy becomes too large, reflective-level control is required. Consider the case of driving an automobile. If the routine level registers a discrepancy between the implicit expectations and what actually happens and if the driver has sufficient expertise, the routine level can quickly launch potential repair procedures, even though such procedures might sometimes be suboptimal. On the other hand, an inexpert driver may have no routine procedures at all to engage under such conditions, in which case slower

reflective-level processes take control and generate a conscious decision as to what to do, which might well be too late. In other words, there is a speed/accuracy tradeoff at work with respect to the two levels.

It is important to emphasize that routine-level expectations are implicit rather than explicit. They are the automatic result of the accumulation of experiences that forms a general model of likely, or "normal," outcomes or events—stored norms which are automatically recruited when anomalies occur (Kahneman & Miller, 1986). The strength of these expectations together with the intensity of valence associated with the current and expected states influence the strength of the ensuing feelings. Expectations might also arise from some kind of continuity-of-experience mechanism—an implicit belief that the future is not apt to deviate much from the recent past. However, whether the expectations are learned or rooted in expectations of experiential continuity, the key point is that the routine level can only detect expectation violations: it cannot interpret them. Only the reflective level can interpret and understand discrepancies and their consequences and then provide active, conscious decisions as to what to do about them. When some discontinuity, potential problem, or disruption of a normal routine is encountered, an interrupt is generated that, in its turn, generally launches other processes and affective reactions. This interrupt might be thought of as a primitive form of surprise. However, in the model, this interrupt is not valenced; therefore, it is categorically not an affective signal or emotion of any kind (Ortony & Turner, 1990). The system needs to do more work before value can be assigned. Thus, we view surprise as the precursor to emotion (Mandler, 1984). This is consistent with the neuroanatomical finding (Kim et al., 2004) that while one region of the amygdaloid complex responds similarly to fear and surprise (suggesting that valence has not yet been assigned), a separate region is responsive to fear but not surprise (suggesting that it is responding to valence per se).

Whereas the reactive level can have only unelaborated positive and negative affect, some minimal elaboration does occur at the routine level. Given our view that the routine level allows for some representation of the future as well as the present, four elementary cognitive categories emerge as a result of crossing these two levels of time with the two levels of valence (positive and negative). These four categories lie at the heart of the rudimentary, primitive emotions that arise at the routine level. In terms of the kinds of emotion specification described by Ortony, Clore, and Collins (1988), these four experientially discriminable primitive emotional states can be characterized as follows:

1. A (positive) feeling about a **good thing** (present)
2. A (negative) feeling about a **bad thing** (present)

3. A (positive) feeling about a **potential good thing** (possible future)
4. A (negative) feeling about a **potential bad thing** (possible future)

If we were to try to assign conventional emotion names to these states (which we think is inadvisable), the first two could be said to correspond roughly to something like "happiness" and "distress" and the second two to primitive forms of "excitement" and "fear," respectively.[5] We call these "primitive emotions," to convey the idea that they are routine-level feelings—affective states which have not yet been interpreted and cognitively elaborated. We think that animal studies of the kind reported by LeDoux (1996) and studies with humans involving unconscious processing of fear-relevant stimuli (e.g., Öhman, Flykt, & Lundqvist, 2000) are studies of routine-level, primitive emotions. As we discuss in the next section, there is an important difference between the "primitive" fear of the routine level and fully elaborated fear, which occurs only at the reflective level. Our analysis, in which four of the primitive emotions result from the product of two levels of valence (positive and negative) and two levels of time (present and future), is also consistent with the proposals of researchers such as Gray (1990), and Rolls (1999; see Chapter 5).

We propose that affective states at the routine level have some, but not all, of the features of a full-fledged emotion and that, at this level, affective states are related to but separable from cognition and motivation. The routine level lacks the cognitive resources necessary to interpret feelings as emotions by making the kind of rich, conscious elaborations of situations (e.g., reasoned, causal attributions) that characterize full-fledged emotions. Sophisticated processes such as these are available only at the reflective level.

We now need to consider the nature of motivation at the routine level. Whereas at the reactive level we had only simple motivations such as drives and approach-and-avoidance tendencies, much richer motivational structures, such as inclinations, urges, restraints, and other, more complex action tendencies, guide behavior at the routine level. These motivations to engage in or inhibit action are now clearly distinct from the actions themselves and related to, but again clearly distinct from, primitive emotions. At the reactive level, motives are entirely driven by cues, whether internal or external, but the motivation disappears when the cue goes away.[6] In contrast, at the routine level, motivations persist in the absence of the associated cue, dissipating only when satisfied. A good historical example of this is the Zeigarnik effect (Zeigarnik, 1927/1967), wherein activities that are interrupted are remembered better than those that are not.

There are, of course, numerous individual differences in the basic parameters of the neuroanatomy at the routine level which translate into differences in the construction and use of routines. Any of the routine-level

subsystems—perception, motor control, learning, memory—will vary in their sensitivity, and capacity for and speed of processing. These, in turn, translate into differences in the rate at which individuals can integrate information, learn skills, or acquire and recall information. Important differences for personality theorists include the sensitivity of the routine level to interruption from below (i.e., reactive level) or to control from above (i.e., reflective level; see Fig. 7.1). There might also be differences in sensitivity to sensory cues and in the tendency to do broad, global processing rather than more narrowly focused processing.

In addition, whereas reactive-level processes are essentially fixed by biology, much of the content at the routine level is learned. Because complex skills are heavily dependent on the substrate of prior learned material, individual differences in experiences and learning accumulated throughout life make for eventual large differences in abilities. Thus, both biological (genetic) and environmental (learned) differences emerge at the routine level.

Affect at the Reflective Level: Cognitively Elaborated Emotions

Reflection is a special characteristic of higher animals, most marked in primates and especially humans. Humans can construct and use mental models of the people, animals, and artifacts with which they do or could interact, as well as models of those interactions. Rich representational structures of this kind enable complex understanding, active predictions, and assessments of causal relations. Humans also have a notion of self; we have self-awareness, consciousness, and importantly, representations of the minds of others. This leads to the possibility of elaborate systems of competition and to the ability to lie and deceive, but it also leads to more sophisticated social cooperation and to a propensity for humor, art, and the like. Monkeys and apes may share some of these cognitive abilities (e.g., deWaal & Berger, 2000), but these abilities remain preeminently human.

The kinds of capability that comprise the enhanced processing of the reflective level depend on the ability of the reflective level to perceive, analyze, and in some cases, alter its own functioning as well as that of the routine and reactive levels. Humans (at least) can examine their own behaviors and mental operations, reflect upon them, and thereby enhance learning, form generalizations, predict future events, plan, problem-solve, and make decisions about what to do. In general, the reflective level comprises consciousness together with all of the advanced cognitive and metacognitive skills that have enabled humans to increase their knowledge cumulatively over the millennia.

We consider the well-established finding that prefrontal regions of the brain subserve the programming, regulation, and verification of activity (e.g., Damasio, 1994; Goldberg, 2001) as support for the separability of the kind of conscious control functions of the reflective level from other, more automatic behaviors. The fact that prefrontal damage does not affect routine behavior or the performance of well-learned skills is also consistent this view. Note that in our model—and in any model that identifies the prefrontal lobes as the locus of such activities—the reflective level neither receives direct perceptual information as input nor directly controls motor output. This means that the reflective level can only bias the levels beneath it. Norman and Shallice (1986) viewed this bias signal as "will." In their model, will is a control signal such that if some activity at a lower level is desired, the control level can add activation signals to it, thereby increasing the likelihood that it will get performed.

It is the power of the reflective level that makes possible the rich emotional experience that we assume is unique to humans. At the reflective level, not only are emotions and their associated behaviors sometimes actually initiated, as when reminiscing about prior experiences can lead to changes in moods and emotions, but less well-defined affective states become elaborated, interpreted, and transformed into full-fledged emotions. Thus, whereas at the reactive level there is only unelaborated proto-affect and at the routine level only feelings and primitive emotions, the reflective level has the capacity to interpret unelaborated proto-affect from the reactive level and primitive emotions and feelings from the routine level so as to generate discrete emotions that can be labeled. This cognitive elaboration comes about by relating higher-level cognitive representations and processes to the kind of internal and external events that induce affect in the first place.

Because the reflective level is the locus of all high-level cognitive processing, it has a rich repertoire of representational and processing resources. In addition to goals, standards, and tastes, the three classes of emotion-relevant representations identified by Ortony, Clore, and Collins (1988), these resources include such things as conscious expectations; plans; mental models and simulations; deductive, inductive, and counterfactual reasoning; and so on. At this level, it is possible to take feelings as objects of thought: we can (sometimes) label them, we try to make sense of them, and we can plan actions around them.

To illustrate this, consider the consequences of reflecting upon realized or unrealized potentials (e.g., fulfilled vs. violated expectations). The two future-oriented emotions, 3 and 4 discussed in the preceding section, have associated with them a further pair of states—one corresponding to the potential being realized (e.g., a confirmed expectation) and the other corresponding to the potential not being realized (e.g., a disconfirmed expecta-

tion). The emotions that derive from 3 (a [positive] feeling about a potential good thing) are:

3.1. A (positive) feeling about a **potential good thing, realized**
3.2. A (negative) feeling about a **potential good thing, not realized**

The emotions that derive from 4 (a negative feeling about a potential bad thing), are:

4.1. A (positive) feeling about a **potential bad thing, not realized**
4.2. A (negative) feeling about a **potential bad thing, realized**

These are four full-fledged emotional states deriving from primitive emotions or emotional feelings originally experienced at the routine level. They are affective because they involve the evaluation of something as good or bad, helpful or harmful, beneficial or dangerous, and so on; they are feelings because they inherit feeling qualities from their lower origins, albeit now changed and augmented by cognition; and they are emotions because they are about something (Clore & Ortony, 2000) and have consciously accessible content.

Of course, as anyone who has ever acted in the heat of the moment knows, strong emotions and their routine-level behavioral concomitants often overwhelm cool reason and its more planful reflective-level responses; but this very fact presupposes, rather than vitiates, the routine–reflective distinction. In fact, there are several reasons why careful, logical planning activities at the reflective level might be thwarted. One such reason is that routine-level responses might become initiated before the reflective level has completed its analysis. Another is that inhibitory signals initiated at the reflective level are too weak to overcome the automatic procedures initiated at the routine level. Finally, the emotional state might cause hormonal states that bias the reflective processes to do more shallow processing, presumably in an effort to quicken their responses, thus generating responses that are logical at the surface but that have severe negative results that would have been predicted had the reflective processes been allowed to continue. Emotional responses are often first-order responses to situations, with poor long-term impact.

It may be informative to consider an example that illustrates the rapid, automatic action at the routine level, preceding both thoughtful planning at the reflective level as well as the delayed interpretation of the resulting affective state. Many years ago, one of the authors spent a year living in a coastal town in tropical Africa. One day, on his way to the beach, he was driving slowly and with considerable difficulty across a shallow, rough, dried-up riverbed with his car windows open. Suddenly, and quite unexpectedly, he saw a huge crocodile that had been lying still on the riverbed, now

disturbed by the approaching car. Panicked, he put his foot on the brake pedal to stop the car, leaned across the unoccupied passenger seat, and frantically rolled up the window on the side where the crocodile was. Having done this, he rolled up the window on his (driver's) side and, shaking and heart pounding, drove, still slowly and with difficulty, out of the riverbed, to what he took to be safety. Then, and only then, did he become aware of how terrified he was.

In this example, a potential threat was perceived and a rapid protective-behavior routine initiated. There was too little time to optimize the selected routine. The system was satisficing rather than optimizing. Realistically, it might have made more sense to just keep going—the crocodile was not likely to climb into a moving car through the passenger door window and devour the driver. Presumably, the driver stopped the car to facilitate the closing of the window, but this was not thought through or planned—it was just done—a sequence of the "car-stopping" routine followed by the "window-closing" routine. Furthermore, the behavior is not well described by saying that it was done in response to, or even as part of, fear. As described, the emotion of fear came only after the driver had engaged in the protective behavior and extricated himself from the situation—only then, on reviewing his racing heart, his panicky and imperfect behavioral reactions, and the situation he had just been in, did he realize how frightened he was. In other words, the emotion was identified (labeled) as fear only after the behavior and concomitant feelings (of bodily changes) had been interpreted and augmented by cognition at the reflective level. The situation is best described by saying that first came the feeling of primitive fear (which includes an awareness of the bodily changes) and then, upon interpretation and additional cognitions, came the full-fledged emotion of fear.

This example not only bears upon several aspects of our three-level model but also speaks to the James-Lange theory of emotions (James, 1884; Lange, 1895/1912), especially with respect to the temporal relationship between emotions and behavior. In our example, the rapid behavior occurred before the emotion was identified, exactly as William James described it with respect to his imaginary bear in the woods:

> the bodily changes follow directly the perception of the exciting fact, and [that] our feeling of the same changes as they occur is the emotion. Common sense says, we lose our fortune, are sorry and weep; we meet a bear, are frightened and run; we are insulted by a rival, are angry and strike. The hypothesis here to be defended says that this order of sequence is incorrect . . . and that the more rational statement is that we feel sorry because we cry, angry because we strike, afraid because we tremble . . . Without the bodily states fol-

> lowing on the perception, the latter would be purely cognitive in form, pale, colorless, destitute of emotional warmth. We might then see the bear, and judge it best to run, receive the insult and deem it right to strike, but we should not actually feel afraid or angry.

Now consider James' example of the emotion that accompanies one's loss of a fortune. In this case, it would seem that the reflective-level analyses come first. The person would start thinking about possible causes of the loss, perhaps reviewing past actions by (formerly) trusted associates and then assessing blame. Such cognitions would be likely to invoke evaluation as a result, for example, of running through various "what-if" scenarios and imagining the responses of family, friends, and colleagues. This kind of cognitively induced introduction of sources of value would be the wellspring of bodily changes, the awareness of which would constitute the underlying emotional feeling. However, if all of this were to lead to anger, the anger would have followed the cognitions. Similarly, James' emotion of "shame" results from self-blame, and this means that it is cognition, not behavior, that is the trigger. All this suggests to us that the question is not whether the James-Lange theory is right or wrong but, assuming that it is at least in part right, under what conditions it is right and under what conditions it is wrong. So, if one asks the question "Which comes first, cognition or behavior?" the answer has to be that it depends. When reactions are triggered from the reactive or routine level, behavior precedes; but when the triggering comes from the reflective level, cognition precedes.

Much as with the routine level, there are many sources of individual differences in the operating parameters of the reflective level. These are likely to include such things as sensitivity, capacity, and processing speed plus the ability of the reflective level to influence lower levels through its control signals of activation and inhibition. We would also expect to find differences in conscious working memory and attentional focus, especially with respect to sensitivity to interruptions and other events. Finally, there will be substantial individual differences in the content at both the behavioral and reflective levels, and inasmuch as the reflective level is the locus of one's self image and much cultural knowledge and self-examination, these differences can be expected to have a significant effect on the way a person interacts with the environment and with others.

IMPLICATIONS FOR PERSONALITY

We have already suggested a number of parameters for which we might expect inter- and intra-individual differences at the different levels of

processing. We view parameters of this kind as the foundations of personality. Inevitable variations in parameter values lead to individuals differing in the ways in which, and the effectiveness with which, they function in the world. However, personality research lacks a consensual account of what personality is (especially with respect to its causal status), so we start our discussion by situating our account in relation to the principal current approaches to personality theory.

Most current research in personality focuses on individual differences in affect and interpersonal behavior while adopting one of two different and largely incompatible perspectives. One of these seeks to identify the primary dimensions in terms of which descriptions of systematic regularities and differences across different times and different places can be parsimoniously but informatively cast. The other perspective views personality as a causal factor in the functioning of individuals and thus seeks to identify deeper explanations of such similarities and differences. We believe that our approach can resolve some of the conflict between these two perspectives and that it moves beyond both by extending the purview of personality theory from affect and interpersonal behavior to include behavior more generally as well as motivation and cognition. For us, personality is a self-tunable system comprised of the temporal patterning of affect, motivation, cognition, and behavior. Personality states and traits (e.g., for anxiety) are a reflection of the various parameter settings that govern the functioning of the different domains at the different levels.

One of the most paradoxical yet profound characterizations of personality is the idea that all people are the same, some people are the same, and no people are the same (Kluckholm & Murray, 1953). In our terms, all people are the same in that everyone is describable in terms of the four domains of functioning (affect, motivation, cognition, and behavior) at the three levels of processing (reactive, routine, and reflective); some people are the same in that they are similar in the way that they function in some or all of the domains; and finally, no one is the same in the unique details of the way in which the four domains interact with each other and at the three processing levels.

With respect to our levels of processing, it is clear that individual differences occur at all three levels. We have already suggested possible dimensions of variability at the different levels. For example, at the reactive level one might expect differences in sensitivity to environmental stimuli, aspects of response strength, and ability to sustain responses. Such differences would manifest themselves as variations in the likelihood of approach and avoidance and in proto-affective responses (Schneirla, 1959). As outside observers, we might characterize some of these as variations in a behavioral trait. For example, one might map observed differences in probabilities of approach

and avoidance onto a boldness–shyness dimension, as do Coleman and Wilson (1998) in their description of pumpkinseed sunfish.[7] More generally, individual differences at this level were discussed long ago by Pavlov and later by others in terms of strength and lability of the nervous system (Pavlov, 1930; Nebylitsyn & Gray, 1972; Robinson, 1996, 2001; Strelau, 1985).

At the routine level, individual differences become more nuanced. Consider an individual who, relative to others, has a high level of positive affect and a high likelihood of approach behaviors, both emanating from the joint effects of reactive- and routine-level processing.[8] This combination of operating parameters is typical of the trait "extraversion." In other words, the descriptive label "extravert" is applied to someone who is high on both the affective and behavioral dimensions. This additive structure will, of course, result in correlations of extraversion with positive affect and with approach behavior but not necessarily to high correlations between responses across the different domains (i.e., of positive affect with approach behaviors). Our view is that the reason that we call someone an extravert is that they tend to do things such as go to lively parties (behavior) and they tend to be happy (affect). Similarly, the descriptive term for an emotionally less stable individual ("neurotic") reflects a larger likelihood of negative affect as well as a higher likelihood of avoidance behaviors. Although many situations that induce negative affect also induce avoidance behaviors, and thus make individual differences in negative affect and avoidance more salient, "neuroticism" is merely the label applied to those who are particularly likely to experience high negative affect while avoiding potentially threatening situations. (A somewhat similar argument was made by Watson, 2000, who emphasized the affective nature of extraversion and neuroticism and considered the functional nature of approach and withdrawal behavior in eliciting affect.) The virtue of this account is that it explains the fact that reliably large correlations across domains of functioning are hard to find. From the point of view of the parameters that control their operation, the domains of functioning are largely independent.

Although there are exceptions, most personality inventories and rating scales are designed to get at what we consider to be routine-level activity (although they do so by soliciting reflective-level responses). Such measures often use items that tap separately the different domains. Thus, an item like "Do you feel nervous in the presence of others?" is an attempt to get at routine-level affect, the item "Do you avoid meeting new people?" addresses routine-level behavior, and the item "Does your mind often wander when taking a test?" addresses routine-level cognition. To be sure, someone who is high on all three of these items is likely to act and feel very differently from someone who is low on all three. However, because for each person the parameter settings in the different domains of functioning are probably

independent, a value on one item (domain) does not predict the value of any others.

At the reflective level, we see the complex interplay of individual differences in motivational structures (e.g., promotion and prevention focus; Higgins, 2000) with cognitive representations (e.g., attributions of success and failure; Elliot & Thrash, 2002) that lead to the complex affective and behavioral responses we think of as effective functioning. It is also at this level that people organize life stories to explain to themselves and others why they have made particular life choices (McAdams, 2001).

We suspect that most, if not all, of the five major domains of the traditional descriptive approach to personality (see John & Srivastava, 1999, for a discussion) can be accounted for by individual differences in the parameters and content of the three levels of processing and the four domains of functioning. As we have already discussed, differences at the reactive level reflect differences in sensitivities to environmental situations. The reactive level is probably also the home of phobias such as fear of heights, crowds, darkness, snakes, spiders, and so on, which might explain why these are relatively easy to acquire but very difficult to extinguish. Routine- and reflective-level differences will exist both at the biological substrate and in learned routines, behavioral strategies, and cultural norms. These will probably determine many of the "Big 5" parameters, with neuroticism and extraversion and parts of agreeableness and conscientiousness probably due to routine-level differences and openness and the more planful parts of conscientiousness due to more reflective-level concerns (see also Arkin's Chapter 9).

By conceptualizing personality in terms of levels of processing and domains of functioning, we believe that we can improve upon prior personality research that has tended to focus on functioning drawn from only one domain at a time (e.g., affect and neuroticism or approach behavior and extraversion). We also think that by applying this approach we will be able to integrate biologically and causally oriented theories with descriptive taxonomies, which, while perhaps lacking explanatory power, have nevertheless been quite useful in predicting functioning in real-life settings (e.g., job performance in the workplace; Barrick & Mount, 1991).

IMPLICATIONS FOR THE DESIGN OF AUTONOMOUS ROBOTS AND OTHER COMPLEX COMPUTATIONAL ARTIFACTS

In animals, affect, motivation, cognition, and behavior are all intertwined as part of an effective functioning system. There is no reason to believe that it

should be any different for intelligent, socialized robots and autonomous agents, physical or virtual. Just as species at different levels of evolutionary complexity differ in their affective and cognitive abilities, so too will different machines differ. A simple artifact, such as a robotic vacuum cleaner, is implemented as a purely reactive-level device. At this level, affect, motivation, and behavior cannot be separated from one another. Such a device has the analog of hard-wired drives and associated goal states. When there is conflict, it can be resolved by the kind of subsumption architecture described by Brooks, which has been implemented in a variety of simple robots (e.g., Brooks, 1986, 2002; see Chapter 10).

More complex artifacts that can perform large numbers of complex tasks under a variety of constraints require routine-level competence. Thus, SOAR, the cognitive modeling system that learns expert skills, is primarily a routine-level system (Rosenbloom, Laird, & Newell, 1993). In fact, expert systems are quintessentially routine-level systems. They are quite capable of expert performance but only within their domain of excellence. They lack higher-level monitoring of ongoing processes and extra-domain supervisory processes. Finally, when HAL, the fictional computer in the movie *2001*, says "I'm afraid, Dave," it is clearly identifiable as a reflective-level computational artifact (assuming that the statement resulted from consideration of its own state). Whether any artifact today operates at the reflective level is doubtful. To address the question of what it would take for this to happen, we now examine how the model of effective functioning that we have sketched might apply to autonomous robots and other complex computational artifacts. In doing so, we will pay special attention to the functional utility of affect for an organism, be it real or synthetic.

We believe that our model, which integrates reactive- and routine-level processing with reflective-level processing and incorporates the crucial functions played by affect, constitutes a good way of thinking about the design of computational artifacts. This is particularly so for artifacts of arbitrary complexity that must perform unanticipated tasks in unpredictable environments. When the task and environment are highly constrained and predictable, it is always appropriate and usually possible to use strong methods (Newell & Simon, 1972) and build a special-purpose device that performs efficiently and successfully, as is current practice with most of today's industrial robots. However, under less constrained tasks and environments, strong methods are inadequate unless the system is capable of producing new mechanisms for itself. A system capable of generating its own, new, special-purpose mechanisms would necessarily employ some weak methods and would probably need an architecture of similar complexity to the one we are proposing.

Implications of the Processing Levels

In the early days of artificial intelligence and cognitive psychology, considerable attention was devoted to how to best represent general and specific knowledge, plans, goals, and other cognitive constructs and how to do higher-order cognitive functioning such as language understanding, problem solving, categorization, and concept formation. To some extent, this ignored motivation, which, of course, is necessary to explain why an organism would establish a goal or develop a plan in the first place. Ironically, behaviorist psychologists—the very people against whom the cognitivists were reacting—had worried about these issues and had even proposed biologically plausible models of the causes of action initiation (e.g., the dynamics of action model; Atkinson & Birch, 1970). We think that recent revivals of this model (e.g., Revelle, 1986) can do a reasonable job of accounting for a good deal of action initiation at our reactive and routine levels.

It is easy to understand why a robot—or any organism, for that matter—acts when confronted with environmental conditions (or internal drives) that demand some kind of response; but what happens when they are not imposing any demands on the organism and it is at or close to homeostasis? Does it then just remain idle until some new action-demanding condition arises that causes it to behave? Animals' motivation systems handle this by letting the resting point of affect be slightly positive so that when there is nothing that needs to be done, the animal is led to explore the environment (see Cacioppo, Gardner, & Berntson, 1997, on positivity offset). This is the affective basis of *curiosity* (an innate motivational force that leads organisms to explore the environment and to try new things). Certainly in humans, curiosity (openness to experience) is a powerful learning aid. So should it be for a robot. Clearly, an autonomous robot is going to need expectations. Perceiving and acting in the world while indifferent to outcomes would hardly be conducive to effective functioning. At the routine level, our model provides implicit expectations (in contrast to the conscious expectations and predictions of the reflective level). Expectations are important not only because their confirmations and disconfirmations are crucial for learning but also because the resulting affect changes the operating characteristics of the other three domains. At the routine level, implicit expectations are tightly bound to their associated routines. They come into play much less often when routines run off successfully than when they fail or are interrupted. Recall that at this level proto-affect from the reactive level becomes partially elaborated as primitive emotions (feeling good or bad about the present or potential future). In designing an autonomous robot, we would need to consider the motivational, cognitive, and behavioral consequences of these primi-

tive emotions. Consider the simplest case, that of feeling good or bad about the present. Part of the power of affective states in general derives from the fact that they are the result of mapping many inputs onto a few or, in the limiting case, two (positive and negative) kinds of internal state. For example, any of a multitude of disconfirmed positive expectations or confirmed negative ones can reduce one to the primitive emotional state that we might call displeasure or distress. This affective state in turn functions as a simple modulator of processing parameters in the other three domains of functioning. Thus, the power of affect, and hence its value for robot design, is its data-reduction capacity and consequent parameter-modulating properties. In animals, the magnitude and even direction of changes that result from an affective state vary from individual to individual and comprise an important part of personality. We would expect to include the potential for such differences in the design of automata.

Finally, we need to consider the implications of adding reflective-level capacities. To do this, we have to enable the robot to have active expectations about outcomes and states of the world. In addition, it will have to be able to reflect on its own actions and states, a capacity that is critical for the formation of generalized knowledge, for abstraction, and for developing principles and new knowledge representations. Some of these representations (e.g., plans, goals, standards, and values) are themselves unique inhabitants of the reflective level, providing the basis for more fine-grained appraisals of emotion-inducing events and the material necessary for interpreting feelings as emotions.

Affect and Emotion

As soon as one raises the topic of affect and emotion in artifacts, one has to confront the probably unanswerable philosophical question of whether robots can have feelings (see Chapter 2). We choose to finesse this question by restricting our attention to the functional utility of affect and emotion. We view feeling as an awareness of a bodily state, a bodily disturbance, or some other bodily change. However, neither we nor anyone else know how to incorporate the experience of such awareness into an inanimate artifact.

With respect to the functional utility of affect, consider first the value of emotion recognition, a crucial capacity for the social aspect of effective functioning. Effective social functioning involves understanding the conditions under which it is or is not appropriate or prudent to interact with other individuals and when it is deemed appropriate, knowing what kind of interaction is expected. However, this ability to recognize, understand, and predict the

current affective state of others, *emotional intelligence* (e.g., Mayer & Salovey, 1997), is not the only determinant of effective social functioning. It is also advantageous to be able to make inferences from a model of the longer-term patterns of affective and motivational states, cognitions, and behavior—that is, from a model of the individual's personality. For example, our reflective characterization of a person as momentarily happy or sad and dispositionally moody or hyperactive contributes to the decisions we might make about our actions and interactions with respect to that person. Thus, a socially savvy robot will need to make inferences from behavior and outward manifestations of emotions (emotional expression), motivations, and cognition as well as from its model of the personality of others, when available, if it is to be capable of effective social functioning.

So, there are good reasons why a robot might need to recognize affect in others; now we need to ask why it might need affect itself. Our answer is that robots need affect for the same reason that humans do. One of the most fundamental functions of affect is as a valenced index of importance, and indeed, there is some neuroscientific evidence that affect is a prerequisite for establishing long-term memories (e.g., McGaugh, Cahill, Ferry, & Roozendaal, 2000). A second important function of affect is that it provides occasions for learning, from quite simple forms of reinforcement learning to complex, conscious planning and experimentation. Affect also has important consequences for the allocation of attention. It is a well-established finding in the psychological literature that negative affect tends to result in the focusing of attention on local details at the expense of global structure. Presumably, this is because in times of stress or threat it is important to be vigilant and to attend to local details, to identify sources of potential danger. Focusing attention on large-scale, global conditions of the environment is not likely to be conducive to these goals. However, such global focus is likely to be valuable in situations that are devoid of threat, danger, or potential harm. Consistent with this idea is the fact that under conditions of positive affect people do tend to focus on the big picture and to engage in more expansive information processing (Ashby, Isen, & Turken, 1999; Gasper & Clore, 2002). All of these (and indeed other) functions of affect are achieved through its capacity to change the operating characteristics of the other domains of functioning—motivation, cognition, and behavior. For example, the negative affect that results from the perception of a threat might modulate motivation by increasing the strength of a self-protecting action tendency, such as running away, relative to, say, an enjoyment-seeking action tendency, such as having a cocktail. Similarly, the affect might modulate cognition by interrupting ongoing cognitive processes and focusing attention on details of the current problem, and of course, it is almost bound to change the ongoing behavior.

CONCLUSION

We have presented a general model of effective functioning conceptualized in terms of three levels of processing (Fig. 7.1), in which four domains of functioning (affect, motivation, cognition, and behavior) are seen as integrated, nonseparable components. The reactive level is the home of rapid detection of states of the world and immediate responses to them. It uses pattern matching to recognize a set of situations and stimuli for which it is biologically prepared. These are essentially the unconditioned stimuli and associated responses of the simplest forms of classical conditioning. The reactive level is essential for mobilizing appropriate responses to the exigencies of the environment, and it can interrupt higher-level processing. The routine level is that of most motor behavior as well as procedural knowledge and automatic skills. It is a complex, rich information-processing and control system. It too interrupts higher-level processing when it encounters unexpected conditions, impasses, or emergencies or when conditions are novel or unknown. The reflective level is that of conscious attention, of higher-level cognitive processes and representations, and of cognitively elaborated, full-blown emotions. It is also the home of reflection and of knowledge about one's own knowledge and behavior. As such, this system continually performs high-level monitoring of ongoing activity at all three levels. The reflective level does not receive direct sensory input nor does it directly control responses: it can only potentiate or inhibit activity at the lower levels.

Within this three-level architecture, we have considered the way in which the four domains of functioning interact, with special attention to the way in which affect is manifested at the different levels. In many respects, labeling these continuous, complex feedback systems in terms of the four common distinctions of affect, motivation, cognition, and behavior is somewhat arbitrary. This is an integrated, holistic system that has evolved to facilitate effective functioning in a complex, dynamic environment. Nature does not necessarily make the sharp distinctions among these levels and domains that we make in order to talk about them. Affect, for example, ranges from proto-affect at the reactive level through primitive emotions at the routine level to full-blown emotions when augmented with the other domains at the reflective level. Thus, full-fledged emotions can involve feelings from the somatic and motor components of the reactive level, interacting with proto-affect from the reactive level and primitive emotions and feelings from the routine level together with cognitive elaboration from the reflective level. Reflective affect without some contribution from lower levels cannot be full-blown, "hot" emotion. For example, the cognitive components of anger without the concomitant feeling components from the lower levels would be what we might call "cold, rational anger." Similarly, the feeling of

primitive fear at the routine level is not a full-blown emotion because it lacks the requisite cognitive elaboration. It is only a feeling (albeit unpleasant) waiting to be "made sense of" by reflective-level processes.

As we indicated at the outset, the model that we have proposed is best thought of as a framework for thinking about how to conceptualize effective functioning. We believe that it is only by considering functioning at all three levels of processing and at all four domains of functioning that we can expect to achieve an understanding of effective functioning that might be useful for the design of fully autonomous robots and agents capable of responding appropriately to the huge array of problems and prospects that their environments might present.

Notes

We thank Ian Horswill, for his helpful comments on an early draft of this chapter, and Tony Z. Tang, for helpful discussions in the early stages of this work.

1. Although some investigators view cognition as a form of behavior (e.g., Fellous, 1999), we prefer to make a sharp distinction between the two.

2. We are well aware that talking of learning at this level simply in terms of classical conditioning is far too simplistic. Razran (1971) provides a brilliant discussion of the complexities of this issue.

3. Following Watson and Tellegen (1985) and others, we view positive affect and negative affect as (at least partially) independent dimensions.

4. Note that although initially bitter tastes are rejected, the system can adapt so that, with sufficient experience, it no longer responds quite so vehemently. Indeed, the higher levels might interpret the taste positively and actively inhibit the lower response—hence, the learned preference for many bitter and otherwise initially rejected foods such as alcoholic beverages and spicy sauces.

5. The key feature of specifying emotions (and emotion-like states, such as 1–4) in this way is that they are characterized in terms of their eliciting conditions with minimal dependence on the use of emotion words. The advantage of doing this can be seen by considering that English does not have a good word to express the affective state characterized by *3*, a positive feeling about a potential good thing. Something like "anticipatory excitement" is much closer to the state than "hope," even though in English hope is usually opposed to fear. In any case, we think it misleading to use conventional emotion names to refer to primitive forms of emotion.

6. By "internal" cue to the reactive level we mean internal to the organism but still external to the reactive-level mechanisms. Thus, in the case of hunger, the internal cue to the reactive level comes from the hunger system.

7. In fact, our preference would be to view boldness and shyness as two independent, unipolar dimensions rather than one bipolar dimension. We also suspect that *timidity* is a better term to capture the construct because it avoids the social connotations of "shyness."

8. Having a high level of positive affect does not mean that the individual is always happy. It means that the median value of positive affective responses is higher for this individual than for most others. The same is true for approach behaviors (indeed, for everything).

References

Arnold, M. (1960). *Emotions and personality* (Vols. I, II). New York: Columbia University Press.

Ashby, F. G., Isen, A. M., & Turken, A. U. (1999). A neuropsychological theory of positive affect and its influence on cognition. *Psychological Review*, *106*, 529–550.

Atkinson, J. W., & Birch, D. (1970). *The dynamics of action*. New York: Wiley.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance. *Personnel Psychology*, *44*, 1–26.

Broadbent, D. E. (1971). *Decision and stress*. London: Academic Press.

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. A. (2002). *Flesh and machines: How robots will change us*. New York: Pantheon.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review*, *1*, 3–25.

Clore, G. L., & Ortony, A. (2000). Cognition in emotion: Always, sometimes, or never? In L. Nadel, R. Lane, & G. L. Ahern (Eds.), *The cognitive neuroscience of emotion*. New York: Oxford University Press.

Coleman, K., & Wilson, D. S. (1998). Shyness and boldness in pumpkinseed sunfish: Individual differences are context-specific. *Animal Behaviour*, *56*, 927–936.

Damasio, A. (1994). *Descarte's error: Emotion, reason, and the human brain*. New York: Putnam.

Damasio, A. R. (2000). A second chance for emotion. In L. Nadel, R. Lane, & G. L. Ahern (Eds.), *The cognitive neuroscience of emotion*. New York: Oxford University Press.

deWaal, F. B. M., & Berger, M. L. (2000). Payment for labour in monkeys. *Nature*, *404*, 563.

Elliot, A. J., & Thrash, T. M. (2002). Approach–avoidance motivation in personality: Approach–avoidance temperaments and goals. *Journal of Personality and Social Psychology*, *82*, 804–818.

Fellous, J.-M. (1999). The neuromodulatory basis of emotion. *The Neuroscientist*, *5*, 283–294.

Gasper, K., & Clore, G. L. (2002). Attending to the big picture: Mood and global versus local processing of visual information. *Psychological Science*, *13*, 34–40.

Goldberg, E. (2001). *The executive brain: Frontal lobes and the civilized mind*. New York: Oxford University Press.

Gray, J. A. (1990). Brain systems that mediate both emotion and cognition. *Cognition & Emotion*, *4*, 269–288.

Halliday, T. R. (1980). Motivational systems and interactions between activities. In F. Toates & T. R. Halliday (Eds.), *Analysis of motivational processes* (pp. 205–220). London: Academic Press.

Higgins, E. T. (2000). Does personality provide unique explanations for behavior? Personality as cross-person variability in general principles. *European Journal of Personality*, *14*, 391–406.

James, W. (1884). What is an emotion? *Mind*, *9*, 188–205.

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–139). New York: Guilford.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to alternatives. *Psychological Review*, *93*, 136–153.

Kim, H., Somerville, L. H., Johnstone, T., Alexander, A., & Whalen, P. J. (2004). Inverse amygdala and medial prefrontal cortex responses to surprised faces. *Neuroreport*, *14*, 2317–2322.

Kluckholm, C., & Murray, H. (1953). *Personality in nature, society, and culture*. New York: Knopf.

Lane, R. (2000). Neural correlates of conscious emotional experience. In L. Nadel, R. Lane, & G. L. Ahern (Eds.). *The cognitive neuroscience of emotion*. New York: Oxford University Press.

Lange, G. (1912). The mechanism of the emotions. In B. Rand (Ed., Trans.), *The classical psychologists: Selections illustrating psychology from Anaxagoras to Wundt* (pp. 672–684). Boston: Houghton Mifflin. (Original work published 1895)

Lazarus, R. S. (1966). *Psychological stress and the coping process*. New York: McGraw-Hill.

LeDoux, J. (1996). *The emotional brain*. New York: Simon & Schuster.

MacLean, P. D. (1990). *The triune brain in evolution*: Plenum.

Mandler, G. (1984). *Mind and body*. New York: Wiley.

Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? In P. Salovey & D. Sluyter (Eds.), *Emotional development and emotional intelligence: Implications for educators* (pp. 3–31). New York: Basic Books.

McAdams, D. P. (2001). The psychology of life stories. *Review of General Psychology*, *5*, 100–122.

McGaugh, J. L., Cahill, L., Ferry, B., & Roozendaal, R. (2000). Brain systems and the regulation of memory consolidation. In J. J. Bolhuis (Ed.), *Brain, perception, memory: Advances in cognitive neuroscience* (pp. 233–251). London: Oxford University Press.

Minsky, M. (in preparation). *The emotion machine*.

Nebylitsyn, V. D., & Gray, J. A. (1972). *The biological basis of individual behavior*. New York: Academic Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self regulation: Advances in research* (Vol. IV). New York: Plenum.

Öhman, A., Flykt, A., & Lundqvist, A. (2000). Unconscious emotion: Evolutionary perspectives, psychophysiological data and neuropsychological mechanisms. In L. Nadel, R. Lane, & G. L. Ahern (Eds.), *The cognitive neuroscience of emotion*. New York: Oxford University Press.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. New York: Cambridge University Press.

Ortony, A., Clore, G. L., & Foss, M. A. (1987). The referential structure of the affective lexicon. *Cognitive Science, 11*, 341–364.

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review, 97*, 315–331.

Pavlov, I. P. (1930). A brief outline of the higher nervous activity. In C. A. Murchinson (Ed.), *Psychologies of 1930*, Worchester, MA: Clark University Press.

Razran, G. (1971). *Mind in evolution*. Boston: Houghton Mifflin.

Revelle, W. (1986). Motivation and efficiency of cognitive performance. In D. R. Brown & J. Veroff (Eds.), *Frontiers of motivational psychology: Essays in honor of John W. Atkinson*. Berlin: Springer.

Revelle, W. (1993). Individual differences in personality and motivation: "Non-cognitive" determinants of cognitive performance. In A. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness and control. A tribute to Donald Broadbent*. Oxford: Oxford University Press.

Robinson, D. L. (1996). *Brain, mind, and behavior: A new perspective on human nature*. London: Praeger.

Robinson, D. L. (2001). How brain arousal systems determine different temperament types and the major dimensions of personality. *Personality and Individual Differences, 31*, 1233–1259.

Rolls, E. T. (1999). *The brain and emotion*. New York: Oxford University Press.

Roseman, I. J. (1984). Cognitive determinants of emotions: A structural theory. In P. Shaver (Ed.), *Review of personality and social psychology* (Vol. 5). Beverly Hills: Sage.

Rosenbloom, P. S., Laird, J. E., & Newell, A. (1993). *The SOAR papers*. Boston: MIT Press.

Sanders, A. F. (1986). Energetical states underlying task performance. In G. R. J. Hockey, A. W. K. Gaillard, & M. G. H. Coles (Eds.), *Energetics and human information processing* (pp. 139–154). Dordrecht: Matinus Nijhoff.

Scherer, K. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer and P. Ekman (Eds.), *Approaches to emotion*. Hillsdale, NJ: Erlbaum.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*, 1–66.

Schneirla, T. (1959). An evolutionary and developmental theory of biphasic pro-

cesses underlying approach and withdrawal. In *Nebraska symposium on motivation* (pp. 27–58). Lincoln: University of Nebraska Press.

Sloman, A., & Logan, B. (2000). Evolvable architectures for human-like minds. In G. Hatano, N. Okada, & H. Tanabe (Eds.), *Affective minds* (pp. 169–181). Amsterdam: Elsevier.

Strelau, J. (1985). Temperament and personality: Pavlov and beyond. In J. Strealu, F. H. Farley, & A. Gale (Eds.), *The biological bases of personality and behavior: Psychophysiology, performance, and application*. Washington, DC: Hemisphere.

Watson, D. (2000). *Mood and temperament*. New York: Guilford.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin, 98*, 219–235.

Zeigarnik, B. (1967). On finished and unfinished tasks. In W. D. Ellis (Ed.), *A source book of Gestalt psychology*. New York: Humanities. (Original work published 1927)

8 The Architectural Basis of Affective States and Processes

AARON SLOMAN, RON CHRISLEY, AND MATTHIAS SCHEUTZ

Much discussion of emotions and related topics is riddled with confusion because different authors use the key expressions with different meanings. Some confuse the concept of "emotion" with the more general concept of "affect," which covers other things besides emotions, including moods, attitudes, desires, preferences, intentions, dislikes, etc. Moreover, researchers have different goals: some are concerned with understanding natural phenomena, while others are more concerned with producing useful artifacts, e.g., synthetic entertainment agents, sympathetic machine interfaces, and the like. We address this confusion by showing how "architecture-based" concepts can extend and refine our folk-psychology concepts in ways that make them more useful both for expressing scientific questions and theories, and for specifying engineering objectives. An implication is that different information-processing architectures support different classes of emotions, different classes of consciousness, different varieties of perception, and so on. We start with high-level concepts applicable to a wide variety of natural and artificial systems, including very simple organisms—namely, concepts such as "need," "function," "information-user," "affect," and "information-processing architecture." For more complex architectures, we offer the CogAff schema as a generic framework that distinguishes types of components that may be in an architecture, operating concurrently with different functional roles. We also sketch H-CogAff, a richly featured special

case of CogAff, conjectured as a type of architecture that can explain or replicate human mental phenomena. We show how the concepts that are definable in terms of such architectures can clarify and enrich research on human emotions. If successful for the purposes of science and philosophy, the architecture is also likely to be useful for engineering purposes, though many engineering goals can be achieved using shallow concepts and shallow theories, e.g., producing "believable" agents for computer entertainments. The more human-like robot emotions will emerge, as they do in humans, from the interactions of many mechanisms serving different purposes, not from a particular, dedicated "emotion mechanism."

Many confusions and ambiguities bedevil discussions of emotion. As a way out of this, we present a view of mental phenomena, in general, and the various sorts of things called "emotions," in particular, as states and processes in an information-processing architecture. Emotions are a subset of affective states. Since different animals and machines can have different kinds of architecture capable of supporting different varieties of state and process, there will be different families of such concepts, depending on the architecture. For instance, if human infants, cats, or robots lack the sort of architecture presupposed by certain classes of states (e.g., obsessive ambition, being proud of one's family), then they cannot be in those states. So the question of whether an organism or a robot needs emotions or needs emotions of a certain type reduces to the question of what sort of information-processing architecture it has and what needs arise within such an architecture.

NEEDS, FUNCTIONS, AND FUNCTIONAL STATES

The general notion of *X* having a need does not presuppose a notion of goal or purpose but merely refers to necessary conditions for the truth of some statement about *X*, *P*(*X*). In trivial cases, *P*(*X*) could be "*X* continues to exist," and in less trivial cases, something like "*X* grows, reproduces, avoids or repairs damage." All needs are relative to something for which they are necessary conditions. Some needs are indirect insofar as they are necessary for something else that is needed for some condition to hold. A need may also be relative to a context since *Y* may be necessary for *P*(*X*) only in some contexts. So "*X* needs *Y*" is elliptical for something like "There is a context, C, and there is a possible state of affairs, *P*(*X*), such that, in C, *Y* is necessary

for *P*(*X*)." Such statements of need are actually shorthand for a complex collection of counterfactual conditional statements about what would happen if . . ."

Parts of a system have a function in that system if their existence helps to serve the needs of the system, under some conditions. In those conditions the parts with functions are sufficient, or part of a sufficient condition for the need to be met. Suppose *X* has a need, *N*, in conditions of type C—i.e., there is a predicate, *P*, such that in conditions of type C, *N* is necessary for *P*(*X*). Suppose moreover that O is an organ, component, state, or subprocess of *X*. We can use *F*(O,*X*,C,*N*) as an abbreviation for "In contexts of type C, O has the function, *F*, of meeting *X's* need, *N*—i.e., the function of producing satisfaction of that necessary condition for *P*(*X*)." This actually states "In contexts of type C the existence of O, in the presence of the rest of *X*, tends to bring about states meeting the need, *N*; or tends to preserve such states if they already exist; or tends to prevent things that would otherwise prevent or terminate such states." Where sufficiency is not achievable, a weaker way of serving the need is to make the necessary condition *more likely* to be true.

This analysis rebuts arguments (e.g., Millikan, 1984) that the notion of function has to be explicated in terms of evolutionary or any other history since the causal relationships summarized above suffice to support the notion of function, independently of how the mechanism was produced.

We call a state in which something is performing its function of serving a need, a *functional state*. Later we will distinguish desire-like, belief-like, and other sorts of functional states (Sloman, 1993). The label "affective" as generally understood seems to be very close to this notion of a desire-like state and subsumes a wide variety of more specific types of affective state, including the subset we will define as "emotional."

Being able to serve a function by producing different behaviors in the face of a variety of threats and opportunities minimally requires (*1*) sensors to detect when the need arises, if it is not a constant need; (*2*) sensors to identify aspects of the context which determine what should be done to meet the need, for instance, in which direction to move or which object to avoid; and (*3*) action mechanisms that combine the information from the sensors and deploy energy to meet the need. In describing components of a system as sensors or selection mechanisms, we are ascribing to them functions that are analyzable as complex dispositional properties that depend on what would happen in various circumstances.

Combinations of the sensor states trigger or modulate activation of need-supporting capabilities. There may, in some systems, be conflicts and conflict-resolution mechanisms (e.g., using weights, thresholds, etc.). Later, we will see how the processes generated by sensor states may be purely reactive in

some cases and in other cases deliberative, i.e., mediated by a mechanism that represents possible sequences of actions, compares them, evaluates them, and makes selections on that basis before executing the actions.

We can distinguish sensors that act as need-sensors from those that act as fact-sensors. *Need-sensors* have the function of initiating action or tending to initiate action (in contexts where something else happens to get higher priority) to address a need, whereas *fact-sensors* do not, though they can modify the effects of need sensors. For most animals, merely sensing the fact of an apple on a tree would not in itself initiate any action relating to the apple. However, if a need for food has been sensed, then that will (unless overridden by another need) initiate a process of seeking and consuming food. In that case, the factual information about the apple could influence which food is found and consumed.

The very same fact-sensor detecting the very same apple could also modify a process initiated by a need to deter a predator; in that case, the apple could be selected for throwing at the predator. In this case, we can say that the sensing of the apple has no motivational role. It is a "belief-like" state, not a "desire-like" state.

INFORMATION-PROCESSING ARCHITECTURES

The *information-processing architecture* of an organism or other object is the collection of information-processing mechanisms that together enable it to perform in such a way as to meet its needs (or, in "derivative" cases, could enable it to meet the needs of some larger system containing it).

Describing an architecture involves (recursively) describing the various parts and their relationships, including the ways in which they cooperate or interfere with one another. Systems for which there are such true collections of statements about what they would do to meet needs under various circumstances can be described as having control states, of which the belief-like and desire-like states mentioned previously (and defined formally below) are examples. In a complex architecture, there will be many concurrently active and concurrently changing control states.

The components of an architecture need not be physical: physical mechanisms may be used to implement virtual machines, in which nonphysical structures such as symbols, trees, graphs, attractors, and information records are constructed and manipulated. This idea of a virtual machine implemented in a physical machine is familiar in computing systems (e.g., running word processors, compilers, and operating systems) but is equally applicable to organisms that include things like information stores, concepts, skills, strategies,

desires, plans, decisions, and inferences, which are not physical objects or processes but are implemented in physical mechanisms, such as brains.[1]

Information-processing virtual machines can vary in many dimensions, for example, the number and variety of their components, whether they use discretely or continuously variable substates, whether they can cope with fixed or variable complexity in information structures (e.g., vectors of values versus parse trees), the number and variety of sensors and effectors, how closely internal states are coupled to external processes, whether processing is inherently serial or uses multiple concurrent and possibly asynchronous subsystems, whether the architecture itself can change over time, whether the system builds itself or has to be assembled by an external machine (like computers and most current software), whether the system includes the ability to observe and evaluate its own virtual-machine processes or not (i.e., whether it includes "meta-management" as defined by Beaudoin, 1994), whether it has different needs or goals at different times, how conflicts are detected and resolved, and so on.

In particular, whereas the earliest organisms had sensors and effectors directly connected so that all behaviors were totally reactive and immediate, evolution "discovered" that, for some organisms in some circumstances, there are advantages in having an indirect causal connection between sensed needs and the selections and actions that can be triggered to meet the needs, i.e., an intermediate state that "represents" a need and is capable of entering into a wider variety of types of information processing than simply triggering a response to the need.

Such intermediate states could allow (1) different sensors to contribute data for the same need; (2) multifunction sensors to be redirected to gain new information relevant to the need (looking in a different direction to check that enemies really are approaching); (3) alternative responses to the same need to be compared; (4) conflicting needs to be evaluated, including needs that arise at different times; (5) actions to be postponed while the need is remembered; (6) associations between needs and ways of meeting them to be learned and used, and so on.

This seems to capture the notion of a system having goals as well as needs. Having a *goal* is having an enduring representation of a need, namely, a representation that can persist after sensor mechanisms are no longer recording the need and that can enter into diverse processes that attempt to meet the need.

Evolution also produced organisms that, in addition to having need sensors, had fact sensors that produced information that could be used for varieties of needs, i.e., "percepts" (closely tied to sensor states) and "beliefs," which are indirectly produced and can endure beyond the sensor states that produce them.

DIRECT AND MEDIATED CONTROL STATES AND REPRESENTATIONS

The use of intermediate states explicitly representing needs and sensed facts requires extra architectural complexity. It also provides opportunities for new kinds of functionality (Scheutz, 2001). For example, if need representations and fact representations can be separated from the existence of sensor states detecting needs and facts, it becomes possible for such representations to be derived from other things instead of being directly sensed. The derived ones can have the same causal powers, i.e., helping to activate need-serving capabilities. So, we get derived desires and derived beliefs. However, all such derivation mechanisms can, in principle, be prone to errors (in relation to their original biological function), for instance, allowing desires to be derived which, if acted on, serve no real needs and may even produce death, as happens in many humans.

By specifying architectural features that can support states with the characteristics associated with concepts like "belief", "desire", and "intention", we avoid the need for what Dennett (1978) calls "the intentional stance," which is based on an assumption of rationality, as is Newell's (1990) "knowledge level." Rather, we need only what Dennett (1978) calls "the design stance," as explained by Sloman (2002). However, we lack a systematic overview of the space of relevant architectures. As we learn more about architectures produced by evolution, we are likely to discover that the architectures we have explored so far form but a tiny subset of what is possible.

We now show how we can make progress in removing, or at least reducing, conceptual confusions regarding emotions (and other mental phenomena) by paying attention to the diversity of architectures and making use of architecture-based concepts.

EMOTION AS A SPECIAL CASE OF AFFECT

A Conceptual Morass

Much discussion of emotions and related topics is riddled with confusion because the key words are used with different meanings by different authors, and some are used inconsistently by individuals. For instance, many researchers treat all forms of motivation, all forms of evaluation, or all forms of reinforcing reward or punishment as emotions. The current confusion is summarized aptly below:

> There probably is no scientifically appropriate class of things referred to by our term emotion. Such disparate phenomena—fear, guilt,

> shame, melancholy, and so on—are grouped under this term that it is dubious that they share anything but a family resemblance. (Delancey, 2002)[2]

The phenomena are even more disparate than that suggests. For instance, some people would describe an insect as having emotions such as fear, anger, or being startled, whereas others would deny the possibility. Worse still, when people disagree as to whether something does or does not have emotions (e.g., whether a fetus can suffer), they often disagree on what would count as evidence to settle the question. For instance, some, but not all, consider that behavioral responses determine the answer; others require certain neural mechanisms to have developed; and some say it is merely a matter of degree and some that it is not a factual matter at all but a matter for ethical decision.

Despite the well-documented conceptual unclarity, many researchers still assume that the word *emotion* refers to a generally understood and fairly precisely defined collection of mechanisms, processes, or states. For them, whether (some) robots should or could have emotions is a well-defined question. However, if there really is no clear, well-defined, widely understood concept, it is not worth attempting to answer the question until we have achieved more conceptual clarity.

Detailed analysis of pretheoretical concepts (folk psychology) can make progress using the methods of conceptual analysis explained in Chapter 4 of Sloman (1978), based on Austin (1956). However, that is not our main purpose.

Arguing about what emotions really are is pointless: "emotion" is a cluster concept (Sloman, 2002), which has some clear instances (e.g., violent anger), some clear non-instances (e.g., remembering a mathematical formula), and a host of indeterminate cases on which agreement cannot easily be reached. However, something all the various phenomena called emotions seem to have in common is membership of a more general category of phenomena that are often called *affective*, e.g., desires, likes, dislikes, drives, preferences, pleasures, pains, values, ideals, attitudes, concerns, interests, moods, intentions, etc., the more enduring of which can be thought of as components of personality, as suggested by Ortony (2002; see also chapter 7, Ortony et al.).

Mental phenomena that would not be classified as affective include perceiving, learning, thinking, reasoning, wondering whether, noticing, remembering, imagining, planning, attending, selecting, acting, changing one's mind, stopping or altering an action, and so on. We shall try to clarify this distinction below.

It may be that many who are interested in emotions are, unwittingly, interested in the more general phenomena of affect (Ortony, 2002). This would account for some of the overgeneral applications of the label "emotion."

Toward a Useful Ontology for a Science of Emotions

How can emotion concepts and other concepts of mind be identified for the purposes of science? Many different approaches have been tried. Some concentrate on externally observable expressions of emotion. Some combine externally observable eliciting conditions with facial expressions. Some of those who look at conditions and responses focus on physically describable phenomena, whereas others use the ontology of ordinary language, which goes beyond the ontology of the physical sciences, in describing both environment and behavior (e.g., using the concepts threat, opportunity, injury, escape, attack, prevent, etc.). Some focus more on internal physiological processes, e.g., changes in muscular tension, blood pressure, hormones in the bloodstream, etc. Some focus more on events in the central nervous system, e.g., whether some part of the limbic system is activated.

Many scientists use shallow specifications of emotions and other mental states defined in terms of correlations between stimuli and behaviors because they adopt an out-of-date empiricist philosophy of science that does not acknowledge the role of theoretical concepts going beyond observation (for counters to this philosophy, see Lakatos, 1970, and Chapter 2 of Sloman, 1978).

Diametrically opposed to this, some define *emotion* in terms of introspection-inspired descriptions of what it is like to have one (e.g., Sartre, 1939, claims that having an emotion is "seeing the world as magical"). Some novelists (e.g., Lodge, 2002) think of emotions as defined primarily by the way they are expressed in thought processes, for instance, thoughts about what might happen; whether the consequences will be good or bad; how bad consequences may be prevented; whether fears, loves, or jealousy will be revealed; and so on. Often, these are taken to be thought processes that cannot be controlled.

Nobody knows exactly how pretheoretical folk psychology concepts of mind work. We conjecture that they are partly architecture-based concepts: people implicitly presuppose an information-processing architecture (incorporating percepts, desires, thoughts, beliefs, intentions, hopes, fears, etc.) when they think about others, and they use concepts that are implicitly defined in terms of what can happen in that architecture. For purposes of scientific explanation, those naive architectures need to be replaced with deeper and richer explanatory architectures, which will support more precisely defined concepts. If the naive architecture turns out to correspond to some aspects of the new architecture, this will explain how naive theories and concepts are useful precursors of deep scientific theories, as happens in most sciences.

A Design-Based Ontology

We suggest that "emotion" is best regarded as an imprecise label for a subset of the more general class of affective states. We can use the ideas introduced in the opening section to generate architecture-based descriptions of the variety of states and processes that can occur in different sorts of natural and artificial systems. Then, we can explore ways of carving up the possibilities in a manner that reflects our pretheoretical folk psychology constrained by the need to develop explanatory scientific theories.

For instance, we shall show how to distinguish affective states from other states. We shall also show how our methodology can deal with more detailed problems, for instance, whether the distinction between emotion and motivation collapses in simple architectures (e.g., see Chapter 7, Ortony et al.). We shall show that it does not collapse if emotions are defined in terms of one process interrupting or modulating the "normal" behavior of another.

We shall also see that where agents (e.g., humans) have complex, hybrid information-processing architectures involving a variety of types of subarchitectures, they may be capable of having different sorts of emotion, percept, desire, or preference according to which portions of the architecture are involved. For instance, processes in a reactive subsystem may be insect-like (e.g., being startled), while other processes (e.g., long-term grief and obsessive jealousy) go far beyond anything found in insects. This is why, in previous work, we have distinguished primary, secondary, and tertiary emotions[3] on the basis of their architectural underpinnings: *primary* emotions (e.g., primitive forms of fear) reside in a reactive layer and do not require either the ability to represent possible but non-actual states of the world, or hypothetical reasoning abilities; *secondary* emotions (e.g., worry, i.e., fear about possible future events) intrinsically do, and for this, they need a deliberative layer; *tertiary* emotions (e.g., self-blame) need, in addition, a layer ("meta-management") that is able to monitor, observe, and to some extent oversee processing in the deliberative layer and other parts of the system. This division into three architectural layers is only a rough categorization as is the division into three sorts of emotion (we will elaborate more in a later section). Further sub-divisions are required to cover the full variety of human emotions, especially as emotions can change their character over time as they grow and subside (as explained in Sloman, 1982). A similar theory is presented in a draft of *The Emotion Machine* (Minsky, 2003).

This task involves specifying information-processing architectures that can support the types of mental state and process under investigation. The catch is that different architectures support different classes of emotion, different classes of consciousness, different varieties of perception, and

different varieties of mental states in general—just as some computer-operating system architectures support states like "thrashing," where more time is spent swapping and paging than doing useful work, whereas other architectures do not, for instance, if they do not include virtual memory or multi processing mechanisms.

So, to understand the full variety of types of emotions, we need to study not just human-like systems but alternative architectures as well, to explore the varieties of mental states they support. This includes attempting to understand the control architectures found in many animals and the different stages in the development of human architectures from infancy onward. Some aspects of the architecture will also reflect evolutionary development (Sloman, 2000a; Scheutz & Sloman, 2001).

VARIETIES OF AFFECT

What are affective states and processes? We now explain the intuitive affective/nonaffective distinction in a general way. Like *emotion*, *affect* lacks any generally agreed upon definition. We suggest that what is intended by this notion is best captured by our architecture-based notion of a desire-like state, introduced earlier in contrast with belief-like and other types of nonaffective state. Desire-like and belief-like states are defined more precisely below.

Varieties of Control States

Previously, we introduced the notion of a control state, which has some function that may include preserving or preventing some state or process. An individual's being in such a state involves the truth of some collection of counterfactual conditional statements about what the individual would do in a variety of possible circumstances.

We define *desire-like* states as those that have the function of detecting needs so that the state can act as an initiator of action designed to produce or prevent changes in a manner that serves the need. This can be taken as a more precise version of the intuitive notion of *affective state*. These are states that involve dispositions to produce or prevent some (internal or external) occurrence related to a need. It is an old point, dating at least back to the philosopher David Hume (1739/1978), that an action may be based on many beliefs and derivatively affective states but must have some intrinsically affective component in its instigation. In our terminology, no matter how many beliefs, percepts, expectations, and reasoning skills a machine or organism has, they will not cause it to do one thing rather than another or even

to do anything at all, unless it also has at least one desire-like state. In the case of physical systems acted on by purely physical forces, no desire-like state is needed. Likewise, a suitably designed information processing machine may have actions initiated by external agents, e.g., commands from a user, or a "boot program" triggered when it is switched on. Humans and other animals may be partly like that insofar as genetic or learned habits, routines, or reflexes permit something sensed to initiate behavior. This can happen only if there is some prior disposition that plays the role of a desire-like state, albeit a very primitive one. As we'll see later in connection with depression, some desire-like states can produce dysfunctional behaviors.

Another common use of *affective* implies that something is being experienced as pleasant or unpleasant. We do not assume that connotation, partly because it can be introduced as a special case and partly because we are using a general notion of affect (desire-like state) that is broad enough to cover states of organisms and machines that would not naturally be described as experiencing anything as pleasant or unpleasant, and also states and processes of which humans are not conscious. For instance, one can be jealous or infatuated without being conscious or aware of the jealousy or infatuation. Being conscious of one's jealousy, then, is a "higher-order state" that requires the presence of another state, namely, that of being jealous. Sloman and Chrisley (2003) use our approach to explain how some architectures support experiential states.

Some people use *cognitive* rather than "*non-affective*," but this is undesirable if it implies that affective states cannot have rich semantic content and involve beliefs, percepts, etc., as illustrated in the apple example above. Cognitive mechanisms are required for many affective states and processes.

Affective versus Nonaffective (What To Do versus How Things Are)

We can now introduce our definitions.

- A *desire-like* state, *D*, of a system, *S*, is one whose function it is to get *S* to do something to preserve or to change the state of the world, which could include part of *S* (in a particular way dependent on *D*). Examples include preferences, pleasures, pains, evaluations, attitudes, goals, intentions, and moods.
- A *belief-like* state, *B*, of a system, *S*, is one whose function is to provide information that could, in combination with one or more desire-like states, enable the desire-like states to fulfill their functions. Examples include beliefs (particular and general), percepts, memories, and fact-sensor states.

Primitive sensors provide information about some aspect of the world simply because the information provided varies as the world changes (another example of sets of counterfactual conditional statements). Insofar as the sensors meet the need of providing correct information, they also serve a desire-like function, namely, to "track the truth" so that the actions initiated by other desire-like states serving other needs can be appropriate to meeting those needs. In such cases, the state *B* will include mechanisms for checking and maintaining the correctness of *B*, in which case there will be, as part of the mechanisms producing the belief-like state, sub-mechanisms whose operation amounts to the existence of another desire-like state, serving the need of keeping *B* true and accurate. In a visual system, this could include vergence control, focus control, and tracking.

In these cases, *B* has a dual function: the primary belief-like function of providing information and the secondary desire-like function of ensuring that the system is in state *B* only when the content of *B* actually holds (i.e., that the information expressed in *B* is correct and accurate.) The secondary function is a means to the first. Hence, what are often regarded as non-desire-like states can be seen as including a special subclass of desire-like states.

We are not assuming that these states have propositional content in the sense in which propositional content can be expressed as predicates applied to arguments or expressed in natural language. On the contrary, an insect which has a desire-like state whose function is to get the insect to find food need not have anything that could be described as a representation or encoding of "I need food." Likewise, the percepts and beliefs (belief-like states) of an insect need not be expressible in terms of propositions. Similar comments could be made about desire-like and belief-like states in evolutionarily old parts of the human information-processing architecture. Nevertheless, the states should have a type of semantic content for which the notion of truth or correspondence with reality makes sense (Sloman, 1996).

In describing states as having functions, we imply that their causal connections are to some extent reliable. However, this is consistent with their sometimes being suppressed or overridden by other states in a complex information-processing system. For instance, although it is the function of a belief-like state to "track the truth," a particular belief may not be removed by a change in the environment if the change is not perceived or if something prevents the significance of a perceived change being noticed. Likewise, the desire to achieve something need not produce any process tending to bring about the achievement if other, stronger desires dominate, if attention is switched to something else, or if an opportunity to achieve what is desired is not recognized. So, all of these notions have interpretations that depend heavily on complex collections of counterfactual conditionals being true: they are inherently dispositional concepts (see also the

discussion of the belief–desire–intention models of teamwork in Nair et al.'s Chapter 11).

Our distinction is closely related to the old notion familiar to philosophers that both desires and beliefs can represent states of the world but they differ in the "direction of fit." When there is a mismatch, beliefs tend to be changed to produce a match (fit) and desires tend to cause something else in the world to be changed to produce or preserve a match:

- A change in the world tends to cause a change in beliefs.
- A change in desires tends to cause a change in the world.

Here, the "world" can include states of the organism.

Belief-like and desire-like states exhaust the variety of possible information states in simple organisms and machines, but in more sophisticated architectures, there are subsystems that provide states that are neither desire-like nor belief-like. Examples include states in which possibilities are contemplated but neither desired nor believed, for instance, in planning or purposeless daydreaming (imagination-like and plan-like states; Sloman, 1993) or some kinds of artistic activity. Such activities have requirements that overlap with requirements for producing belief-like and desire-like states. For instance they require possession of a collection of concepts and mechanisms for manipulating representations. Language considerably enhances such capabilities.

In other words, the evolution of sophisticated belief-like and desire-like states required the evolution of mechanisms whose power could also be harnessed for producing states that are neither. Such resources can then produce states that play a role in more complex affective states and processes even though they are not themselves affective. For instance, the ability to generate a certain sort of supposition might trigger states that are desire-like (e.g., disgust or desire) or belief-like (e.g., being reminded of something previously known). What we refer to as secondary and tertiary emotions can also use such mechanisms.

Positive versus Negative Affect

There are many further distinctions that can be made among types of affective state. Among the class of affective (i.e., desire-like) states, we can distinguish positive and negative cases, approximately definable as follows:

- Being in a state *N* of a system *S* is a negatively affective state if being in *N* or moving toward *N* changes the dispositions of *S* so as to cause processes that reduce the likelihood of *N* persisting or tend to resist processes that bring *N* into existence.

- Being in a state *P* of a system *S* is a positively affective state if being in *P* or moving toward *P* changes the dispositions of *S* so as to cause processes that increase the likelihood of *P* persisting or tend to produce or enhance processes that bring *P* into existence or maintain the existence of *P*.

For example, being in pain is negatively affective since it tends to produce actions that remove or reduce the pain. Enjoying eating an apple is positively affective since it involves being in a state that tends to prolong the eating and tends to resist things that would interfere with the eating. In both cases, the effects of the states can be overridden by other factors including further states involving mechanisms that tend to suppress or remove the affective states, such as satiety mechanisms in animals; that is why the definitions have to be couched in terms of dispositions, not actual effects. For instance, masochistic mechanisms can produce pain-seeking behavior, and various kinds of religious indoctrination can cause states of pleasure to produce guilt feelings that interfere with those states.

There are many subdivisions and special cases that would need to be discussed in a more complete analysis of information-processing systems with affective and nonaffective states. In particular, various parts of the above definitions could be made more precise. We could also add further details, such as defining the intensity of an affective state, which might involve things like its ability to override or be overridden by other affective states and perhaps how many parts of the overall system it affects. Here, we mention only three important points.

First, we can distinguish direct and mediated belief-like and desire-like states. This amounts to a distinction between states without and with an explicit instantiation in some information structure that the system can create, inspect, modify, store, retrieve, or remove. If the state is merely implicit (i.e., direct, unmediated), then the information state cannot be created or destroyed while leaving the rest of the system unchanged.

In other words, explicit mental states are instantiated in, but are not part of, the underlying architecture (although they can be acquired and represented within it), whereas implicit mental states are simply states of the architecture that have certain effects. Note that "explicit" does not mean "conscious", as it is possible for a system to have an explicit instantiation of an information structure without being aware of it (i.e., while the information structure is used by some process, there is no process that notices or records its presence).

Second, some belief-like states and desire-like states are derivative substates in that they result from a process that uses something like premises (i.e., preexisting explicit/mediated states) and a derivation of a new explic-

itly represented state. Others are nonderivative substates because they are produced without any process of reasoning or derivation of one representation from others but merely arise out of activation of internal or external sensors and their effects on other subsystems. Derivative states, as defined here, are necessarily also explicit (but not necessarily conscious). The derivative ones might also be described as "rational" and the nonderivative ones as "nonrational" insofar as the former, but not the latter, are produced by possibly very primitive reasoning processes.

A third point concerns a causal connection between two states that does not include explicit reasoning but something more like reinforcement learning. For example, associative learning may bring about a certain type of action, *A*, as the "content" of a desire-like state, *S*, because *S* is repeatedly followed by a previously desired state, *S'*. Thus, *S*, in which *A* is desired, arises because *A* has been found to be a means to *S'*. For instance, a rat can be trained to press a lever because that has been associated with acquiring food. Thus a desire-like state that tends to cause food-seeking might produce a desire-like state whose content is pressing the lever. This does not require the rat to have an explicit belief that pressing the lever causes food, from which it infers the result of pressing the lever. Having such a belief would support a different set of possible mental processes from the set supported by the mere learned desirability of pressing the lever. For instance, the explicit belief could be used in making predictions as well as selecting actions.

Likewise, a result of associative learning may be that a particular kind of sensory stimulation produces a belief-like state because the organism has learned to associate the corresponding situations with those stimuli. For instance, instead of only the sound or smell of food producing the belief or expectation that food will appear, the perception of the lever going down could produce that belief.

In summary, we have distinguished merely associatively triggered belief-like and desire-like states from those that are derived by a process of reasoning, making use of explicit representations rather than simply the causal consequences of implicit desire-like and belief-like states. The distinction between derivative and associative affective states will later be of assistance when distinguishing between different kinds of emotion.

Positive and Negative Affect and Learning

We have defined positive and negative affective states in terms of tendencies or dispositions to achieve/preserve (positive) or avoid/remove (negative) some state of affairs. It might be thought tempting to define affect in

terms of the ability to produce learning, For example, by defining positive affective states (rewards) as those that tend to increase the future likelihood of behaviors that produce or maintain those states and negative affective states (punishments) as those that tend to increase the future likelihood of behaviors that prevent or remove those states (see Chapter 5, Rolls).

However, there is no need to introduce these effects on learning as part of the definition of affective state since those causal connections follow from the more general definitions given above. If predictive associative learning is possible in an organism, that is, if it can discover that some state of affairs *S* tends to produce another state of affairs *S'* that is positively or negatively affective, then actions that tend to produce or to avoid *S* will have the consequence of producing or avoiding a positively or negatively affective state and will therefore themselves tend to be supported or opposed (from the definitions of positive and negative affect). Therefore, if *S'* is positively affective, so will *S* be; and if *S'* is negatively affective, so will *S* be.

States associated with affective states may themselves become associative affective states. Of course, the relationships become far more complex and subtle in more sophisticated organisms with multiple goals, context-sensitive conflict-resolution strategies, explicit as opposed to implicit affective states and belief-like states, derivation processes, and so on.

Complex Affective States

Depression would seem to be a counterexample to our analysis of positive and negative affective states.[4] It is clearly a negative affective state, yet some forms of depression do not prompt action that tends to remove the state, as our analysis of negative affective states requires. Indeed, depression often prompts behaviors which have functional roles that perpetuate the state, the defining characteristic of positive affect. How can depression be accommodated under our account?

The answer lies in viewing depression as a complex affective state. A possible explanation that employs this view follows.

Having an in-built desire to minimize the perceived obstacles to one's action is a plausible feature for autonomous systems. Such a system might be capable of having a negative affective state, *N*, such that it goes into *N* when it perceives that its set of possible actions is being restricted; and when *N* occurs, a mechanism, *E*, is reliably triggered, which generates a variety of attempts to escape from *N* by escaping from the restrictions. Hence, the state *N* has the function of making the system engage in activity that tends to remove or diminish *N*.

Now suppose that there are some situations in which an overall damping of action is adaptive: for instance, hibernation, being in the presence of a dominant conspecific, or having a brutal parent who reacts violently on the slightest provocation. The adaptivity of restricting actions in such situations might result in the evolution of a damping mechanism, *D*, that, when activated, globally reduces the possibilities for action, via internal controls. So, when the system detects a situation in which such damping would be advantageous, this produces state *P* (an example of a mood) where *P* reliably activates *D*, which in turn both activates or enhances the negative affective state *N* and enhances *P*. While those conditions in which damping is advantageous persist, *P* would be a positively affective state: it can be desirable to lie low in a dangerous situation even though it is not desirable to be in a dangerous situation and lying low is not normally desirable (e.g., when hungry). So, there will be a conflict between *P*, whose function is to reduce activity, and *N*, whose function is to increase possibilities for action; but *P* wins in certain circumstances. In some cases, positive feedback mechanisms could make it very difficult to break out of *P*, even after the initiating conditions have been removed and continuation of damping would no longer be advantageous.

The actual nature of depression is probably far more complex; this explanatory sketch is offered only to show that there is no incompatibility, in principle, between complex states like depression and our analysis of affect.

Incidentally, this explanation sketch also shows that what we call positively or negatively affective states need not be consciously experienced as pleasant or unpleasant. In fact, the state itself need not be recognized, even though some of its consequences are.

Crucial to this explanation is the fact that if two affective substates coexist, one positive and one negative (or if there are two positive or two negative affective states that tend to produce conflicting actions), their effects do not in general "sum up" or "cancel out" as if they were coexisting physical forces. It is even possible for one substate to have the specific function of disabling the normal effects of another, for instance, when being paralyzed by fear prevents the normal escape behavior that would reveal one's location, as in freezing in rats. More generally, vector summation is often not suitable either for combining the effects of coexisting affective states or for dealing with conflicts. Instead of summing, it is normally sensible to select one from a set of desirable but incompatible actions since any "summing" could produce disastrous effects, like Buridan's proverbial ass placed halfway between food and drink. More intelligent organisms may invent ways of satisfying two initially incompatible desires, instead of merely selecting one of them.

Varieties of Affective States and Processes

Within the context of a sufficiently rich (e.g. human-like) architecture, we can distinguish a wide range of affective states, depending on factors such as:

- whether they are directed (e.g., craving an apple) or nonspecific (e.g., general unease or depression)
- whether they are long-lasting or short-lived
- how fast they grow or wane in intensity
- what sorts of belief-like, desire-like, and other states they include
- which parts of an architecture trigger them
- which parts of the architecture they can modulate
- whether their operation is detected by processes that monitor them
- whether they in turn can be or are suppressed
- whether they can become dormant and then be reawakened later
- what sorts of external behaviors they produce
- how they affect internal behaviors (e.g., remembering, deciding, dithering, etc.)
- whether they produce second-order affective states (e.g., being ashamed of being angry)
- what sorts of conceptual resource they require

Many of these distinctions, like the distinctions in the taxonomy in Ortony, Clore, and Collins (1988), cannot be applied to organisms or robots with much simpler architectures than an adult human one. For instance it is not clear that the architecture of a newborn human infant can support long-term affective states that are sometimes dormant because attention is diverted, like long-term grief or intense patriotism.

ARCHITECTURAL CONSTRAINTS ON AFFECT

The precise variety of mental states and processes (affective and nonaffective) that are possible for an individual or a species will depend on the information-processing architecture of that individual or species. Insofar as humans at different stages of development, humans with various kinds of pathology, animals of different kinds, and robots all have different sorts of architecture, that will constrain the classes of affective and other kinds of state they support. The fact that different sorts of architecture support different classes of mental state may mean that care is needed in talking about things like desires, emotions, perception, and learning in different types of organisms and machines, e.g., insects, rodents, primates, human infants, human adults, or

robots of various types. The varieties of emotion, desire, or consciousness that can occur in a newborn infant are different from those that are possible in adults. Unfortunately, there is no agreed-upon terminology for discussing varieties of architecture so that we can pose questions about which sorts of mental state and process are possible in which sorts of architecture. We therefore present a schematic framework for describing architectures, named "CogAff" because it was developed in the Cognition and Affect project at the University of Birmingham. This schema defines a high-level ontology for components and connections between components, in a wide range of information-processing architectures, though it does not cover all possibilities.

CogAff: A Schema Allowing Multiple Types of Emotion

The generic CogAff architecture schema sketched in Figures 8.1 and 8.2 covers a wide variety of types of possible (virtual machine) architectures for organisms or robots, which vary in the types of sophistication in their perceptual mechanisms, their motor mechanisms, and their "central" processing mechanisms, as well as in the kinds of connectivity between submechanisms.

For instance, central processes can be purely reactive, in the sense of producing immediate (internal or external) actions without the use of any

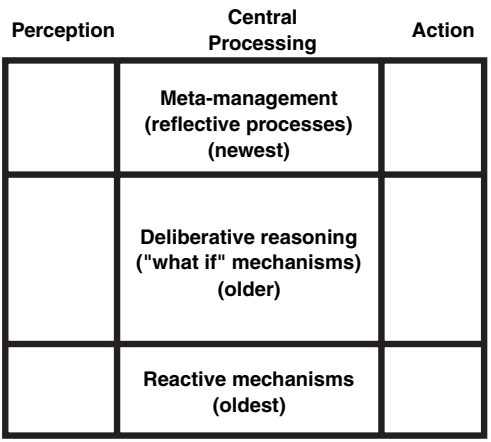


Figure 8.1. The CogAff schema developed in the Cognition and Affect project: two kinds of architectural subdivision are superimposed. One distinguishes perception, central processing, and action. The other (more distinctive) distinguishes three levels: reactive, deliberative, and reflective. Many information flow-paths are possible between the boxes.

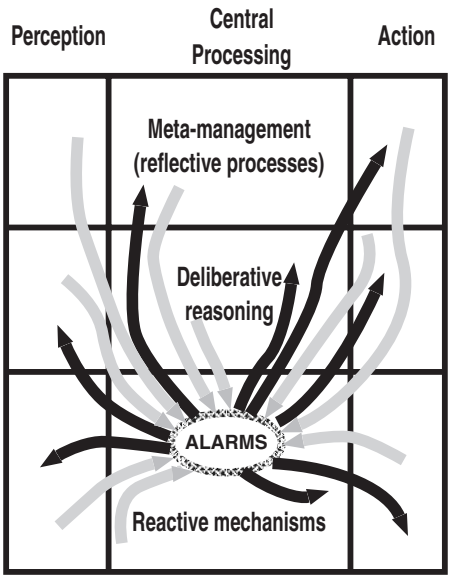


Figure 8.2. This elaborates the CogAff schema of Figure 8.1 to include reactive alarms that detect situations where rapid global redirection of processing is required. There may be many varieties with different input and output connections.

mechanisms for constructing and comparing alternative possible multistep futures. Alternatively, they may be deliberative in the sense of using explicit hypothetical representations of alternative possible futures, predictions, or explanations; comparing them; and selecting a preferred option. This requires highly specialized and biologically costly mechanisms, including short-term memory for temporary structures of varying complexity. Very few animals seem to have these deliberative mechanisms, though simple reactive mechanisms in which two inconsistent reactions are simultaneously activated and then one selected by a competitive mechanism could be described as "proto-deliberative." Another subdivision among central processes concerns meta-management mechanisms, which use architectural features that allow internal processes to be monitored, categorized (using an appropriate ontology for information-processing states and processes), evaluated, and in some cases controlled or modulated. This requires the "meta-semantic" capability to represent and reason about states and processes with semantic content.

These are not mutually exclusive categories since ultimately all processes have to be implemented in reactive mechanisms. Moreover, meta-management processes may be either reactive or deliberative.

Corresponding to the different kinds of processing mechanism and semantic resource available in the central subsystems, we can also distinguish layers of abstraction in the perceptual and action subsystems. For instance, a deliberative layer requires perceptual mechanisms that can discretize, or "chunk," the environment into categories between which associations can be learned that play a role in planning and predicting future events. It is not always appreciated that without such discretization, multistep planning would require consideration of branching continua, which appears to be totally infeasible. Another sort of correspondence concerns the ability of organisms to perceive others as information-users. Doing this requires perceptual processes to use concepts for other agents that are similar to those the meta-management system uses for self-categorization.[5] Examples might be seeing another as happy, sad, attentive, puzzled, undecided, angry, looking to the left, etc. Similarly, layers of abstraction in an action system could evolve to meet the varying needs of central layers.

Superimposing two threefold distinctions gives a grid of nine possible sorts of component for the architecture, providing a crude, high-level classification of submechanisms that may be present or absent. Architectures can vary according to which of these "boxes" are occupied, how they are occupied, and what sorts of connection there are between the occupants of the boxes. Further distinctions can be made as follows:

- whether the components are capable of learning or fixed in their behavior
- whether new components and new linkages develop over time
- which forms of representation and semantic content are used in the various boxes

In Figure 8.2 we indicate the possibility of a reactive component that receives inputs from all the other components and sends outputs to all of them. This could be a design for an "alarm" system that detects situations where rapid global redirection of processing is required, one of the ways of thinking about the so-called "limbic system" (discussed in Chapter 3 by Kelley and in Chapter 4 by Fellous and LeDoux), although there can be many more specialized alarm systems in a complex architecture, such as a protective blinking reflex.

This schema provides a generic framework relative to which particular architectures can be defined by specifying types of components, types of links, types of formalisms, and types of mechanisms used in the various components. This subsumes a very wide variety of types of architectures, and within each type a wide variety of architectures of that type. See also Sloman and Logan, 2000 and Sloman, 2000b.

Many architectures that have been investigated in recent years are purely reactive (Nilsson, 1994). Some purely reactive architectures have layers of

control, where all the layers are merely reactive subsystems monitoring and controlling the layers below them (Brooks, 1991; see also Chapter 10). Some early artificial intelligence (AI) systems had purely deliberative architectures, for instance, planners, theorem provers, and early versions of the SOAR architecture (Laird, Newell, & Rosenbloom, 1987). Some architectures have different sorts of central processing layer but do not have corresponding layers of abstraction in their perception and action subsystems. An information flow diagram for such a system would depict information coming in through low-level perceptual mechanisms, flowing up and then down the central processing tower, and then going out through low-level action mechanisms. This sort of flow diagram is reminiscent of the Greek Ω, so we call these "omega architectures" (e.g. Cooper and Shallice, 2000).

Different Architectures Support Different Ontologies

For each type of architecture, we can analyze the types of state and process that can occur in instances of that type, whether they are organisms or artifacts, and arrive at a taxonomy of types of emotion and other state that the architecture can support. For instance, one class of emotions (primary emotions) might be triggered by input from low-level perceptual mechanisms to an alarm system (shown in Fig. 8. 2), which interrupts normal processing in other parts of the reactive subsystem to deal with emergency situations (we return to this below). What we are describing as "normal" processing in the other parts is simply what those parts would do to meet whatever needs they have detected or to perform whatever functions they normally fulfill.

Another class of emotions (secondary emotions) might be triggered by inputs from internal deliberative processes to an alarm system, for instance if a process of planning or reasoning leads to a prediction of some highly dangerous event or a highly desirable opportunity for which special action is required, like unusual caution or attentiveness. Recognition of such a situation by the alarm mechanism might cause it immediately to send new control signals to many parts of the system, modulating their behavior (e.g., by pumping hormones into the blood supply). It follows that an architecture that is purely reactive could not support secondary emotions thus defined.

However, the CogAff framework does not determine a unique class of concepts describing possible states, although each instance of CogAff does.

A theory-generated ontology of states and processes need not map in a simple way onto the pretheoretical collection of more or less confused concepts (emotion, mood, desire, pleasure, pain, preference, value, ideal, attitude, and so on). However, instead of simply rejecting the pre-theoretical concepts, we use architecture-based concepts to refine and extend them. There

are precedents for this in the history of science: a theory of the architecture of matter refines and extends our pretheoretical classifications of types of matter and types of process; a theory of how evolution works refines and extends our pretheoretical ways of classifying kinds of living things, for example, grouping whales with fish; and a theory of the physical nature of the cosmos changes our pretheoretical classifications of observable things in the sky, even though it keeps some of the distinctions, for example, between planets and stars (Cohen, 1962).

The general CogAff framework should, in principle, be applicable beyond life on earth, to accommodate many alien forms of intelligence, if there are any. However, as it stands, it is designed for agents with a located body, and some aspects will need to be revised for distributed agents or purely virtual or otherwise disembodied agents.

If successful for the purposes of science and philosophy, the architecture schema is also likely to be useful for engineering purposes, though many engineering goals can be achieved using shallow concepts (defined purely behaviorally) and shallow theories (linking conditions to observable behaviors). For instance, this may be all that is required for production of simple but effective "believable" agents for computer entertainments (see also Chapter 10).

Intermediate cases may, as pointed out by Bates (1994), use architectures that are broad in that they encompass many functions but shallow in that the individual components are not realistic. Exploring broad and initially shallow, followed by increasingly deep, implementations may be a good way to understand the general issues. In the later stages of such research, we can expect to discover mappings between the architectural functions and neural mechanisms.

When Are Architectural Layers/Levels/Divisions the Same?

Many people produce layered diagrams that indicate different architectural slices through a complex system. However, close textual analysis reveals that things that look the same can actually be very different. For example, there is much talk of "three-layer" models, but it is clear that not all three-layered systems include the same sorts of layers. The model presented in Chapter 7 (Ortony et al.) has three layers (reactive, routine, and reflective), but none of these maps directly onto the three layers of the CogAff model. For example, their middle layer, the routine layer, combines some aspects of what we assign to the lowest layer, the reactive layer (e.g., learned, automatically executable strategies), and their reflective layer (like the reflective layer in Minsky, 2003) includes mechanisms that we label as part of the deliberative

layer (e.g., observing performance of a plan and repairing defects in the plan), whereas our third layer would contain only the ability to observe and evaluate internal processes, such as the planning process itself, and to improve planning strategies, like Minsky's (2003) "self-reflective" layer. Moreover, what we call "reactive" mechanisms occur in all three layers in the sense that everything ultimately has to be implemented in purely reactive systems.

More importantly, in the model of Ortony et al., the reflective layer receives only preprocessed perceptual input and does not do any perceptual processing itself, whereas CogAff allows for perceptual and action processing in the meta-management layer, for instance, seeing a face as happy or producing behavior that expresses a high-level mental state, such as indecision.

Even when people use the same labels for their layers, they often interpret them differently: for example, some people use "deliberative" to refer to a reactive system which can have two or more simultaneously triggered, competing reactions, one of which wins over the other (e.g., using a "winner takes all" neural mechanism). We call that case "protodeliberative," reserving the label "deliberative" for a system that is able to construct and compare structured descriptions with compositional semantics, where the descriptions do not have a fixed format but can vary according to the task (e.g., planning trees, theories, explanations of an observed event, etc.). Another example is the tendency of some researchers to use "reactive" to imply "stateless." Unfortunately, we do not yet have a good theoretical overview of the space of possible designs comprising both purely reactive and fully deliberative designs. There are probably many interesting intermediate cases that need to be studied if we are to understand both evolution and individual development.

H-CogAff: A Special Case of CogAff

We are currently developing H-CogAff (depicted in Fig. 8.3), a first-draft version of a specific architecture, which is a special case of the CogAff schema, conjectured to cover the main features of the virtual information-processing architecture of normal (adult) humans, though there are still many details to be worked out.

This architecture allows us to define a variety of classes of human emotions, which differ with regard to which component of the architecture triggers them and which components they affect. In addition to primary and secondary emotions, we distinguish tertiary emotions, which perturb or have a disposition to perturb the control of attention in the meta-management subsystem, as explained at length elsewhere (Wright, Sloman, & Beaudoin, 1996/2000). The layers in H-CogAff are also intended to mark significant evolutionary steps. For example, the architecture of H-CogAff assumes that

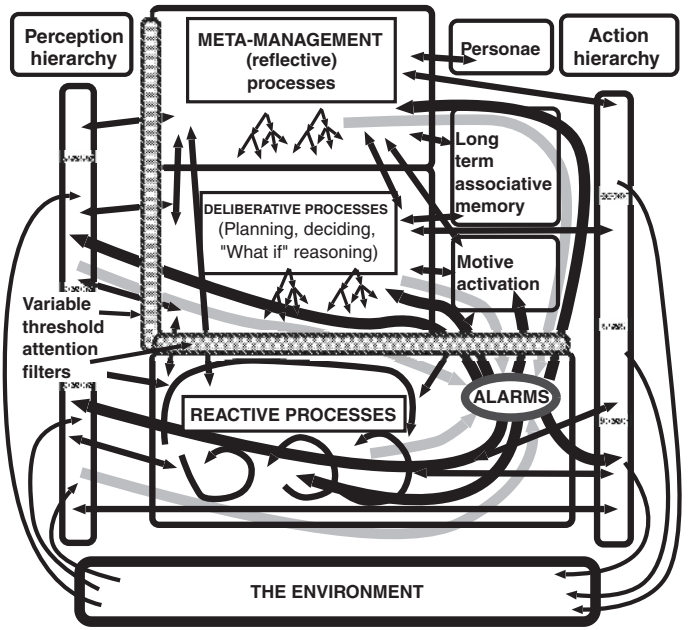


Figure 8.3. The H-CogAff architecture is a version of the CogAff architecture of Figure 8.2, which has many of the features posited for the cognitive architecture of adult humans. Note particularly the representation of personae, the activation of motives, the long-term associative memory, and the attentional filters that modify not only the treatment of sensory data but also the interactions between different levels of sensory processing. Meta-management may be able to inspect intermediate states in perceptual layers, e.g., sensory quality. Indeed, the architecture of H-CogAff assumes that the evolution of the meta-management layer made possible the evolution of additional layers in perceptual and action systems related to the needs and capabilities of the meta-management layer. Not all possible links between boxes are shown.

the evolution of the meta-management layer made possible the evolution of additional layers in perceptual and action systems related to the needs and capabilities of the metamanagement layer (e.g., using the same ontology for labeling internal states and perceived states of others; see Chapter 9 of Sloman, 1978; Sloman, 1989, 2001b; Sloman & Chrisley, 2003).

Architectural Presuppositions

Our above conjectures imply that our folk-psychological concepts and theories all have architectural presuppositions. However, since those presuppositions

are sometimes unclear, inarticulate, confused, or inconsistent, the clarity and consistency of our use of concepts like emotion, attention, learning, and so on will be undermined. So, scientists, engineers, and philosophers who use those concepts to ask questions, state theories, or propose practical goals are likely to be confused or unclear. If we use architecture-based concepts, by defining new, more precise versions of our old mental concepts in terms of the types of processes supported by an underlying architecture, we may hope to avoid arguing at cross purposes, e.g. about which animals have emotions, or how consciousness evolved. (Similar comments may be made about using architecture-based analysis to clarify some technical concepts in psychology, e.g. drive, executive function.)

Where to Begin?

We agree with Turner & Ortony (1992) that the notion of "basic emotion" involves deep muddles. Searching for a small number of basic emotions from which others are composed is a bit like searching for a small number of chemical reactions from which others are composed. It is the wrong place to look. To understand a wide variety of chemical processes, a much better strategy is to look for a collection of basic physical processes in the physical mechanisms that underly the chemical reactions and see how they can be combined. Likewise, with emotions, it is better to look for an underlying collection of processes in information-based control systems (a mixture of virtual and physical machines) that implement a wide variety of emotional (and other affective) states and processes, rather than to try to isolate a subset of emotions to provide the basis of all others, for example, by blending or vector summation (see Chapter 10, Breazeal & Brooks).

The kinds of architectural presupposition on which folk psychology is based are too vague and too shallow to provide explanations for working systems, whether natural or artificial. Nevertheless, folk psychology is a useful starting point as it is very rich and includes many concepts and implicit theories that we use successfully in everyday life. However, as scientists and engineers, we have to go beyond the architectures implicit in folk psychology and add breadth and depth.

Since we do not know enough yet to get our theories right the first time, we must be prepared to explore alternative architectures. In any case, there are many types of organism with many similarities and differences in their architectures. Different artificial systems will also need different architectures. So, there are many reasons for not attending exclusively to any one kind of architecture. Many different conjectured architectures can be inspired by empirical evidence regarding biological systems, including humans at dif-

ferent stages of development. Moreover, humans have many subsystems that evolved long ago and still exist in other animals, where they are sometimes easier to study. We should also be open to the possibility of biological discoveries of architectures that do not fit our schema, for which the schema will have to be extended. Moreover, we are not restricted to what is biologically plausible. We can also consider architectures for future possible robots.

EXAMPLES OF ARCHITECTURE-BASED CONCEPTS

We are extending folk-psychological architectures in the framework of the CogAff schema (Fig. 8.1), which supports a wide variety of architectures. An example is our tentatively proposed special case, the H-CogAff architecture offered as a first draft theory of the human virtual information processing architecture. In the more specific context of H-CogAff, we can distinguish more varieties of emotions than are normally distinguished (and more varieties of perceiving, learning, deciding, attending, acting). However, it is likely that the ontology for mental states and processes that will emerge from more advanced versions of H-CogAff (or its successors) will be far more complex than anyone now imagines.

We shall offer some examples of words normally regarded as referring to emotions and show how to analyze them in the context of an architecture. We start with a proposal for a generic definition of emotion that might cover many states that are of interest to psychologists who are trying to understand emotions in human as well as to roboticists intending to study the utility of emotional control in artifacts. This is an elaboration of ideas originally in Simon (1967/1979).

Toward a Generic Definition of "*Emotion*"

We start from the assumption that in any information-processing system there are temporally extended processes that sometimes require more time to complete a task than is available because of the speed with which external events occur. For example, the task of working out how to get some food that is out of reach may not be finished by the time a large, fast-approaching object is detected, requiring evasive action. An operating system might be trying to write data to a memory device, but the user starts disconnecting the device before the transfer is complete. It may be useful to have a process which detects such cases and interrupts normal functioning, producing a very rapid default response, taking high priority over everything else, to avoid file corruption. In Figure 8.2, we used the label "alarm mechanism"

for such a fast-acting system which avoids some danger or grasps some short-lived opportunity.

In an animal or robot, such an alarm mechanism will have to use very fast pattern-triggered actions using relatively unsophisticated reasoning. It is therefore likely sometimes to produce a less appropriate response than the mechanism which it interrupts and overrides would have produced if it had had sufficient time to complete its processing. However, the frequency of wrong responses might be reduced by training in a wide variety of circumstances. This notion can also be generalized to cases where, instead of interrupting, the alarm mechanism merely modulates the normal process (e.g., by slowing it down or turning on some extra resources which are normally not needed, such as mechanisms for paying attention to details).

We can use the idea of an alarm system to attempt a very general definition of *emotion*: an organism is in an *emotional state* if it is in an episodic or dispositional state in which a part of it, the biological function of which is to detect and respond to abnormal states, has detected something which is either

1. actually (episodic) interrupting, preventing, disturbing, or modulating one or more processes which were initiated or would have been initiated independently of this detection, or
2. disposed (under certain conditions) to interrupt, prevent, disturb, etc. such processes but currently suppressed by a filter (Fig. 8.3) or priority mechanism.

We have given examples involving a speed requirement, but other examples may involve detection of some risk or opportunity that requires an ongoing action to be altered but not necessarily at high speed, for instance, noticing that you are going to be near a potentially harmful object if you do not revise your course.

This architecture-based notion of "emotion" (involving actual or potential disruption or modulation of normal processing) falls under the very general notion of "affective" (desire-like) state or process proposed above. It encompasses a large class of states that might be of interest to psychologists and engineers alike. In the limiting cases, it could even apply to relatively simple organisms such as insects, like the fly whose feeding is aborted by detection of the fly-swatter moving rapidly toward it or the woodlouse that quickly rolls up into a ball if touched by a pencil. For even simpler organisms (e.g. a single-celled organism), it is not clear whether the information-processing architecture is rich enough to support the required notions.

This generic notion of emotion as "actual or potential disturbance of normal processing" can be subdivided into many different cases, depending on the architecture involved and where in the architecture the process is

initiated, what it disturbs, and how it does so. There is no implication that the disturbance will be externally visible or measurable, though often it will be if the processes that are modified include external actions.

Previous work (Sloman, 2001a) elaborated this idea by defining *primary emotions* as those entirely triggered within a reactive mechanism, *secondary emotions* as those triggered within a deliberative system, and *tertiary emotions* (referred to as "perturbances" in the analysis of grief by Wright, Sloman, & Beaudoin, 1996/2000) as states and processes that involve actual or dispositional disruption of attention-control processes in the meta-management (reflective) system. That is just a very crude, inadequate, first-draft high-level subdivision which does not capture the rich variety of processes colloquially described as "emotions" or "emotional."

Within the framework of an architecture as rich as H-CogAff, many more subdivisions are possible, including subdivisions concerning different time scales, different numbers of interacting subprocesses, different etiologies, different sorts of semantic content, etc. This overlaps with the taxonomy in Ortony, Clore, and Collins (1988).

An Architecture-Based Analysis of "Being Afraid"

Many specific emotion concepts (e.g., fear, joy, disgust, jealousy, infatuation, grief, obsessive ambition, etc.) share some of the polymorphism and indeterminacy of the general concept. For example, "fear" and "afraid" cover many types of state and process. Consider being

1. afraid of spiders
2. afraid of large vehicles
3. afraid of a large vehicle careering toward you
4. afraid of a thug asking you to hand over your wallet
5. afraid your favorite party is going to lose the next election
6. afraid you have some horrible disease
7. afraid of growing old
8. afraid that your recently published proof of Goldbach's conjecture has some hidden flaw

Each of these different forms of "being afraid" requires a minimal set of architectural features (i.e., components and links among them). For example, there are instances of the first four forms that involve perceptions that directly cause the instantiation of the state of being afraid, while the other four do not depend on perception to cause their instantiation (e.g., merely remembering that your proof has been published might be sufficient to cause fear that the proof has a hidden flaw). There are states that inherently come

from mental processes other than current perception (e.g., embarrassment about what you said yesterday).

Furthermore, the above states vary in cognitive sophistication. The first, for example, might only require a reactive perceptual process that involves a matcher comparing current perceptions to innate patterns (i.e., those of spiders), which in turn triggers an alarm mechanism. The alarm mechanism could then cause various visceral processes (e.g., release of hormones, widening of the pupils) in addition to modifications of action tendencies and dispositions (e.g., the disposition to run away or to scream; cf. LeDoux, 1996, and Fellous & LeDoux's Chapter 4).

The second, for example, could be similar to the first in that large objects cause anxiety, or it could be learned because fast-approaching vehicles in the past have caused state 3 to be instantiated, which in turn formed an association between it and large vehicles so that the presence of large vehicles alone can instantiate state 3. State 2 then involves a permanent dispositional state by virtue of the learned associative connection between large vehicles and state 3. State 2 is activated upon perceiving a large vehicle, regardless of whether it is approaching or not.

The fourth involves even more in that it requires projections concerning the future and is instantiated because of possible negative outcomes. Consequently, a system that can instantiate state 4 will have to be able to construe and represent possible future states and maybe assess their likelihood. Note, however, that simple forms of state 4 might be possible in a system that has learned a temporal association only (namely, that a particular situation, e.g., that of a thug asking for one's wallet, is always preceded by encountering a thug). In that case, a simple conditioning mechanism might be sufficient.

For the remaining examples, however, conditioning is not sufficient. Rather, reasoning processes of varying complexity are required that combine various kinds of information. In state 6, this may be evidence from one's medical history, statements of doctors, common-sense knowledge, etc. The information needs to be corroborated in some way (whether the corroboration is valid or not does not matter) to cause the instantiation of these states. For the last three, it is likely that additional reflective processes are involved, which are capable of representing the very system that instantiates them in different possible contexts and evaluating future outcomes with respect to these contexts and the role of the system in them (e.g., a context in which the disease has manifested itself and how friends would react to it or how colleagues would perceive one's failure to get the proof right).

The above paragraphs are, of course, only very sketchy outlines that hint at the kind of functional analysis we have in mind, which eventually leads to a list of functional components that are required for an affective state of a

particular kind to be instantiable in an architecture. Once these requirements are fixed, it is possible to define the state in terms of these requirements and to ask whether a particular architecture is capable of instantiating the state. For example, if reflective processes that observe, monitor, inspect, and modify deliberative processes are part of the last three states, then architectures without a meta-management layer (as defined in CogAff) will not be capable of instantiating any of them.

This kind of analysis is obviously not restricted to the above states but could be done for any form of anger (Sloman, 1982), fear, grief (Wright, Sloman, & Beaudoin, 1996/2000), pride, jealousy, excited anticipation, infatuation, relief, various kinds of joy, schadenfreude, spite, shame, embarrassment, guilt, regret, delight, or enjoyment (of a state or activity). Architecture-based analyses are also possible for nonemotional, affective states such as attitudes, moods, surprise, expectation, and the like.

DISCUSSION

Our approach to the study of emotions in terms of properties of agent architectures can safely be ignored by engineers whose sole object is to produce "believable" mechanical toys or displays that present appearances that trigger, in humans, the attribution of emotional and other mental states. Such "emotional models" are based on shallow concepts that are exclusively defined in terms of observable behaviors and measurable states of the system. This is in contrast to deep concepts, which are based on theoretical entities (e.g., mechanisms, information structures, types of information, architectures, etc.) postulated to generate those behaviors and states but not necessarily directly observable or measurable (as most of the theoretical entities of physics and chemistry are not directly observable).

Implementing shallow models does not take much if, for example, the criteria for success depend only on human ratings of the "emotionality" of the system, for we, as human observers, are predisposed to confer mental states even upon very simple systems (as long as they obey basic rules of behavior, e.g., Disney cartoons). At the same time, shallow models do not advance our theoretical understanding of the functional roles of emotions in agent architectures as they are effectively silent about processes internal to an agent. Shallow definitions of emotions would make it impossible for someone whose face has been destroyed by fire or whose limbs have been paralyzed to have various emotional states that are defined in terms of facial expressions and bodily movements. In contrast, architecture-based notions would allow people (or robots) to have joy, fear, anguish, despair, and relief despite lacking any normal way of expressing them.

The majority view in this volume seems to be that we need explanatory theories that include theoretical entities whose properties may not be directly detectable, at least using the methods of the physical sciences or the measurements familiar to psychologists (including button-pushing events, timings, questionnaire results, etc.). This is consistent with the generic definition of *emotion* proposed in this chapter, based on internal processes that are capable of modulating other processes (i.e., initiating or interrupting them, changing parameters that give rise to dispositional changes, etc.). Such a definition should be useful both for psychologists interested in the study of human emotions and for engineers implementing deep emotional control systems for robots or virtual agents. While the definition was not intended to cover all aspects of the ordinary use of the word *emotion* (nor could it cover them all given that "emotion" is a cluster concept), it can be used as a guideline that determines the minimal set of architectural features necessary to implement emotions (as defined in this paper). Furthermore, it allows us to determine whether a given architecture is capable of implementing such emotions and, if so, of what kinds (as different emotion terms are defined using architectural features). This is different from much research in AI, where it is merely taken as obvious that a system of a certain sort is indeed emotional.

More importantly, our definition also suggests possible roles of mechanisms that generate what are described as "emotions" in agent architectures (e.g., as interrupt controllers, process modifiers, action initiators or suppressors, etc.) and, hence, when and where it is appropriate and useful to employ such control systems. This is crucial for a general understanding of the utility of what is often referred to as "emotional control" and consequently the adaptive advantage of the underlying mechanisms in biological systems, even though many of the emotions they produce may be dysfunctional.

Do Robots Need Emotions and Why?

One of the questions some robot designers address is whether there is any principled reason why their robots need emotions to perform a given task (assuming some clear definition of *emotion*). However, there is a more general question: whether there is any task that cannot be performed by a system that is not capable of having emotional states.

The answer to this question is certainly nontrivial in the general case. For simple control systems satisfying a particular definition of *emotional*, it may be possible to define a finite-state machine that has exactly the same input–output behavior but does not instantiate any emotion in the specified sense. Most so-called emotional agents currently developed in AI would probably fall under this category.

While this idea applies in principle to agents of all levels of complexity, in practice there are limits to the approach, and the situation will already be very different for more complex agents. For one, implementing the control system as a finite-state controller will not work as the number of states of a complex agent (e.g., with thousands of condition–action rules involving complex representations) will likely be too large for the state table to fit into a standard computer. Hence, the control system needs to be implemented in a virtual machine that supports multiple finite-state machines with substates and connections among them. In short, a complex architecture with complex states will have to be implemented in a virtual machine that supports the required complexity. While transitions are immediate in finite-state machines, many physical steps may be required for a complex virtual machine transition (like a computer updating a simulated neural net). Finite-state machines do not need alarm systems to interrupt normal processing in order to react to unforeseen events: they simply transit into a state where they deal with the circumstance. Complex systems with multiple finite-state machines with complex substates, however, need a way of coordinating state transitions (especially if they have different lengths, might take different amounts of time, or might even occur asynchronously). In that case, special mechanisms need to be added to improve the reactivity of the system (i.e., the time it takes to respond to critical environmental changes).

Following this reasoning, one would expect to find something like alarm mechanisms in complex agents that need to react quickly in real time to unforeseen events. Such systems might lead to internal interactions instantiating emotional states as defined above which the designers did not intend (e.g., an operating system with a mechanism that terminates processes or limits and reallocates resources in response to an overload might delete processes urgently required for some subtask).

Returning to the question of whether robots need or should have emotions, the answer will depend on the task and environment for which the robot is intended. This *niche*, or set of requirements to be satisfied, will in turn determine a range of architectures able to satisfy the requirements. The architectures will then determine the sorts of emotions that are possible (or desirable) for the robot. Here are some examples of questions designers may ask:

- Will the robot be purely for entertainment?
- Will it have a routine practical task, for example, on a factory floor or in the home (cleaning carpets)?
- Will it have to undertake dangerous tasks in a dynamic and unpredictable environment (as in the Robocup Rescue project)?
- Will it have to cooperate with other agents (robots and humans/animals)?

- Will it be a long-term friend or helper for one or more humans (e.g., robots to help the disabled or infirm)?
- Will its tasks include understanding the humans with whom it interacts?
- Will it need to fit into different cultures or subcultures with different tastes, preferences, values, etc.?
- Will the designers be able to anticipate all the kinds of problem and conflict that can arise during the "life" of the robot?
- Will it ever need to resolve ethical conflicts on its own, or will it always refer such problems to humans? (Maybe there will not be time or communication links if it is down a mine or in a spacecraft on a distant planet.)
- Will it need to be able to provide explanations and justifications for its goals, preferences, decisions, etc.?
- Is the design process aimed primarily at scientific goals, such as trying to understand how human (and other animal) minds work, or are the objectives practical, like how to get some task done? (We are mainly interested in the science, whereas some people are primarily interested in practical goals.)

To say that certain mechanisms, forms of representation, or architectural organization are required for an animal or robot is to say something about the niche of that animal or robot and what types of information-processing capabilities, and behaviors, increase the individual's chance of doing well (surviving, flourishing, reproducing successfully, achieving individual goals, etc.) in that niche. A full treatment will require a survey of niche-space and design-space and the relationships between them (see Breazeal & Brooks, Chapter 10, for an attempt at classifying them). (This is also required for understanding evolutionary and developmental trajectories.)

How Are Emotions Implemented?

Another important, recurring question raised in the literature on emotions (in AI) is whether a realistic architecture needs to include some particular, dedicated emotion mechanism. Our view (as argued elsewhere: Sloman & Croucher, 1981; Sloman, 2001a) is that, in realistic human-like robots, emotions of various types will emerge, as they do in humans, from various types of interaction between many mechanisms serving different purposes, not from a dedicated emotion mechanism.

Another issue is whether emotions are necessarily tied to visceral processes, as assumed in biological theories that construe notions like emotion, affect, and mood as characterizing physical entities (animal bodies, including brains,

muscles, skin, circulatory system, hormonal systems, etc.). If the presence of an emotion requires a body of a particular type (e.g., with chemical hormones), then there may never be (nonbiological) robots with emotions.

Alternatively, one could take emotion terms to refer to states and processes in virtual machines that happen to be implemented in these particular physical mechanisms but might in principle be implemented in different mechanisms. In that case, nonbiological artifacts may be capable of implementing emotions as long as they are capable of implementing all relevant causal relationships that are part of the definition of the emotion term. The above alternatives are not mutually exclusive, for there is nothing to rule out the combination of

- deep, implementation-neutral, architecture-based concepts of emotion, definable in terms of virtual machine architectures without reference to implementation-dependent properties of the physical substratum
- special cases (i.e., subconcepts) that are implementation-dependent and defined in terms of specific types of body and how they express their states (e.g., snarling, weeping, grimacing, tensing, changing color, jumping up and down, etc.).

LeDoux (1996) and Panksepp (1998) present such special cases, where emotions are defined in terms of particular brain regions and pathways. These definitions are intrinsically dependent on a particular bodily make-up (i.e., anatomical, physiological, chemical, etc.). Hence, systems that do not possess the required type of body cannot, by definition, implement them.

The conceptual framework of Ortony, Clore, and Collins (1988; and see Chapter 7), however, is an example of an implementation-neutral conception, where emotions are defined in terms of an ontology that distinguishes events, objects, and agents and their different relationships to the system that has the emotion. It is interesting to note that if emotions are reactions to events, agents, or objects (as Ortony and co-workers claim), then their agent-based emotions (i.e., emotions elicited by agents) cannot occur in architectures that do not support representations of the ontological distinction between objects and agents. Such systems could consequently never be jealous (as being jealous involves other agents). This is a virtual machine design constraint, not an implementation constraint.

Comparison with Other Work

There is now so much work on emotions in so many disciplines that a comparison with alternative theories would require a whole book. Readers of this volume will be able to decide which of the other authors have explicitly

or implicitly adopted definitions of *emotion* that take account of the underlying architecture and the processes that the architecture can support, which have assumed that there is a clear and unambiguous notion of "emotion" and which have not, which are primarily interested in solving an engineering design problem (e.g., producing artifacts that are entertaining or demonstrate how humans react to certain perceived behaviors), and which are attempting to model or explain naturally occurring states and processes. One thing that is relatively unusual that we have attempted is to produce a generic framework to accommodate a wide variety of types of organism and machine. We hope that more researchers will accept that challenge, and the challenge of developing a useful ontology for describing and comparing different architectures so that work in this area can grow into a mature science instead of a large collection of ad hoc and loosely related studies that are hard to compare and contrast.

The view we have propounded contradicts some well-known theories of emotions, in particular Jamesian theories (James, 1890; Damasio, 1994), according to which having an emotion involves sensing some pattern in one's physiological state. The claim that many emotions involve changes to physiological states (e.g., blood pressure, muscular tension, hormones in the bloodstream) is perfectly consistent with what we have said about emotions, but not the claim that such processes are *necessary* conditions for emotions. Theories proposing such necessary conditions have a hard problem accommodating long-term emotional states that are often temporarily suppressed by other states and processes, for instance, long-term grief, long-term concern about a threat to one's job, or intense long-term devotion to a political project.

However, others do present architectural ideas partly similar to our own, though arrived at from a completely different standpoint (see Barkley, 1997, for an example from neuropsychiatry). Our emphasis on the link between the concept of emotion and mechanisms that produce strong dispositions to disrupt and redirect other processing also fits much folk psychology and features of emotions that make them the subject of novels. Changes in blood pressure, galvanic skin responses, and levels of hormones are not usually of much interest to readers of great literature compared to changes in thought processes, preferences, evaluations, how much people can control their desires, the extent to which their attention is strongly held by someone or something, and the consequences thereof. These are features of what we have called "tertiary" emotions, which usually involve rich semantic content as well as strong control states. When a robot first tells you in detail why it is upset by your critical analysis of the poems it has written, you will be far more likely to believe it has emotions than if it merely blushes, weeps, and shakes its head. Even ducking to avoid being hit by a large moving object might just be a simple planned

response to a perceived threat, in a robot whose processing speeds are so great that it needs no alarm mechanism. It is arguable, then, that only linguistic expression (see Arbib, Chapter 12) is capable of conveying the vast majority of tertiary emotions, whereas most current research on detecting emotions focuses on such "peripheral" phenomena as facial expression, posture, and other easily measurable physiological states.

THE NEXT STEPS

Emotions in the sense we have defined them are present in many control systems, where parts of the control mechanism can detect abnormal states and react to them (causing a change in the normal processing of the control system, either directly through interruption of the current processing or dispositionally through modification of processing parameters). Emotions thus defined are not intrinsically connected to living creatures, nor are they dependent on biological mechanisms; e.g., operating systems running on standard computers have several emotions in our technical sense, although they lack many of the detailed features of the sorts of emotion to which our folk concepts are applied.

What *is* special about at least a subset of emotions so defined (compared to other non-emotional control states) is that they (1) form a class of useful control states that (2) are likely to evolve in certain resource-constrained environments and, hence, (3) may also prove useful for certain AI applications (e.g., robots that have only limited processing resources, which impose severe constraints on the kinds of control mechanism that can be implemented on them).

Useful affective control mechanisms are likely to evolve if there are many evolutionary trajectories that, given various sets of well-specified initial conditions and fitness functions, will lead to those control systems (e.g., Scheutz, 2001; Scheutz & Schermerhorn, 2002). A subset of those will be control mechanisms that can produce emotional states suited to coping with emergencies or unexpected situations as they occur in dynamic, unpredictable, real-world environments.

It is not yet clear which of the more subtle and complex long-term emotional states, such as grief, ambition, jealousy, infatuation, and obsession with a difficult problem, are merely side effects of desirable mechanisms and which are states that can be intrinsically useful in relation to either the needs of individuals or the needs of a social group or species. Human aberrations make it clear, however, that machines containing useful mechanisms are capable of getting into highly dysfunctional states through the interactions of those mechanisms. As machines become more human-like, we can expect

some undesirable emotional states to be hard to avoid in certain contexts if the machines have affective control mechanism that interact in complex ways.

Detailed studies of design and niche space, in which the relationships between classes of designs and classes of niches for these designs in a variety of environments are investigated, should clarify the costs and benefits. For this, we need experiments with agent architectures that complement theoretical, functional analyses of control systems by systematic studies of performance–cost tradeoffs, which will reveal utility or disadvantages of various forms of control in various environments.

Finally, the main utility in AI of control systems producing states conforming to our suggested definition of *emotional* does not lie in systems that need to interact with humans or animals (e.g., by recognizing emotions in others and displaying emotions to others). There is no reason to believe that such control mechanisms (where something can modulate or override the normal behavior of something else) are necessary to achieve "believable interactions" among artifacts and humans. Large sets of condition–action rules, for example, may produce convincing behavioral expressions that give the appearance of sympathy or surprise without implementing the kinds of control mechanism that we called "emotional." Hence, such systems may appear to be emotional without actually having emotions in our sense, but appearances will suffice for many applications, especially in computer games and entertainments, as they do in human stage performances and in cartoon films.

In contrast, control mechanisms capable of producing states conforming to our proposed definition of *emotional* will be useful in systems that need to cope with dynamically changing, partly unpredictable and unobservable situations where prior knowledge is insufficient to cover all possible outcomes. Specifically, noisy and/or faulty sensors, inexact effectors, and insufficient time to carry out reasoning processes are all limiting factors with which real-world, real-time systems have to deal. As argued in Simon (1967/1979) and Sloman & Croucher (1981), architectures for such systems will require mechanisms able to deal with unexpected situations. In part, this trivializes the claim that emotional controls are useful since they turn out to be instances of very general requirements that are obvious to engineers who have to design robust and "failsafe" systems to operate in complex environments. What is nontrivial is which systems are useful in different sorts of architectures and why.

There is much work in computer science and robotics that deals with control systems that have some features in common with what we call affective mechanisms, from real-time operating systems that use timers and alarm mechanisms to achieve time-critical tasks to robot control systems that

drive an autonomous unmanned vehicle and need to react to and correct different kinds of error at different levels of processing (Albus, 2000).

As our field matures, it should be possible to explicate this practical wisdom developed in the engineering sciences and compare it to findings in psychology and neuroscience about the control architectures of biological creatures. For this, we need a conceptual framework in which we can express control concepts useful in the description of neural circuits, in the description of higher-level mental processes, and in control theory and related fields. Such a conceptual framework will allow us to see the commonalities and differences in various kinds of affective and nonaffective control mechanism found in biological systems or designed into machines. Systematic studies of architectural tradeoffs will help us understand the kinds of situation where emotional control states should be employed because they will be beneficial, situations where they should be avoided because they are harmful, and situations where they arise unavoidably out of interactions between mechanisms that are useful for other reasons.

Notes

This work was funded by grant F/94/BW from the Leverhulme Trust, for research on "evolvable virtual information processing architectures for human-like minds." The ideas presented here were inspired especially by the work of Herbert Simon and developed with the help of Luc Beaudoin, Ian Wright, Brian Logan, Marvin Minsky, Ruth Kavanagh, and many students, colleagues, and friends. We are grateful for the comments and suggestions from the editors and for their patience (i.e., lack of emotion).

1. The attribute "virtual" here is in contrast to "physical;" i.e., a running virtual machine is an abstract machine containing abstract components that may be capable of running on different physical machines. Virtual machine states can have causal powers, for instance, the power to deliver e-mail or to detect and prevent access violations.

2. There are many variants of this point in the emotions literature. Give a search engine: "emotion + natural kind." Oatley and Jenkins (1996) comment on the diversity of definitions of *emotion* in the psychology literature.

3. Extending terminology used by Damasio (1994), Goleman (1996), and Picard (1997).

4. Thanks to Brian Logan for drawing this to our attention.

5. An interesting research question is whether self-descriptive mechanisms or descriptions of others as information-users evolved first or whether they evolved partly concurrently (Sloman & Logan, 2000). The ability to describe something as perceiving, reasoning, attending, wanting, choosing, etc. seems to require representational capabilities that are neutral between self-description and other-description (see Jeannerod, Chapter 6, for more on assessing the mental states of others).

References

Albus, J. S., Juberts, M., & Szabo, S. (1992, June). RCS: A reference model architecture for intelligent vehicle and highway systems. In *Proceedings of the 25th Silver Jubilee International Symposium on Automotive Technology and Automation*, Florence, Italy.

Austin, J. (1956). A plea for excuses. In J. O. Urmson, & G. J. Warnock (Eds.), *Philosophical papers* (pp. 175–204). Oxford: Oxford University Press.

Barkley, R. A. (1997). *ADHD and the nature of self-control*. New York: Guildford.

Bates, J. (1994). The role of emotion in believable agents. *Communications of the Association for Computing Machinery*, *37*, 122–125.

Beaudoin, L. (1994). Goal processing in autonomous agents. Birmingham, UK: University of Birmingham. Dissertation. (Available at http://www.cs.bham.ac.uk/research/cogaff/).

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*, 139–159.

Cohen, L. (1962). *The diversity of meaning*. London: Methuen.

Cooper, R., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*, 297–338.

Damasio, A. (1994). *Descartes' error: Emotion reason and the human brain*. New York: Putnam.

Delancey, C. (2002). *Passionate engines: What emotions reveal about the mind and artificial intelligence*. Oxford: Oxford University Press.

Dennett, D. C. (1978). *Brainstorms: Philosophical essays on mind and psychology*. Cambridge, MA: MIT Press.

Goleman, D. (1996). *Emotional intelligence: Why it can matter more than IQ*. London: Bloomsbury.

Hume, D. (1978). *A treatise of human nature* (2nd ed.). New York: Oxford University Press. (Original work published 1739)

James, W. (1890). *The principles of psychology*. New York: Holt.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, *33*, 1–64.

Lakatos, I. (1970). *Criticism and the growth of knowledge*. New York: Cambridge University Press.

LeDoux, J. (1996). *The emotional brain*. New York: Simon & Schuster.

Lodge, D. (2002). *Consciousness and the novel: Connected essays*. London: Secker & Warburg.

Millikan, R. (1984). *Language, thought, and other biological categories*. Cambridge, MA: MIT Press.

Minsky, M. L. (2003). *The emotion machine*. Draft available online (http://web.media.mit.edu/~minsky/)

Newell, A. (1990). *Unified theories of cognition*. Boston: Harvard University Press.

Nilsson, N. (1994). Teleo-reactive programs for agent control. *Journal of Artificial Intelligence Research*, *1*, 139–158.

Oatley, K., & Jenkins, J. (1996). *Understanding emotions*. Oxford: Blackwell.

Ortony, A. (2002). On making believable emotional agents believable. In R. Trappl, P. Petta, P., & S. Payr (Eds.), *Emotions in humans and artifacts* (pp. 189–211). Cambridge, MA: MIT Press.

Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of the emotions*. New York: Cambridge University Press.

Panksepp, J. (1998). *Affective neuroscience. The foundations of human and animal emotions*. Oxford: Oxford University Press.

Picard, R. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Sartre, J.-P. (1939). *The emotions: A sketch of a theory*. New York: Macmillan.

Scheutz, M. (2001). The evolution of simple affective states in multi-agent environments. In D. Cañamero (Ed.), *Proceedings of the American Association for Artificial Intelligence fall symposium 01* (pp. 123–128). Falmouth, MA: AAAI Press.

Scheutz, M., & Schermerhorn, P. (2002). Steps towards a systematic investigation of possible evolutionary trajectories from reactive to deliberative control systems. In R. Standish (Ed.), *Proceedings of the 8th Conference of Artificial Life*. Cambridge, MA: MIT Press.

Scheutz, M., &Sloman, A. (2001). Affect and agent control: Experiments with simple affective states. In N. Zhong, et al. (Eds.), *Intelligent agent technology: Research and development* (pp. 200–209). River Edge, NJ: World Scientific.

Simon, H. A. (1979). Motivational and emotional controls of cognition. In *Models of thought* (pp. 29–38). New Haven, CT: Yale University Press. (Original work published 1967)

Sloman, A. (1978). *The computer revolution in philosophy*. Hassocks, UK: Harvester Press (Humanities Press). Available online (http://www.cs.bham.ac.uk/research/cogaff/crp)

Sloman, A. (1982). Towards a grammar of emotions. *New Universities Quarterly*, *36*, 230–238. (http://www.cs.bham.ac.uk/research/cogaff/0-INDEX96-99.html#47)

Sloman, A. (1989). On designing a visual system (towards a gibsonian computational model of vision). *Journal of Experimental and Theoretical AI*, *1*, 289–337.

Sloman, A. (1993). The mind as a control system. In C. Hookway & D. Peterson (Eds.), *Philosophy and the cognitive sciences* (pp. 69–110). Cambridge: Cambridge University Press.

Sloman, A. (1996). Towards a general theory of representations. In D. M. Peterson (Ed.), *Forms of representation: An interdisciplinary theme for cognitive science* (pp. 118–140). Exeter, UK: Intellect.

Sloman, A. (2000a). Interacting trajectories in design space and niche space: A philosopher speculates about evolution. In M. Schoenauer et al. (Eds.), *Parallel problem solving from nature—PPSN VI, lecture notes in computer science* (No. 1917, pp. 3–16). Berlin: Springer-Verlag.

Sloman, A. (2000b). Models of models of mind. In M. Lee (Ed.), *Proceedings of symposium on how to design a functioning mind, AISB'00* (pp. 1–9). Birmingham: Artificial Intelligence and Simulation of Behaviour.

Sloman, A. (2001a). Beyond shallow models of emotion. *Cognitive Processing: International Quarterly of Cognitive Science*, *2*, 177–198.

Sloman, A. (2001b). Evolvable biologically plausible visual architectures. In T. Cootes & C. Taylor (Eds.), *Proceedings of British machine vision conference* (pp. 313–322). Manchester: British Machine Vision Association.

Sloman, A. (2002). Architecture-based conceptions of mind. In *In the Scope of Logic, Methodology, and Philosophy of Science* (Vol. II, pp. 403–427; Synthese Library Vol. 316). Dordrecht: Kluwer.

Sloman, A., & Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies, 10*, 113–172.

Sloman, A., & Croucher, M. (1981). Why robots will have emotions. In *Proceedings of the 7th International Joint Conference on AI* (pp. 197–202), Vancouver: Morgan-Kaufman.

Sloman, A., & Logan, B. (2000). Evolvable architectures for human-like minds. In G. Hatano, N. Okada, & H. Tanabe (Eds.), *Affective minds* (pp. 169–181). Amsterdam: Elsevier.

Turner, T., & Ortony, A. (1992). Basic emotions: Can conflicting criteria converge? *Psychological Review, 99*, 566–571.

Wright, I., Sloman, A., & Beaudoin, L. (2000). Towards a design-based analysis of emotional episodes. *Philosophy Psychiatry and Psychology, 3*(2): 101–126. Reprinted in R. L. Chrisley (Ed.), *Artificial intelligence: Critical concepts in cognitive science* (Vol. IV). London: Routledge. (Original work published 1996)

9 Moving Up the Food Chain

Motivation and Emotion in Behavior-Based Robots

RONALD C. ARKIN

This article investigates the relationship between motivations and emotions as evidenced by a broad range of animal models, including humans. Emotions constitute a subset of motivations that provide support for an agent's survival in a complex world. Both motivations and emotions affect behavioral performance, but motivation can additionally lead to the formulation of concrete goal-achieving behavior, whereas emotions are concerned with modulating existing behaviors in support of current activity. My focus is placed on how these models can have utility within the context of working robotic systems. Behavior-based control serves as the primary vehicle through which emotions and motivations are integrated into robots ranging from hexapods to wheeled robots to humanoids. In this framework, motivations and emotions dynamically affect the underlying control of a cybernetic system by altering its underlying behavioral parameters.

I review actual robotic examples that have, each in their own way, provided useful environments where questions about emotions and motivations can be addressed. I start with a description of models of the sowbug that provided the first testbed for asking questions about the use of parallel streams of sensory information, goal-oriented behaviors, motivation and emotions, and developmental growth. I then move on in some detail to a model of the praying mantis, in which explicit motivational

state variables such as fear, hunger, and sex affect the selection of motivated behaviors. Moving on to more-complex systems, I review the progress made in using attachment theory as a basis for robot exploration. I then describe the attempts at using canine ethology to design doglike robots that use their emotional and motivational states to bond with their human counterparts. Finally, I describe an ongoing modeling effort to address the issue of time varying affect-related phenomena such as personality traits, attitudes, moods, and emotions.

It has been a while since I have had to wrestle with the nebulous, relatively unscientific term *emotions*. Previously (Arkin, 1998), I stated that "Modifying Associate U.S. Supreme Court Justice John Paul Stevens' famous quotation, we can't define emotion, but we know it when we see it." Granted, significant advances have been recently made in understanding the neural underpinnings of emotional structure in humans (Dolan, 2002), where "emotions represent complex psychological and physiological states that, to a greater or lesser degree, index occurrences of value" (where *value* refers to "an organism's facility to sense whether events in its environment are more or less desirable.") Much of this recent work, however, is concerned with discovering the neurobiological underpinnings of emotions in humans, and it is somewhat far removed from the more immediate needs of roboticists, whose goal is to design functioning, reliable artifacts in the real world.

While many scientists and philosophers argue long and hard about the definitions of this term, classifying theories as "shallow" or "deep" (see Sloman et al., Chapter 8), most roboticists tend to be far more pragmatic and somewhat irreverent toward biology. We instead ask what capabilities can emotions, however defined, endow a robot with that an unemotional robot cannot possess? Minsky (1986) put a spin on this research hypothesis in stating "The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions."

Unfortunately, the situation is even worse than stated above regarding the relevance of emotion to robotics. Most *ethologists* (scientists who study animal behavior in a natural setting) generally use the term *motivation* instead of *emotion*. As much of our group's research historically has been derived from ethological studies, there is a strong tendency to continue to use this term in this chapter over the seemingly vague word *emotions*, even in the context of describing human behavior. Human behavior, at least from my perspective as a roboticist, can be characterized to a great extent through ethological studies (e.g., Blurton Jones, 1972) that abstract away from neural models of brain function in favor of observation. It is also unclear, at least

to me, where within the range of animal species the ability to possess emotions begins. Does a paramecium, sowbug, or dog express emotion? These questions seem better left to others than myself as I am unconvinced that their pursuit will lead to more intelligent robots, which some consider a new species in their own right (Menzel & D'Alusio, 2000).

Motivations, however, tend to be more general than emotions, especially when concerned with human performance (see Chapters 3 [Kelley] and 5 [Rolls]). They often involve the articulation of goals that result in the performance of goal-achieving behavior. Thus, when pressed to define the distinction between emotions and motivations, I state the following working definition (caveat: this is a roboticist speaking): emotions constitute a subset of motivations that provide support for an agent's survival in a complex world. They are not related to the formulation of abstract goals that are produced as a result of deliberation. Motivations and emotions affect behavioral performance, but motivation can additionally lead to the formulation of concrete goal-achieving behavior, at least in humans, whereas emotions are concerned with modulating existing behaviors in support of current activity. In this regard, motivations might additionally invoke specific behaviors to accomplish more deliberate tasks or plans (e.g., strategies for obtaining food).

It is my view that motivations (and emotions) affect the underlying control of a cybernetic system by altering the underlying behavioral parameters of the agent, whether it is biological or artifactual (i.e., a robot). Certain internal states, which are used to represent various motivation/emotional qualities, are maintained by processes that reflect the agent's time course through the environment as well as its perception of the immediate situation. Using this definition, it then becomes our goal, as roboticists, to design systems that can maintain this internal motivational state and use it to produce behavior in ways that are consistent with intelligent performance in the real world.

Motivations/emotions provide two potentially crucial roles for robotics:

1. *Survivability*: Emotions serve as one of the mechanisms to complete autonomy and to help natural systems cope with the world. Darwin (1872/1965) postulated that emotions serve to increase the survivability of a system. Often, a critical situation does not allow time for deliberation, and emotions modulate the behavioral response of the agent directly.
2. *Interaction*: Many robots that are created to function in close proximity to people need to be able to relate to them in predictable and natural ways. This is primarily a limitation of the human, whom we do not have the luxury of reprogramming.

> In order to make robots interact effectively and efficiently with people, it is useful for them to react in ways with which humans are familiar and comfortable.

This chapter will present a range of research results that address the issues above while spanning the phylogenetic complexity of various animal models: i.e., moving up the food chain. We first look at the lowly sowbug as a basis for incorporating motivational behavior, then move up to predatory insects, specifically the praying mantis. Moving into the realm of humans, we then investigate intraspecies behavior, the mother–child relationship, and then interspecies interaction in the relationship of a robotic dog with its owner. Finally, we summarize a relatively complex model of motivations that includes multiple time scales formulated in terms of traits, attitudes, moods, and emotions. Hopefully, the journey through these various biological entities and their robotic counterparts will demonstrate a basis for the commonality of emotion and motivation across all species, while simultaneously encouraging others to loosen their definitional belt a bit regarding emotions.

TOLMAN'S SCHEMATIC SOWBUG AND ITS ROBOTIC COUNTERPART

Our first study looks at a psychological model of the behavior of a sowbug. Tolman introduced his concept of a schematic sowbug, which was a product of his earlier work on purposive behaviorism developed in the early 1920s.

> Initially, in Tolman's purposive behaviorism, behavior implied a performance, the achievement of an altered relationship between the organism and its environment; behavior was functional and pragmatic; behavior involved motivation and cognition; behavior revealed purpose. (Innis, 1999)

Motivation was incorporated into the tight connection between stimulus and response that the prevailing behaviorist view largely ignored. While Tolman also developed the notion of the cognitive map, this was not used in his sowbug model. Instead, motivation was used to create additional inputs to his overall controller, something that more traditional behaviorists tended to ignore. These relations were expressed (Tolman, 1951) in the determination of a behavioral response (B) as a function receiving inputs from environmental stimuli (S), physiological drive (P, or motivation), heredity (H), previous training (T), and age (A).

Tolman (1939) initially proposed the concept of the schematic sowbug as a thought experiment to express the concepts of purposive behaviorism. It was infeasible at the time to think of physically constructing a robotic implementation of his sowbug, even if he had had this as a goal. Space prevents a detailed description of his model (Tolman, 1939; Endo & Arkin, 2001). The sowbug was, in Tolman's view, a simple creature capable of moving autonomously through an obstacle-strewn area in search of food to maintain its existence, where its performance is affected by drives, stimuli, and the other factors mentioned earlier. Described at a high level, Tolman's schematic sowbug consisted of the following:

- A receptor organ: a set of multiple photosensors that perceive light (or any given stimulus) in the environment, mounted on the front-end surface of the sowbug, forming an arc.
- Movement was based on the following components for determining the sowbug's behavioral response:
  - Orientation distribution that indicates the output values of the photosensors serving as environmental stimuli.
  - Orientation/progression tensions, which correspond to the motivational demand. The term *tension* here refers to the readiness of the sowbug to pursue a stimulus, the readiness in this case being derived from hunger (i.e., the hungrier it is [motivation], the greater is its readiness [tension]). *Orientation tension* refers to its readiness to turn, while *progression tension* indicates its readiness to move forward.
  - Orientation vector, which when generated rotates the sowbug in the appropriate direction.
  - Progression distribution, which reflects the strength (or certainty) of a specific stimulus.
  - Progression vectors, which represent the velocities of the left and right motors of the sowbug, similar to what Braitenberg (1984), in his book *Vehicles*, described nearly 40 years later.

The orientation/progression tensions are the key underlying mechanisms by which motivation is introduced into the model. These tensions modulate the sowbug's response to given environmental objects (e.g., food) by representing a variable for motivational drives (e.g., hunger). The result is the orientation need, which directly alters the strength of the agent's motor response to a given stimulus.

What is remarkable is that Tolman's schematic sowbug was the first prototype in history that actually described a behavior-based robotics architecture, to the best of our knowledge. It was a half-century before Brooks (1986) developed the subsumption architecture. Past training and internal

motivational state also affect the sowbug's behavior. Endo and Arkin (2001) created a partial robotic incarnation of Tolman's schematic sowbug. Their software, called eBug[1] (emulated sowbug), supports both simulations (Fig. 9.1, *top*) and robot experiments (Fig. 9.1, *bottom*) that reflect many of the details of Tolman's model.

Tolman's purposive behaviorism spoke to many of the same issues that modern behavior-based robotics architectures address (Arkin, 1998): how to produce intelligent behavior from multiple concurrent and parallel sensorimotor (behavioral) pathways, how to coordinate their outputs meaningfully, how to introduce the notion of goal-oriented behavior, how to include motivation and emotion, and how to permit stages of developmental growth to influence behavior. Tolman's approach has yet to be fully exploited for

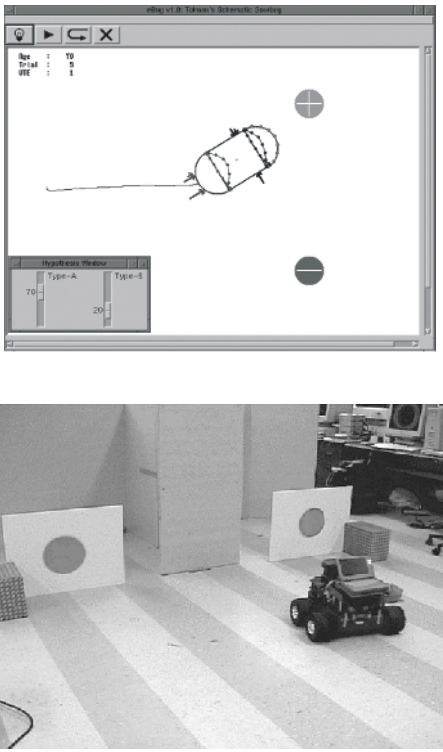


Figure 9.1. (*Top*) eBug simulation of schematic sowbug. Initially, the two colored objects appear to be potential food sources. The robot learns over time that the light-colored object is an attractive (food) stimulus and the darker object is an aversive one and changes its behavior as a result. (*Bottom*) eBug controlling an actual robot to react to stimuli of the same color. The robot is attracted to the object on the right.

motivational and emotional control in real robots and, indeed, may never be as more modern theories of affect now exist, but the work is certainly of historical significance to both roboticists and those in psychology who have the vision to create computational models that can serve as the basis for robotic intelligence.

MOTIVATIONAL BEHAVIOR IN MANTIDS AND ROBOTS

Moving on from the lowly sowbug to animals that prey on insects (i.e., the praying mantis), we strive to understand how the basic drives (motivations to ethologists) affect their behavior and to create similar models for robotic systems. One behavioral model based on schema theory and earlier work applying schema theory to frog behavior (Arbib, 1992) has been used to represent the insect's participation with its world. This involves extension of our robotic schematic–theoretic approach (Arkin, 1989) to incorporate internal motivational processes in addition to external perception. Fortunately, schema theory is quite amenable to this strategy for the mantis, which we have demonstrated both in simulations and in actual robotic experiments (Arkin, Ali, Weitzenfeld, & Cervantes-Perez, 2000). One early example of integrating motivation into navigation was forwarded by Arbib and Lieblich (1977), who made explicit use of drives and motivations integrated into spatial representations that, although not implemented, explained a variety of data on rats running mazes (see also Arbib's Chapter 12). Others (e.g., Steels, 1994; McFarland, & Bosser, 1993) have also explored similar issues experimentally. Our overall approach (Arkin, 1990) is also related to ecological and cognitive psychology (as formulated by Gibson, 1977, and Neisser, 1976, respectively).

As we have seen earlier, the efficacy of visual stimuli to release a response (i.e., type of behavior, intensity, and frequency) is determined by a range of factors: the stimulus characteristics (e.g., form, size, velocity, and spatiotemporal relationship between the stimulus and the animal); the current state of internal variables of the organism, especially those related to motivational changes (e.g., season of the year, food deprivation, time interval between feeding and experimentation); and previous experience with the stimulus (e.g., learning, conditioning, and habituation).

In a joint project between researchers at Georgia Tech and the Instituto Technológico Autónomo de México in Mexico City, a behavioral model was created (Arkin, Cervantes-Perez, & Weitzenfeld, 1998) that captures the salient aspects of the mantid and incorporates four different visuomotor behaviors (a subset of the animal's complete behavioral repertoire):

- ***Prey acquisition***: This behavior first produces orienting, followed by approach (if necessary), then grasping when the target is within reach.
- ***Predator avoidance***: At the most abstract level, this produces flight of the insect, but when considered in more detail there are several forms of avoidance behavior. A large flying stimulus can yield either a ducking behavior or a fight-type response, referred to as "deimatic behavior," where the insect stands up and opens its wings and forearms to appear larger than it is.
- ***Mating***: This is an attractive behavior generated by a female stimulus during the mating season that produces an orienting response in the male, followed by approach, then actual mating.
- ***Chantlitaxia***[2]: This involves an agent's search for a proper habitat for survival and growth. The praying mantis climbs to higher regions (e.g., vegetation) when older, actively searching for a suitable place to hunt.

This model incorporates motivational variables that affect the selection of these motivated behaviors. For predator avoidance, fear is the primary motivator; for prey acquisition, hunger serves a similar purpose; while for mating, the sex drive dominates. These variables are modeled quite simply in this instance, but they may be extended to incorporate factors such as diurnal, seasonal, and climatic cycles and age-related factors as discussed in some of the other models that follow in this chapter. The behavioral controller (Cervantes-Perez, Franco, Velazquez, & Lara, 1993; see also Fig. 9.2) was implemented on a small hexapod robot.

In this implementation, rather than responding to movement, as is the case for the mantis, the robot responds to colors. Green objects represent predators, purple objects represent mates, orange objects that are at least twice as tall as they are wide represent hiding places, and all other orange objects represent prey. The robot maintains three motivational variables that represent its hunger, fear, and sex drive. Initially, the value of each of these variables is set to 0. Arbitrarily, the hunger and sex-drive levels increase linearly with time, with hunger increasing at twice the rate of sex drive. When the robot has contacted a prey or mate, it is considered to have eaten or mated with the object and the relevant motivational variable resets to 0. Contact is determined by the position of the prey or mate color blob in the image captured by the camera mounted on the front of the robot. In this case, the object is considered to be contacted when the bottom of the object blob is in the lower 5% of the image. The fear level remains 0 until a predator becomes visible. At that time, the fear variable is set to a predetermined high value. When the predator is no longer visible, the fear level resets to 0.

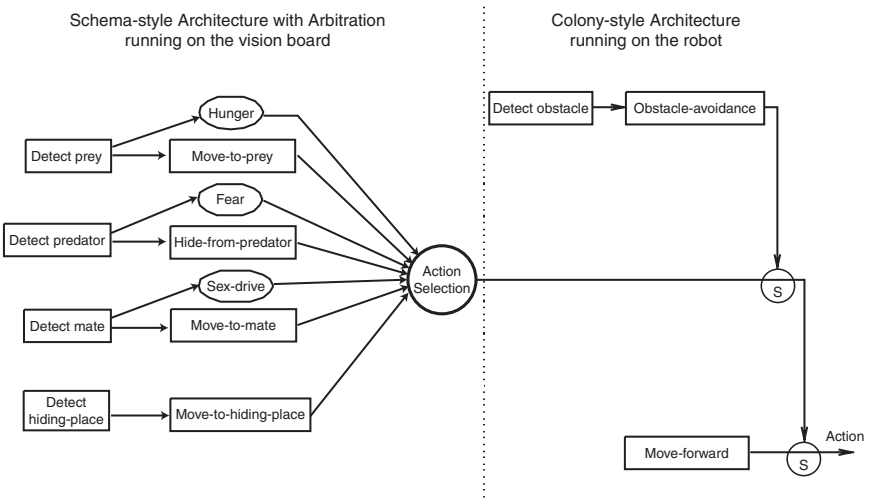


Figure 9.2. Implemented behavioral model. A schema-style architecture (*left*) involves fusing the outputs of the behaviors together cooperatively, while a colony-style architecture (*right*) involves competitive priority-based arbitration for action selection (Arkin, 1998).

Grey Walter's (1953) turtle had an alternative perspective for being motivated, avoiding the use of state variables and encapsulating them within the behaviors themselves: e.g., avoid light unless you are hungry for a recharge. This type of motivation affects the action selection of the behaviors based on external immediate stimuli, whereas a more explicit representation format is encoded in the mantis model.

In the robotic implementation of the mantis model, motivational values directly influence the underlying behavioral control parameters, altering them in a manner consistent with the agent's needs. For example, when the robot has a high hunger level and food appears as a stimulus, the behavior associated with moving toward food is strong. However, if the robot does not have a high hunger motivational value, it will ignore the food stimulus. Similar behavioral reactions occur for the fear motivational value in the presence of predators and the sex-drive variable when a mate is present. The behavior that dominates the overall performance of the hexapod is determined to a great extent by the internal motivational variables and is no longer solely driven by the visual stimuli present in the robot's current field of view. Figure 9.3 illustrates one of many examples of the robot's behavior using this model (Arkin, Ali, Weitzenfeld, & Cervantes-Perez, 2000). It is relatively straightforward to incorporate more complex motivational modeling if it were available, although this work was reserved for more complex species,

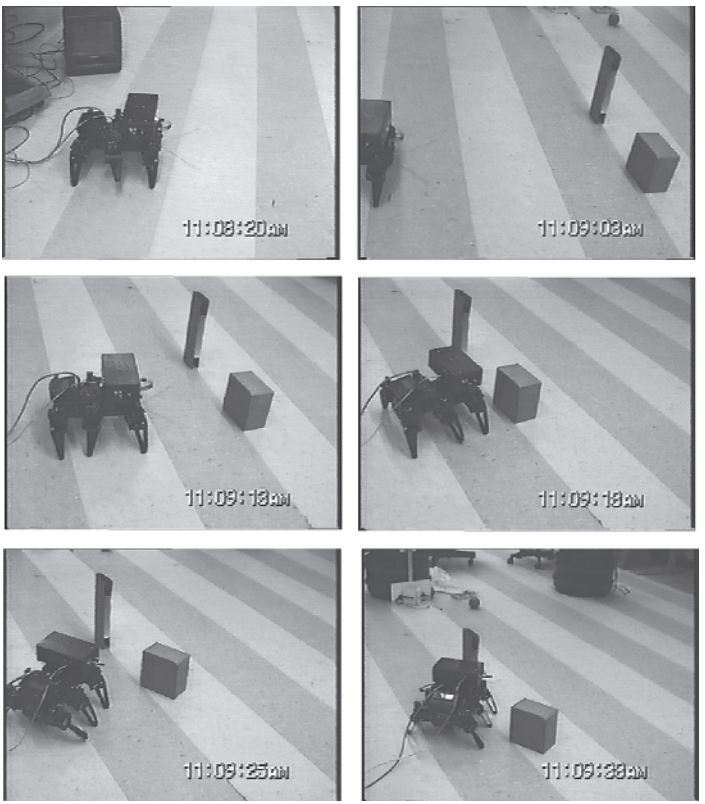


Figure 9.3. *Top photos*: The robot is initially still due to a high fear level as a predator object is in view off to the right. *Middle photos*: Once the predator object is removed, the robot moves toward a prey object to satisfy its hunger. *Bottom photos*: Once its appetite has been satisfied, it ignores the food object and moves toward the tall object, which represents a mate.

as discussed below. Examples of more complex relationships might include an increase of the hunger variable after sex or perhaps a concomitant increase in fear sensitivity during sex. These relationships can in principle be determined neurologically, if not behaviorally. The relationships may change in time and could possibly be said to define the emotional state of the mantis, if emotions were to be attributed to insects.

What this work illustrates is the ability to readily integrate a range of motivational variables, some of which might be construed as emotions (reluctantly, due to concerns over the inherent looseness of this term, as mentioned at the beginning of this chapter) into a behavior-based architecture. The net effect is that the behaviors are no longer solely driven by external perceptions but now also by internal state. These states can be affected by

elapsed time, perceived events, or other factors. Drawing on biologically inspired models of robotic control provides an easy mechanism by which motivational states can be introduced into robots that at least imitate lower life forms in some respects. We now move forward and upward to look at how humans interact with each other and other species as a basis for capturing additional nuances of emotional behavior in robotic systems.

ATTACHMENT THEORY AS A BASIS FOR ROBOT EXPLORATION

We now investigate the relationship of parent and child as a basis for the child's emotional state and its potential for use within robotics. In particular, we look at work that focuses on emotional attachment, which is reflected in what we refer to as "comfort level" (Bowlby, 1969). In humans, both external (exogenous) environmental conditions and internal (endogenous) states determine this level of comfort, reflecting the perceived degree of safety in the current environment and the degree of normal functioning of our internal system.

The input features of comfort consist of these two components, at least for infants (Dunn, 1977). Endogenous factors include hunger, body temperature, pain, and violent or sudden stimulation received by any of the infant's sensors. One of the most significant exogenous factors is environmental familiarity. Hebb's (1946) discrepancy theory states that fear and discomfort are evoked by events that are very different from previous experiences. Dunn (1977) elaborates that whether the past experience with the current situation was pleasant, neutral, or unpleasant is significant. An infant brings to the evaluation of any situation a predisposition threshold on whether to react with pleasure or fear (Stroufe, Waters, & Matas, 1974).

Bowlby (1969) created a theory of attachment in which he pointed out that infants associate certain individuals (caregivers) with security and comfort. They use these people as sources of comfort. In their early years, children want to maintain close proximity to their caregiver, and the degree to which they want to maintain this proximity depends on the circumstances.

> The behavioral hallmark of attachment is seeking to gain and to maintain a certain degree of proximity to the object of attachment, which ranges from close physical contact under some circumstances to interaction or communication across some distance under other circumstances. (Ainsworth & Bell, 1970)

The mother–child relationship is the primary exemplar of this interaction, where the child prefers to be close to the mother in unfamiliar situations,

especially when young. Every attachment object has an attachment bond between itself and the child, where the force of the attachment is situationally dependent and is directed toward reducing the distance of separation between the child and the attachment object (Bowlby, 1969).

Likhachev and Arkin (2000) extrapolated these ideas to working robotic systems. Taking some liberty with formal attachment theory, we now view that an infant (robot) maximizes its exogenous and endogenous comfort components by being physically collocated with its mother (object of attachment). The attractive force is a function of the attachment bond that corresponds to the object, the individual's overall comfort level, and the separation distance.

The intent here is not to create a robotic model of a human child but, rather, to produce useful behavior in autonomous robotic systems (see Chapter 4, Fellous and LeDoux). While robots are not children, there are nonetheless advantages to maintaining an attachment bond with certain individuals or objects (e.g., caregivers, owners, a military base, a fuel supply, or familiar end-users) as they typically satisfy the robot's endogenous needs (e.g., energy) while also providing a high level of familiarity and predictability in the environment. Each attachment object has an associated attachment bond. The degree to which the robot bonds to a particular object depends on how its needs are met by that object and the level of comfort it can achieve. It is interesting to note that brain research on attachment is emerging, though at this point it is not clear how it can be applied to robotics (see Chapter 4, Fellous and LeDoux).

COMPUTATIONAL MODEL OF ATTACHMENT

The result of attachment behavior is a response directed toward increasing or maintaining proximity with the attachment object (Colin, 1996). In a schema-based model, this results in an attractive vector directed toward the object of attachment of varying magnitude dependent upon the separation distance. This vector magnitude (A) represents the intensity of the attachment, which is functionally as follows:

$$A = f(C,\alpha,d)$$

where C is the overall comfort level of a robot, α is the attachment bonding quality between the robot and the particular attachment object in question, and d is the distance between the robot and the attachment object. Specifically, A is defined as the product of the normal attachment maximum level (N), the quality of attachment (α), and the amplification of the comfort component in the function by a proximity factor (D):

$$A = N^{*}\alpha^{*}D^{*}\varphi(\mathrm{C})$$

The normal attachment maximum level N defines the maximum magnitude of the attachment intensity when the object of attachment is a normal "mother," so to speak. The other factors in the function, with the exception of α, are normalized. The attachment bonding quality (α) should be dependent on the quality of care that the attachment object provides for the robot but is set arbitrarily in advance for the results reported below. Setting α to 1 corresponds to a "normal mother" attachment object. Setting α to greater than 1 corresponds to "overcaring mother," whereas decreasing α below 1 corresponds to "undercaring mother." Setting α to 0 corresponds to "no-care mother," resulting in the complete absence of attachment behavior in a robot.

The relationship between A and C expressed in the comfort component ϕ(C) is drawn from the following two sources. Feeney and Noller (1996) describe comfort-seeking intensity in adults as a function of anxiety and fear that is linear for secure adults, where adults typically form attachments with parents, siblings, friends, and partners. It is similarly treated linearly for the robot experiments that appear below. Colin (1996) identifies low-level attachment behavior activation where the behavior has almost no effect but only monitors the proximity and strong activation where the behavior's output overrides virtually all other behaviors in the system when the distance of separation becomes significant. Mathematically, this relationship can be described as follows:

$$\varphi(\mathrm{C}) = \begin{cases} A_l & \text{if } \mathrm{C} > \mathrm{C}_l \\ \dfrac{A_h - A_l}{\mathrm{C}_h - \mathrm{C}_l} * \mathrm{C} - \dfrac{A_h - A_l}{\mathrm{C}_h - \mathrm{C}_l} * \mathrm{C}_h + A_h & \text{if } \mathrm{C}_h < \mathrm{C} < \mathrm{C}_l \\ A_h & \text{if } \mathrm{C} < \mathrm{C}_h \end{cases}$$

where C_l and C_h define low and high comfort activation levels, respectively, and A_l and A_h are the corresponding low and high activation levels.

The proximity factor D is a function of the distance d from the robot to the attachment object. As defined, when the robot is very near the attachment object, the proximity factor is set to 0, in effect negating the attachment force since the robot is already sufficiently close to its object of attachment. This results in a *safe zone*, which is a secure area where the robot receives maximum comfort. When the robot moves outside this safe zone, the proximity factor grows, increasing the overall attachment force, until reaching a maximum at some distance. This area between the safe zone and the distance where the maximum proximity factor occurs is called the *comfort zone*, (Fig. 9.4), which constitutes the normal working region for the robot. Outside of this comfort zone, the attachment force is quite large and generally forces the robot to move into and stay within its comfort zone.

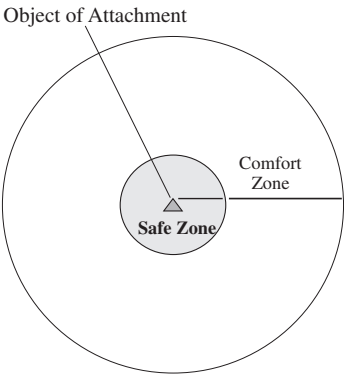


Figure 9.4. The safe and comfort zones of the robot around the object of attachment.

ROBOTIC EXPERIMENTS

Figure 9.5 depicts the effects of various settings of comfort levels on a simulated robot's performance during an exploration task in the presence of an attachment object. As the robot becomes less comfortable, it remains closer to its object of attachment. A complete statistical analysis of these and other results is presented in Likhachev and Arkin (2000). Similar results were obtained during actual robotic experiments (Fig. 9.6).

The notion of emotional comfort as a basis for modulating behavior can have significant impact in controlling a robot's performance as it moves through the world. This is not only of value in ensuring that the robot does not stray from a given task or area with which it is familiar but can also provide a basis for establishing interspecies bonding in entertainment robotics, where creating a pet robot that can effectively relate to a human is of great importance. The next section focuses on various aspects of human–robot interaction in this new application domain for robots.

CANINE ETHOLOGY IN SUPPORT OF HUMAN–ROBOT BONDING

One of the principal goals of entertainment robotics is to provide the illusion of life in a robot to a human. A strategy we have chosen to follow, in joint work with Sony Corporation (Arkin, Fujita, Takagi, & Hasegawa, 2001, 2003), is to develop a computational model of behavior based on ethology. In this work, we engage the concept of motivational behavior in nonhuman animals, specifically dogs, with that of emotionality experienced in humans. One of the goals is to produce appropriate emotional responses in people through observation and interaction with a robotic artifact. This requires

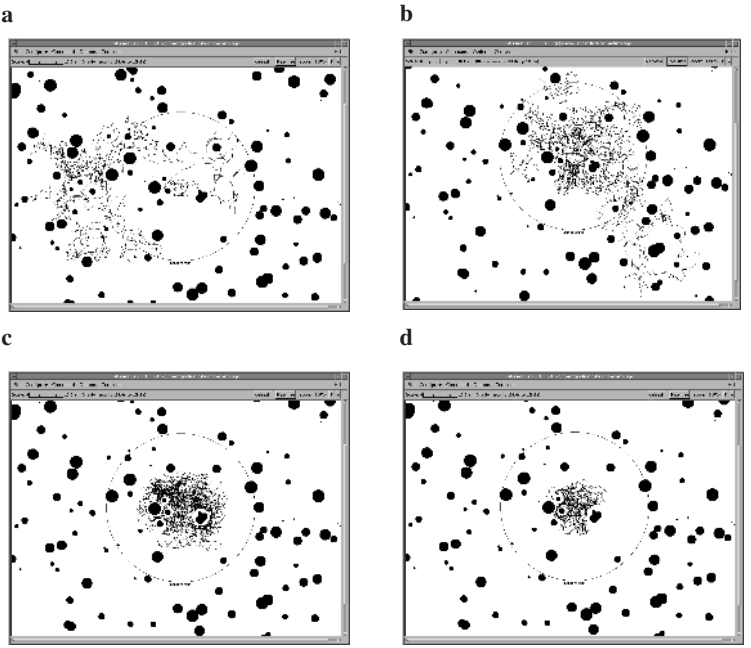


Figure 9.5. Three-minute runs of exploration behavior with the attachment object of attachment located at the center of the circle that defines the comfort zone. (*a*) No attachment behavior. (*b*) Attachment behavior with comfort level set at 1.0 (maximum comfort). (*c*) Comfort level set at 0.0 (neutral comfort). (*d*) Comfort level set at –1.0 (maximum discomfort).

generating natural behaviors as well as maintaining motivational/emotional states within the robot. Studies of the manifestation of emotions in humans and their similar occurrence as motivational behavior in animals can provide support for effective interactivity between a robot and a human (Breazeal, 2002; Dautenhahn & Billard, 1999; Fujita et al., 2001; see also Chapter 10, Breazeal). By incorporating aspects of emotional and motivational behavior into a robotic architecture, we and others (e.g., Breazeal and Scassellati, 1999) contend that a greater ability to relate to the end-user is provided.

The primary robotic system used for this work is Sony's AIBO, a highly successful commercial product (Fig. 9.7). A broad range of behaviors is available, organized into multiple subsystems (Arkin, Fujita, Takagi, & Hasegawa, 2001). Their selection is related to the motivational state of the robot, maintained in what is referred to as the instinct/emotion (I/E) model. The model of Ekman and Davidson (1994) has been influential in this work and consists of six basic emotional states: happiness, anger, sadness, fear, surprise, and disgust (cf. Chapter 5 and Ekman's dimension as illustrated by Kismet

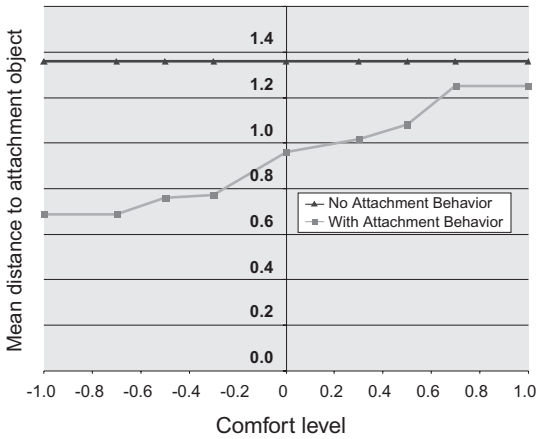


Figure 9.6. (*Top*) Nomad robot conducting 5-minute explorations. The object of attachment is the tree. (*Bottom*) Results showing how average distance from attachment object increases as robot's comfort level increases.

in Breazeal's Chapter 10). Takanishi's (1999) approach is also used to reduce the overall internal state space into three dimensions: pleasantness, arousal, and confidence. The six basic emotional states are located within this three-dimensional space. By establishing predefined levels of internal variables, such as hunger and thirst, and determining how the current state of the robot relates to those thresholds, pleasantness can be assessed. If these variables remain within the regulated range, the pleasantness is high. Arousal is controlled by both circadian rhythm and unexpected stimuli, while confidence is determined by the certainty of recognized external stimuli.

The resulting emotional values affect the action-selection process for behavior eligibility for execution. Drawing on aspects of both McFarland's

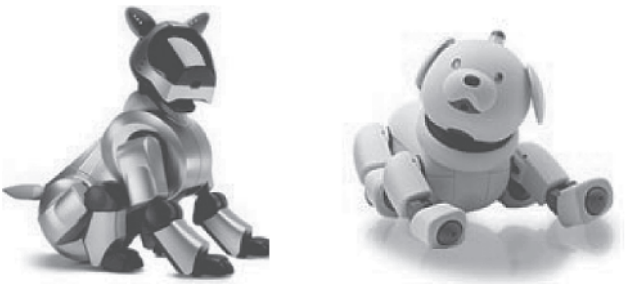


Figure 9.7. Various AIBO robots.

(1974) motivational space and Blumberg's (1994) action-selection mechanisms, particular behaviors are scheduled for execution by the robot that are consistent with the current set of environmental stimuli and the internal state of the robot itself. In the action-selection module, a behavior is selected based on inputs derived from external stimuli (releasing mechanisms, cf. Chapter 7, Ortony et al., and Chapter 8, Sloman et al.) and the robot's current motivational state variables. A state-space diagram represents the mapping from these inputs onto the appropriate behavior to be activated (Arkin, Fujita, Takagi, & Hasegawa, 2003).

Further extension of this research has resulted in the emotionally grounded (EGO) architecture (Fujita et al., 2001a,b), leading to potential applications in humanoid robots (Arkin, Fujita, Takagi, & Hasegawa, 2003). The generated motion patterns can be affected by the emotions themselves. Fujita et al. (2001b) specifically addresses the symbol-grounding problem (Harnard, 1990) in this architecture, allowing the robot to learn to associate behaviors with specific symbols through the use of an emotionally grounded symbol, where the physically grounded symbol is associated with the change of internal state that occurs when the robot applies a behavior in response to the object. For example, when the robot hears the symbol's name spoken, it knows which behavior(s) is associated with that symbol and can produce a change in its internal motivations. Thus, in a sense, the robot knows the meaning of the symbol in the way in which it affects both its internal state and what behaviors are correct to use in the associated object's presence. The symbols are grounded not only perceptually, by associating the correct perceptual stimuli with the spoken symbol, but also behaviorally, by producing the appropriate behavioral response in the presence of the stimuli that acts in a manner to produce a change in internal variables consistent with the I/E model. This use of symbols for emotional modeling diverges somewhat from strict ethology, especially when compared to the more faithful canine behavioral modeling employed (Arkin, Fujita, Takagi, & Hasegawa,

2003), but the intent is to create robotic artifacts that successfully entertain and engender human–robot emotional bonding.

More recent work has expanded this architecture into humanoid behavior for the Sony Dream Robot (SDR-4X) (Fig. 9.8), which is also capable of emotional expression. Extensions to the EGO architecture include the introduction of a deliberative layer capable of planning (Arkin, Fujita, Takagi, & Hasegawa, 2003). Proprietary issues prevent a more thorough discussion at this time.

AIBO and SDR are neither dogs nor humans. They are entertainment robots intended to provide interesting and engaging experiences with people. The use of emotions in these systems bears that particular purpose in mind. An underlying belief is that if a robot is capable of expressing itself not only through speech but also emotionally, it is more likely to be accepted by consumers.

In all of the robots discussed thus far, "internal state" refers to the maintenance of a set of variables that reflect the emotional/motivational state of the machine. This appears to be consistent with Dolan's (2002) definition of emotion, which appears at the beginning of this chapter, with the exception that there is relatively little complexity involved in our implementation. These variables are updated continuously by arriving sensory information and, in some cases, by circadian rhythms and other factors. This set of states acts on the behavioral regime of the robot to modulate and/or select behaviors that best reflect the current set of emotional conditions. Considerably more complex models are possible, one of which is discussed in the next section.

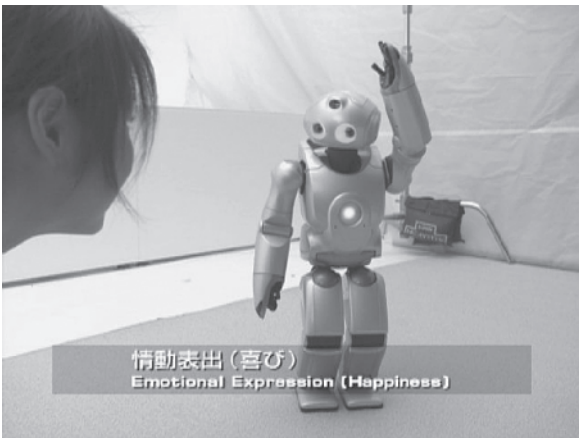


Figure 9.8. Sony Dream Robot (SDR-4X) manifesting a happy greeting consistent with its emotional state

A NEW MODEL: SUMMARY AND CONCLUSION

We conclude this chapter with the presentation of a new model under development in our laboratory at Georgia Tech and then a summary of the overarching themes.

Traits, Attitudes, Moods, and Emotions

We are not currently aware of any single computational model that captures the interaction between a wide range of time-varying, affect-related phenomena, such as personality traits, attitudes, moods, and emotions. In humans, each of these components performs a distinct adaptive function. It is our research hypothesis that providing autonomous robots with similar, easily recognizable affective cues may facilitate robot–human interaction in complex environments.

Moshkina and Arkin (2003) proposed a new affect model which incorporates and unites traits, attitudes, moods, and emotions (TAME) as separate components with well-characterized interfaces in order to produce multiscale temporal affective behavior for use in a behavior-based robotic system. The TAME model draws from a number of related theories of personality, mood, emotion, and attitudes, but it is intended to serve primarily as a basis for producing intelligent robotic behavior and not as a cognitive model of affect and personality.

The personality and affect module is currently being integrated into the autonomous robot architecture (AuRA) (Arkin & Balch, 1997) as embodied in the *MissionLab*[3] mission specification system (Mackenzie, Arkin, & Cameron, 1997). The personality and affect module modifies the underlying behavioral parameters, which directly affect currently active behaviors, similar to earlier work from our laboratory in homeostatic control (Arkin, 1992) and the work on the praying mantis described above. The conceptual view of the TAME model is presented in Figure 9.9.

Psychologists have factored affective responses into at least these four components: traits, attributes, moods, and emotions. By applying temporal attributes to this differentiated set of affective response factors, the generation of affective behavior in robots can be clarified through the generation of new computational models that enable the composition of these time-varying patterns. This will hopefully give rise to mechanisms by which a robot's responses can be more appropriately attuned to a human user's needs, in both the short and long terms. It is not intended to validate any particular theories of human affective processing but, rather, to assist in creating better human–robot interaction.

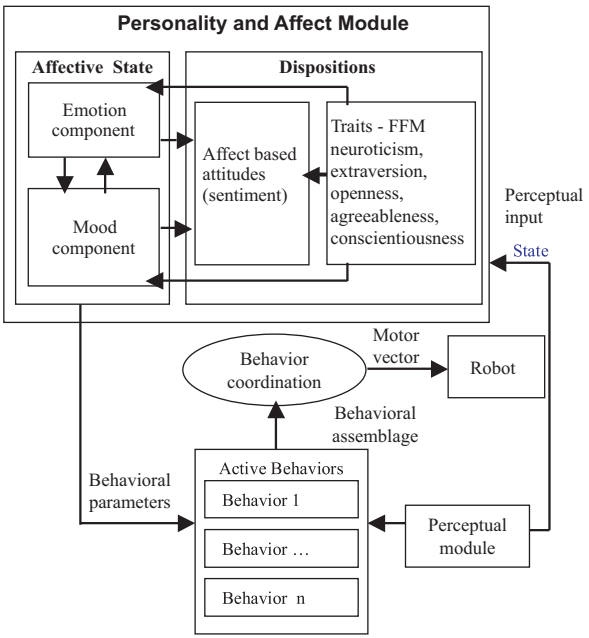


Figure 9.9. Integrated model of personality and affect (traits, attitudes, moods, and emotions [TAME] model). FFM, five-factor model.

The four major components operate in different time and activation scales. Emotions are high-activation and short-term, while moods are low-activation and relatively prolonged. Traits and attitudes determine the underlying disposition of the robot and are relatively time-invariant. The basis for each of these four components is discussed briefly below (see also Ortony et al., Chapter 7).

Traits serve as an adaptation mechanism to specialized tasks and environments, whereas emotions mobilize the organism to provide a fast response to significant environmental stimuli. The five-factor model of personality developed by McCrae and Costa (1996) serves as the basis for the trait components. Trait dimensions include openness (O), agreeableness (A), conscientiousness (C), extroversion (E), and neuroticism (N). Traits influence a wide range of behaviors and are not limited to emotionally charged situations.

Emotion, in the TAME context, is an organized reaction to an event that is relevant to the needs, goals, or survival of the organism (Watson, 2000). It is short in duration, noncyclical, and characterized by a high activation state and significant energy and bodily resource expenditure. A typical set of emotions to which we subscribe includes joy, interest, surprise, fear, anger, sadness, and disgust (Watson, 2000); these are continuously dynamically generated as emotion-eliciting stimuli are detected.

Moods bias behavior according to favorable/unfavorable environmental conditions and are defined by the two independent categories of positive and negative affect (Revelle, 1995; see also Chapter 7, Ortony et al.). They constitute a continuous affective state that represents low activation and low intensity and, thus, expends less energy and bodily resources than emotion. Moods are mainly stimulus-independent and exhibit cyclical (circadian) variation according to time of day, day of the week, and season.

An *attitude* is a "learned predisposition to respond in a consistently favorable or unfavorable manner with respect to a given object" (Breckler & Wiggins, 1989). Attitudes guide behavior toward desirable goals and away from aversive objects and facilitate the decision-making process by reducing the decision space complexity. They are relatively time-invariant, object/situation-specific, and influenced by affect and result in a certain behavior toward the object.

To test the TAME model, a partial integration of the personality and affect module into the MissionLab system was undertaken, which is a supplemented version of the AuRA (Arkin & Balch, 1997; see also Moshkina & Arkin, 2003, for additional details). Research is now under way on the administration of formal usability studies to determine whether this form of affect can play a significant role in improving a user's experience with a robot.

SUMMARY AND CONCLUSION

In the end, what can we learn from this journey through a broad range of motivations/emotions that span multiple species? I propose the following:

- Emotions, at least to a roboticist, consist of a subset of motivations that can be used to dynamically modulate ongoing behavioral control in a manner consistent with survival of the robotic agent (Arkin & Vachtsevanos, 1990). The nuances surrounding which species possess *emotions* versus *motivations* and the terminological differences between these terms are best left to nonroboticists in my opinion as it is unclear if the resolution to these semantic differences will have any impact whatsoever on our ability to build more responsive machines. Our community, however, sorely needs more and better computational models and processes of affect that effectively capture these components within a behavioral setting.
- Human–robot interaction can be significantly enhanced by the introduction of emotional models that benefit humans as much as robots.

- Motivational/emotional models can be employed that span many different organisms and that can match the requirements of an equally diverse robotic population, ranging from vacuum cleaners to military systems to entertainment robots and others. All of these systems need to survive within their ecological niche and must respond to a broad range of threats toward their extinction or obsolescence. The principle of biological economy would argue that emotions/motivations exist in biology to serve a useful purpose, and it is our belief that robots can only benefit by having a similar capability at their disposal.
- The diversity of emotional models is something to celebrate and not lament as they all can potentially provide fodder for robotic system designers. As I have often said, I would use phlogiston as a model if it provided the basis for creating better and more intelligent robots, even if it does not explain natural phenomena accurately.

Finally, there is much more work to be done. This branch of robotics has been enabled due to major computational and hardware advances that have come into existence only within the past few decades. As such, it is an exciting time to be studying these problems in the context of artificial entities.

Notes

The author is deeply indebted to the researchers whose work is reported in this chapter, including Yoichiro Endo, Khaled Ali, Francisco Cervantes-Perez, Alfredo Weitzenfeld, Maxim Likhachev, Masahiro Fujita, Tsubouchi Takagi, Rika Hasegawa, and Lilia Moshkina. Research related to this article has been supported by the National Science Foundation, Defense Advanced Research Projects Agency (DARPA), and the Georgia Tech Graphic, Visualization and Usability (GVU) Center. The author also thanks Dr. Doi, the director of Digital Creatures Laboratory, Sony, for his continuous support for our research activity.

1. eBug is available on line (http://www.cc.gatech.edu/ai/robot-lab/research/ebug/)

2. Etymology: from the Nahuatl word *chantli*, which means shelter or refuge, and word *taxia*, the Latin for attraction (Cervantes-Perez, personal communication, 2003).

3. MissionLab is freely available on line (www.cc.gatech.edu/ai/robot-lab/research/MissionLab.html)

References

Ainsworth, M. D. S., & Bell, S. M. (1970). Attachment, exploration and separation: Illustrated by the behaviour of one-year-olds in a strange situation. *Child Development*, *41*, 49–67.

Arbib, M. A. (1992). Schema theory. In S. Shapiro (Ed.), *Encyclopedia of artificial intelligence* (2nd ed., pp. 1427–1443). New York: Wiley.

Arbib, M. A., & Lieblich, I. (1977). Motivational learning of spatial behavior. In J. Metzler (Ed.), *Systems neuroscience* (pp. 221–240). London: Academic Press.

Arkin, R. C. (1989). Motor schema-based mobile robot navigation. *International Journal of Robotics Research*, *8*, 92–112.

Arkin, R. C. (1990). The impact of cybernetics on the design of a mobile robot system: A case study. *IEEE Transactions on Systems, Man, and Cybernetics*, *20*, 1245–1257.

Arkin, R. C. (1992). Homeostatic control for a mobile robot: Dynamic replanning in hazardous environments. *Journal of Robotic Systems*, *9*, 197–214.

Arkin, R. C. (1998). *Behavior-based robotics*. Cambridge, MA: MIT Press.

Arkin, R. C., Ali, K., Weitzenfeld, A., & Cervantes-Perez, F. (2000). Behavioral models of the praying mantis as a basis for robotic behavior. *Journal of Robotics and Autonomous System*, *32*, 39–60.

Arkin, R. C., & Balch, T. (1997). AuRA: Principles and practice in review. *Journal of Experimental and Theoretical Artificial Intelligence*, *9*, 175–189.

Arkin, R. C., Cervantes-Perez, F., & Weitzenfeld, A. (1998). Ecological robotics: A schema-theoretic approach. In R. C. Bolles, H. Bunke, & H. Noltemeier (Eds.), *Intelligent robots: Sensing, modelling and planning* (pp. 377–393). Singapore: World Scientific.

Arkin, R., Fujita, M., Takagi, T., & Hasegawa, R. (2003). An ethological and emotional basis for human–robot interaction. *Robotics and Autonomous Systems*, *42*, 191–201.

Arkin, R. C., Fujita, M., Takagi, T., & Hasegawa, R. (2001). Ethological modeling and architecture for an entertainment robot. In *2001 IEEE International Conference on Robotics and Automation* (pp. 453–458). Seoul, Korea: IEEE.

Arkin, R. C., & Vachtsevanos, G. (1990). Techniques for robot survivability. *Proceedings of the 3rd International Symposium on Robotics and Manufacturing* (pp. 383–388).Vancouver, BC.

Blumberg, B. (1994). Action-selection in hamsterdam: Lessons from ethology. In D. Cliff et al. (Eds.), *From Animals to Animats 3* (pp. 108–117). Cambridge, MA: MIT Press.

Blurton Jones, N. (1972). *Ethological studies of child behavior*. London: Cambridge University Press.

Bowlby, J. (1969). *Attachment and loss. I: Attachment*. London: Hogarth.

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: MIT Press.

Breazeal, C. (2002). *Designing sociable robots*. Cambridge, MA: MIT Press.

Breazeal, C., & Scassellati, B. (1999). How to build robots that make friends and influence people. In *Proceedings of the International Conference on Intelligent Robots and Systems 1999* (pp. 858–863).

Breckler, S. J., & Wiggins, E. C. (1989). On defining attitude and attitude theory: Once more with feeling. In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function* (pp. 407–429). Mahwah, NJ: Erlbaum.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation, RA-2*, 14–23.

Cervantes-Perez, F., Franco, A., Velazquez, S., & Lara, N. (1993). A schema theoretic approach to study the "chantitlaxia" behavior in the praying mantis. In *Proceedings of the First Workshop on Neural Architectures and Distributed AI: From Schema Assemblages to Neural Networks*, USC.

Colin, V. L. (1996). *Human attachment*. New York: McGraw-Hill.

Darwin, C. (1965). *The expression of the emotions in man and animals*. Chicago: University of Chicago Press. (Original work published 1872)

Dautenhahn, K., & Billard, A. (1999). Bringing up robots or psychology of socially intelligent robots: From theory to implementation. In *Proceedings of the 3rd International Conference on Autonomous Agents* (pp. 366–367). Seattle, WA: Association for Computing Machinery.

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science*, *298*, 1191–1194.

Dunn, J. 1977). *Distress and comfort*. Cambridge, MA: Harvard University Press.

Ekman, P., & Davidson, R. J. (1994). *The nature of emotion*. Oxford: Oxford University Press.

Endo, Y., & Arkin, R. C. (2001). Implementing Tolman's schematic sowbug: behavior-based robotics in the 1930s. *2001 IEEE International Conference on Robotics and Automation* (pp. 487–484). Seoul, Korea.

Feeney, J., & Noller, P. (1996). *Adult attachment*. London: Sage.

Fujita M., Hasegawa, R., Costa, G., Takagi, T., Yokono, J., & Shimomura, H. (2001a). Physically and emotionally grounded symbol acquisition for autonomous robots. In *Proceedings of the AAAI Fall Symposium: Emotional and Intelligent II* (pp. 43–46).

Fujita M., Hasegawa, R., Costa, G., Takagi, T., Yokono, J., & Shimomura, H. (2001b). An autonomous robot that eats information via interaction with human and environment. In *Proceedings of the IEEE ROMAN-01*.

Gibson, J. J. (1977). The theory of affordances. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing*. Mahwah, NJ: Erlbaum.

Harnard, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346.

Hebb, D. O. (1946). On the nature of fear. *Psychological Review*, *53*, 259–276.

Innis, N. K. (1999). Edward C. Tolman's purposive behaviorism. In *Handbook of behaviorism* (pp. 97–117). New York: Academic Press.

Likhachev, M., & Arkin, R. C. (2000). Robotic comfort zones. In *Proceedings of the SPIE Sensor Fusion and Decentralized Control in Robotic Systems III* (pp. 27–41).

MacKenzie, D., Arkin, R. C., & Cameron, R. (1997). Multiagent mission specification and execution. *Autonomous Robots*, *4*, 29–52.

McCrae, R. R., & Costa, P. T. (1996). Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In *Five-factor model of personality* (pp. 51–87). New York: Guilford.

McFarland, D. (Ed.) (1974). *Motivational control systems analysis*. London: Academic Press.

McFarland, D., & Bosser, T. (1993). *Intelligent behavior in animals and robots*. Cambridge, MA: MIT Press.

Menzel, P., & D'Alusio, F. (2000). *Robo sapiens*. Cambridge, MA: MIT Press.

Minsky, M. (1986). *The society of the mind*. New York: Simon & Schuster.

Moshkina, L., & Arkin, R. C. (2003). On TAMEing robots. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*. Piscataway, NJ, 2003.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. New York: Freeman.

Revelle, W. (1995). Personality processes. *Annual Review of Psychology, 46*, 295–328.

Steels, L. (1994). A case study in the behavior-oriented design of autonomous agents. In D. Cliff, et al. (Eds.), *From Animals to Animats 3* (pp. 445–452). Cambridge, MA: MIT Press.

Stroufe, L. A., Waters, E., & Matas, L. (1974). Contextual determinants of infant affective response. In M. Lewis & L. A. Rosenblum (Eds.), *The origins of fear*. New York: Wiley.

Takanishi, A. 1999). An anthropomorphic robot head having autonomous facial expression function for natural communication with human. In *9th International Symposium of Robotics Research* (*ISRR99*) (pp. 197–304). London: Springer-Verlag.

Tolman, E. C. (1939). Prediction of vicarious trial and error by means of the schematic sowbug. *Psychological Review, 46*, 318–336.

Tolman, E. C. (1951). *Behavior and psychological man*. Berkeley: University of California Press.

Walter, G. (1953). *The living brain*. New York: Norton.

Watson, D. (2000). *Mood and temperament*. New York: Guilford.

This page intentionally left blank

10 Robot Emotion

A Functional Perspective

CYNTHIA BREAZEAL
AND RODNEY BROOKS

Robots are becoming more and more ubiquitous in human environments. The time has come when their ability to intelligently and effectively interact with us needs to match the level of technological sophistication they have already achieved, whether these robots are tools, avatars (human surrogates), partners (as in a team) or cyborg extensions (prostheses). Emotion-inspired mechanisms can improve the way autonomous robots operate in a human environment with people, and can improve the ability of these robots to effectively achieve their own goals. Such goals may be related to accomplishing tasks or satisfying motivations, and they may be achieved either autonomously or in cooperation with a person.

In order to do this in a way that is natural for humans, the robot needs to be designed with a social model in mind. We illustrate this concept by describing in detail the design of Kismet, an anthropomorphic robot that interacts with a human in a social way, focusing on its facial and vocal expressions and gaze direction. Kismet's architecture includes a cognitive system that is tightly coupled to a separate emotive system. Each is designed as interacting networks of "specialists" that are activated when specific conditions are met. The cognitive component is

responsible for perceiving and interpreting events, and for selecting among a hierarchy of goal-achieving behaviors, in accordance with its current motivational drives. It is primarily concerned with homeostasis and "well being." The emotive system implements eight basic emotions that are proposed to exist across species. It detects those internal and external events that have affective value, and motivates either task-based or communicative behavior to pursue beneficial interactions and to avoid those that are not beneficial by modulating the operation of the cognitive component. When placed in a realistic social setting, these two systems interact to achieve lifelike attention bias, flexible decision making, goal prioritization and persistence, and effective communication where the robot interacts in a natural and intuitive way with the person to achieve its goals.

Why should there be any serious research at all on the possibility of endowing robots with emotions? Surely this is the antithesis of engineering practice that is concerned with making functional devices rather than ones which invoke emotions in people—the latter is the realm of art or, at best, design.

Over the last 100 years, the average home in the Western world has seen the adoption of new technologies that at first seemed esoteric and unnecessarily luxurious. These include electricity, refrigeration, running hot water, telephone service, and most recently wideband internet connections. Today, the first few home robots are starting to be sold in conventional retail stores. Imagine the world 50 years from now, when robots are common in everybody's home. What will they look like, and how will people interact with them?

Electricity, refrigeration, and running hot water are utilities that are simply present in our homes. The first robots that have appeared in homes are largely also a utility but have a presence that triggers in some people responses that are normally triggered only by living creatures. People do not name their electrical outlets and talk to them, but they do name their robots and, today at least, carry on largely one-sided social interactions with them. Will it make sense, in the future, to capitalize on the tendency of people for social interactions in order to make machines easier to use?

Today's home-cleaning robots are not aware of people as people. They are not even aware of the difference between a moving obstacle and a static obstacle. So, today they treat people as they would any other obstacle, something to be skirted around while cleaning right up against them.

Imagine a future home-cleaning robot, 50 years from now, and the capabilities it might plausibly have. It should be aware of people as people

because they have very different behaviors from static objects such as chairs or walls. When a person happens to be walking down the hallway where a robot is cleaning, it might determine that there is a person walking toward it. A reasonable behavior would be to then get out of the way. By estimating the person's mood and level of attention to his or her surroundings, the robot can determine just how quickly it should get out of the way. If a person points to the corner of a room and says "Clean more over here," the robot should understand the focus of attention of the person as that room corner, rather than the end of the finger used to point. If the person utters an angry "No" directed toward the robot, it should reevaluate its recent actions and adapt its future behavior in similar circumstances.

These are examples of the robot reading cues, some emotional, from a person; but a robot might also interact with people more easily if it provides social cues.

As the robot notices a person approaching down a corridor, it might, like a dog, make brief eye contact with the person (here, the person has to estimate its gaze direction) and then give a bodily gesture that it is accommodating the person and the person's intent. When cleaning that dirty corner, it might express increased levels of both frustration and determination as it works on a particularly difficult food spill on the carpet so that the person who directed its attention to that corner can rest assured that it has understood the importance of the command and that it will do what it takes to fulfill it. Upon hearing that angry "No," the robot may express its chagrin in an emotional display so that the person knows intuitively that the command has been heard and understood.

A robot with these sorts of capabilities would seem luxuriously out of place in our current homes, but such capabilities may come to be expected as they will make interaction with robots more natural and simple.

In order to get to these future levels of social functionality, we need to investigate how to endow our robots with emotions and how to enable them to read social and emotional cues from people.

FUNCTIONAL ROLES OF EMOTIONS

All intelligent creatures that we know of have emotions. Humans, in particular, are the most expressive, emotionally complex, and socially sophisticated of all (Darwin, 1872).

To function and survive in a complex and unpredictable environment, animals (including humans) are faced with applying their limited resources (e.g., muscles, limbs, perceptual systems, mental abilities, etc.) to realize multiple goals in an intelligent and flexible manner (Gould, 1982). Those

species considered to be the most intelligent tend to exist in complex and dynamic social groups, where members have to communicate, cooperate, or compete with others.

Two conceptually distinct and complementary information-processing systems, cognition and emotion, evolved under such social and environmental pressures to promote the health and optimal functioning of the creature (Damasio, 1994; Izard & Ackerman, 2000). As argued in Chapter 7 (Ortony et al.), the cognitive system is responsible for interpreting and making sense of the world, whereas the emotive system is responsible for evaluating and judging events to assess their overall value with respect to the creature (e.g., positive or negative, desirable or undesirable).

Emotion theorists agree that the cognitive and emotive systems are interrelated. One view privileges the cognitive system where the cognitive processes of appraisal and attribution recruit emotions. Others see emotion and cognition as being reciprocally interrelated, recognizing that each emotion often recruits and organizes cognitive processes and behavioral tendencies in a specific manner to the adaptive advantage of the creature (Izard, 1993). For instance, according to Izard (1993), a unique function of sadness is its ability to slow the cognitive and motor systems. Termine and Izard (1988) found that mothers' display of sorrow through facial and vocal expression during face-to-face interactions with their 9-month-old infants significantly decreased their babies' exploratory play. In adults, the slowing of cognitive processes may enable a more careful and deliberate scrutiny of self and circumstances, allowing the individual to gain a new perspective to help improve performance in the future (Tomkins, 1963).

Numerous scientific studies continue to reveal the reciprocally interrelated roles that cognition and emotion play in intelligent decision making, planning, learning, attention, communication, social interaction, memory, and more (Isen, 2000). Emotion plays an important role in signaling salience, to guide attention toward what is important and away from distractions, thereby helping to effectively prioritize concerns (Picard, 1997). Isen (2000) has studied the numerous beneficial effects that mild positive affect has on a variety of decision-making processes for medical diagnosis tasks, for example, facilitating memory retrieval (Isen, Shalker, Clark, & Karp, 1978); promoting creativity and flexibility in problem solving (Estrada, Isen, & Young, 1994); and improving efficiency, organization, and thoroughness in decision making (Isen, Rosenzweig, & Young, 1991). As argued by Isen (1999), negative affect allows us to think in a highly focused way when under negative, high-stress situations. Conversely, positive affect allows us to think more creatively and to make broader associations when in a relaxed, positive state.

Furthermore, scientists are finding that whereas too much emotion can hinder intelligent thought and behavior, too little emotion is even more

problematic. The importance of emotion in intelligent decision making is markedly demonstrated by Damasio's (1994) studies of patients with neurological damage that impairs their emotional systems. Although these patients perform normally on standardized cognitive tasks, they are severely limited in their ability to make rational and intelligent decisions in their daily lives. For instance, they may lose a great deal of money in an investment. Whereas healthy people would become more cautious and stop investing in it, these emotionally impaired patients do not. They cannot seem to learn the link between bad feelings and dangerous choices, so they keep making the same bad choices again and again. The same pattern is repeated in their relationships and social interactions, resulting in loss of jobs, friends, and more (Damasio, 1994; Picard, 1997). By looking at highly functioning autistics, we can see the crucial role that emotion plays in normal relations with others. They seem to understand the emotions of others like a computer, memorizing and following rules to guide their behavior but lacking an intuitive understanding of others. They are socially handicapped, not able to understand or interpret the social cues of others to respond in a socially appropriate manner (Baron-Cohen, 1995).

Emotion Theory Applied to Robots

This chapter presents a pragmatic view of the role emotion-inspired mechanisms and capabilities could play in the design of autonomous robots, especially as it is applied to human–robot interaction (HRI). Given our discussion above, many examples could be given to illustrate the variety of roles that emotion-inspired mechanisms and abilities could serve a robot that must make decisions in complex and uncertain circumstances, working either alone or with other robots. Our interest, however, concerns how emotion-inspired mechanisms can improve the way robots function in the human environment and how such mechanisms can improve robots' ability to work effectively in partnership with people. In general, these two design issues (robust behavior in the real world and effective interaction with humans) are extremely important given that many real-world autonomous robot applications require robots to function as members of human–robot teams.

We illustrate these advantages with a design case study of the cognitive and emotion-inspired systems of our robot, Kismet. We have used the design of natural intelligence as a guide, where Kismet's cognitive system enables it to figure out what to do and its emotive system helps it to do so more flexibly in complex and uncertain environments (i.e., the human environment), as well as to behave and interact with people in a socially acceptable and natural manner.

This endeavor does not imply that these emotion-based or cognition-based mechanisms and capabilities must be in some way identical to those in natural systems. In particular, the question of whether or not robots could have and feel human emotions is irrelevant to our purposes. Hence, when we speak of robot emotions, we do so in a functional sense. We are not claiming that they are indistinguishable from their biological correlates in humans and other animals. Nonetheless, we argue that they are not "fake" because they serve a pragmatic purpose for the robot that mirrors their natural analogs in living creatures.

Furthermore, the insights these emotion-based and affect-based mechanisms provide robot designers should not be glossed over as merely building "happy" or entertaining robots. To do so is to miss an extremely important point: as with living creatures, these emotion-inspired mechanisms modulate the cognitive system of the robot to make it function better in a complex, unpredictable environment—to allow the robot to make better decisions, to learn more effectively, to interact more appropriately with others—than it could with its cognitive system alone.

EMOTION-INSPIRED ABILITIES IN HUMAN–ROBOT INTERACTION

There are a diverse and growing number of applications for robots that interact with people, including surgery, scientific exploration, search and rescue, surveillance and telepresence, museum docents, toys, entertainment, prosthetics, nursemaids, education, and more.

The demands that the robot's architecture must address depend on a number of issues. For instance, is the robot completely autonomous, teleoperated by a human, or somewhere in between? Is the robot's environment controlled and predictable, or is it complex and unpredictable, even potentially hostile? Is the robot designed to perform a specific task, or must it satisfy multiple and potentially competing goals? Is the robot expected to function in complete isolation or in cooperation with others? Does the human use the robot to mediate his or her own actions (i.e., as a tool or prosthesis)? Does the human cooperate with the robot as a teammate? Does the robot provide some form of companionship, such as a pet? Is it expected to enter into long-term relationship with a particular person, such as a nursemaid?

In Breazeal (2003d), we classify these applications into four different paradigms of interaction (see below). Each is distinguished from the others based on the mental model a human has of the robot when interacting with it. Furthermore, for each there is a wide assortment of advantages that giving robots skills and mechanisms associated with emotion could play.

- Intelligent behavior in a complex, unpredictable environment
- The ability to sense and recognize affect and emotion
- The ability to express affect and internal state in familiar human terms
- The ability to respond to humans with social adeptness

Robot as Tool

In the first paradigm, the human views the robot as a device that is used to perform a task. The amount of robot autonomy varies (and, hence, the cognitive load placed on the human operator) from complete teleoperation to a highly self-sufficient system that need only be supervised at the task level. Consider a specialist who uses a robot to perform tasks autonomously in complex and often hazardous environments. This might be a scientist interacting with a robot to explore planetary surfaces, the ocean depths, etc. Alternatively, it could be a fireman working with a search-and-rescue robot to survey a disaster site. In both of these cases, the communication between the robot and the human is very limited (e.g., large delays in transmissions or limited bandwidth). As a result, the robot must be self-sufficient enough to perform a number of tasks in difficult environments where the human supervises the robot at the task level.

Much like an animal, the robot must apply its limited resources to address multiple concerns (performing tasks, self preservation, etc.) while faced with complex, unpredictable, and often dangerous situations. For instance, balancing emotion-inspired mechanisms associated with interest and fear could produce a focused yet safe searching behavior for a routine surveillance robot. For this application, one could take inspiration from the classic example of Lorenz (1950) regarding the exploratory behavior of a raven when investigating an object on the ground starting from a perch high up in a tree. For the robot just as for the raven, interest encourages exploration and sustains focus on the target, while recurring low levels of fear motivate it to retreat to safe distances, thereby keeping its exploration within safe bounds. Thus, an analogous exploratory pattern for a surveillance robot would consist of several iterative passes toward the target: on each pass, move closer to investigate the object in question and return to a launching point that is successively closer to the target.

Robot as Cyborg Extension

In the second paradigm, the robot is physically merged with the human to the extent that the person accepts it as an integral part of his or her body.

For example, the person would view the removal of his or her robotic leg as an amputation that leaves him or her only partially whole. Consider a robot that has an intimate physical connection with one's body, such as an exoskeleton for a soldier or a prosthetic for an amputee. Emotions play a critical role in connecting the mind with the body and vice versa. Note that the performance of one's physical body changes depending on whether one is relaxed or exhilarated. Although a robotic leg would not "have emotions" itself, it is important that it adapt its characteristics to match those of the rest of the body in accordance with the emotional state of the human to avoid imbalance. If the robotic extension were able to sense and recognize the person's emotional state (perhaps via physiological changes in the body), it could adapt its operating characteristics appropriately. At calmer times, the robotic extension could go into energy-conservation mode since power demands are lower. However, in high-stress situations, the robot could change its operation parameters to significantly augment the person's strength or speed.

Robot as Avatar

In the third paradigm, a person projects him- or herself through the robot to remotely communicate with others (the next best thing to being there). The robot provides a sense of physical presence to the person communicating through it and a sense of social presence to those interacting with it. Consider robot-mediated communication with others at a distance. Technology-mediated communication today is rather impoverished compared to face-to-face conversation, limiting our diverse communication channels to a select few. The advantage and appeal of a robot avatar is the ability to have a more fully embodied experience for the user and a greater physical and social presence to others (including touch, eye contact, physical proximity, movement and gesture within a shared space, etc.). The cognitive load and physical coordination required to directly control the many degrees of freedom for physical skills such as locomotion, object manipulation, gesture, and facial expression is overwhelming for the user. Hence, the robot would need to take high-level direction from the human and to be responsible for the performance of these physical and expressive abilities. To do so, the robot avatar would not need to have emotions itself, but it would need to be able to perceive and recognize the affective and linguistic intent of the user's message and to possess the ability to faithfully express and convey this message to others (see also Chapter 11, Nair et al.).

Robot as Partner

In the last paradigm, a person collaborates with the robot in a social manner as he or she would with a teammate (Grosz, 1996). Robots that interact with people as capable partners need to possess social and emotional intelligence so that they can respond appropriately. A robot that cares for an elderly person should be able to respond appropriately when the patient shows signs of distress or anxiety. It should be persuasive in ways that are sensitive, such as reminding the patient when to take medication, without being annoying or upsetting. It would need to know when to contact a health professional when necessary. Yet, so many current technologies (animated agents, computers, etc.) interact with us in a manner characteristic of socially or emotionally impaired people. In the best cases, they know what to do but often lack the social–emotional intelligence to do it in an appropriate manner. As a result, they frustrate us and we quickly dismiss them even though they can be useful. Given that many exciting applications for autonomous robots in the future place them in a long-term relationship with people, robot designers need to address these issues or people will not accept robots into their daily lives.

WHY SOCIAL/SOCIABLE ROBOTS?

In order to interact with others (whether it is a device, a robot, or even another person), it is essential to have a good conceptual model for how the other operates (Norman, 2001). With such a model, it is possible to explain and predict what the other is about to do, its reasons for doing it, and how to elicit a desired behavior from it. The design of a technological artifact, whether it is a robot, a computer, or a teapot, can help a person form this model by "projecting an image of its operation," through either visual cues or continual feedback (Norman, 2001). Hence, there is a very practical side to developing robots that can effectively convey and communicate their internal state to people for cooperative tasks, even when the style of interaction is not social.

For many autonomous robot applications, however, people will most likely use a social model to interact with robots in anthropomorphic terms. Humans are a profoundly social species. Our social–emotional intelligence is a useful and powerful means for understanding the behavior of, and for interacting with, some of the most complex entities in our world, people and other living creatures (Dennett, 1987). Faced with nonliving things of sufficient complexity (i.e., when the observable behavior is not easily understood in terms of physics and its underlying mechanisms), we often apply a social model to explain, understand, and predict their behavior, attributing

mental states (i.e., intents, beliefs, feelings, desires, etc.) to understand it (Reeves & Naas, 1996; Premack & Premack, 1995). Right or wrong, people rely on social models (or fluidly switch between using a social model and other mental models) to make complex behavior more familiar, understandable, and intuitive. We do this because it is enjoyable for us and often surprisingly quite useful (Dennett, 1987).

From a design perspective, the emotive system would implement much of the style and personality of a robot, encoding and conveying its attitudes and behavioral inclinations toward the events it encounters. Designing robots with personality may help provide people with a good mental model for them. According to Norman (2001), personality is a powerful design tool for helping people form a conceptual model that channels beliefs, behaviors, and intensions in a cohesive and consistent set of behaviors. The parameters of the personality must fall within recognizable human (or animal) norms, however; otherwise, the robot may appear mentally ill or completely alien. The robot's personality must be designed such that its behavior is understandable and predictable to people. Natural behavior can be a useful guide in this respect.

This raises an important question: to what extent does the robot's design support the social model? Simply stated, does applying a social mental model to understand and interact with the robot actually work? Many early examples of "social robots" (i.e., robot toys or entertainment robots) only projected the surface appearance of social and emotional intelligence. This may be acceptable for a sufficiently structured scenario (e.g., as theme park entertainment, etc.) where the environment and the audience's interaction with the robot are highly constrained. However, as the complexity of the environment and the interaction scenario increases, the social sophistication of the robot will clearly have to scale accordingly. Once the robot's behavior fails to support the social model a person has for it, the usefulness of the model breaks down. Ideally, the robot's observable behavior will continue to adhere to a person's social expectations of it during natural interactions in the full complexity of the human environment. We argue that it will not be possible to achieve this degree of scalability without tackling the (hard) problem of developing "deep" architectures for socially and emotionally intelligent robots (see Chapter 8, Sloman et al.).

DESIGNING SOCIABLE ROBOTS

The Sociable Robots Project develops expressive anthropomorphic robots to explore scientific questions and to address engineering challenges of building socially and emotionally intelligent robots. Their social and emotive qualities are integrated deep into the core of their design and serve not only

to "lubricate" the interface between themselves and their human interlocutors but also to promote survival, self-maintenance, learning, decision making, attention, and more (Breazeal, 2002a, 2003c,d). Hence, social and affective interactions with people are valued not just at the interface but at a pragmatic and functional level for the robot as well.

Humans, however, are the most socially and emotionally advanced of all species. As one might imagine, an autonomous anthropomorphic robot that could interpret, respond, and deliver human-style social and affective cues is quite a sophisticated machine. We have explored the simplest kind of human-style social interaction (guided and inspired by what occurs between a human infant with its caregiver) and have used this as a metaphor for building a sociable robot, called Kismet (shown in Fig. 10.1). The robot has been designed to support several social and emotive skills and mechanisms that are outlined in the rest of this chapter. Kismet is able to use these capabilities to enter into rich, flexible, dynamic interactions with people that are physical, affective, and social.

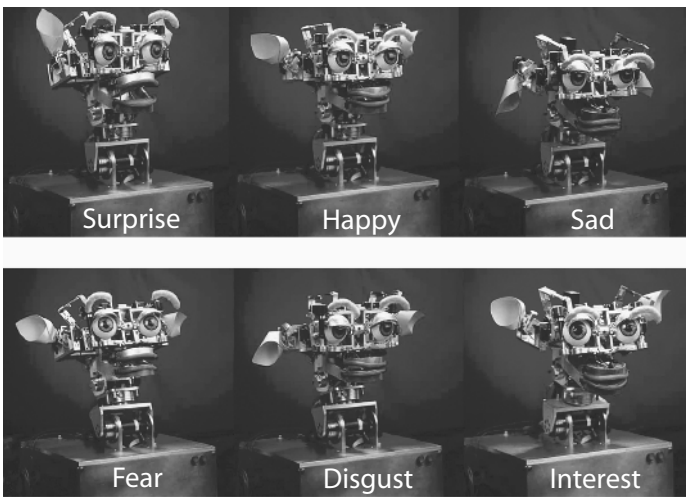


Figure 10.1. A sample of Kismet's facial expressions for basic emotions (see text). Kismet is about 1.5 times the size of an adult human head and has a total of 21 degrees of freedom. The robot perceives a variety of natural social cues from visual and auditory channels. Kismet has four cameras to visually perceive its environment: one behind each eye for postattentive visual processing (e.g., face detection), one between the eyes to provide a wide peripheral view (to track bright colors, skin tone, and movement), and one in the "nose" that is used in stereo with the peripheral-view camera to estimate the distance to targeted objects. A human wears a lavalier microphone to speak to the robot. (Images courtesy of Sam Ogden, © 2000.)

Expression of Affective State

Kismet can communicate its emotive state and other social cues to a human through facial expressions (Breazeal, 2003a), body posture, gaze direction (Breazeal, Fitzpatrick, & Scassellati, 2001), and quality of voice (Breazeal, 2003b). We do not have sufficient space to explain in detail how each of these is implemented, but they all contribute to the readability of Kismet's expression and ability to communicate its internal state to a human in a natural and intuitive way.

We have found that the scientific basis for how emotion correlates with facial or vocal expression is very useful in mapping Kismet's emotive states to its face actuators (Breazeal, 2003a) and to its articulation-based speech synthesizer (Breazeal, 2003b). In HRI studies (Breazeal, 2002b), we have found that these expressive cues are effective at regulating affective/intersubjective interactions (Trevarthen, 1979) and proto-dialogs (Tronick, Als, & Adamson, 1979) between the human and the robot that resemble their natural correlates during infant–caregiver exchanges.

With respect to communicating emotion through the face, psychologists of the componential theory of facial expression posit that these expressions have a systematic, coherent, and meaningful structure that can be mapped to affective dimensions that span the relationship between different emotions (Smith & Scott, 1997) (see Table 10.1). Some of the individual features of expression have inherent signal value. Raised brows, for instance, convey attentional activity for the expression of both fear and surprise. By considering the individual facial action components that contribute to the overall facial display, it is possible to infer much about the underlying properties of the emotion being expressed. This promotes a signaling system that is robust, flexible, and resilient. It allows for the mixing of these components to convey a wide range of affective messages, instead of being restricted to a fixed pattern for each emotion.

Inspired by this theory, Kismet's facial expressions are generated using an interpolation-based technique over a three-dimensional affect space (see Fig. 10.2). The three dimensions correspond to arousal (high/low), valence (good/bad), and stance (advance/withdraw), the same three attributes that are used to affectively assess the myriad environmental and internal factors that contribute to Kismet's overall affective state (see Affective Appraisal, below). There are nine basic postures that collectively span this space of emotive expressions.

The current affective state of the robot (as defined by the net values of arousal, valence, and stance) occupies a single point in this space at a time. As the robot's affective state changes, this point moves around this space and the robot's facial expression changes to mirror this. As positive valence increases, Kismet's lips turn upward, the mouth opens, and the eyebrows

Table 10.1. A Possible Mapping of Facial Movements to Affective Dimensions Proposed by Smith & Scott (1997)

Meaning	Eyebrow Frown	Raise Eyebrows	Raise Upper Eyelid	Raise Lower Eyelid	Raise Lip Corners	Open Mouth	Tighten Mouth	Raise Chin
Pleasantness	↓				↑	↑	↓	↓
Goal obstacle/discrepancy	↑							
Anticipated effort	↑							
Attentional activity		↑	↑					
Certainty		↓		↑		↑		
Novelty		↑	↑					
Personal agency/control		↓	↓			↓		

Up arrow indicates that the facial action is hypothesized to increase with increasing levels of the affective meaning dimension. Down arrow indicates that the facial action increases as the affective meaning dimension decreases. For instance, the lip corners turn upward as pleasantness increases and downward with increasing unpleasantness.

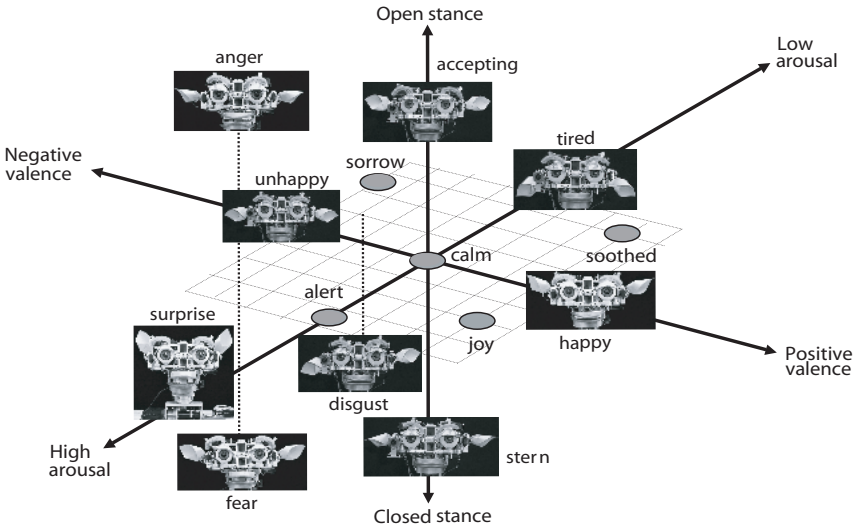


Figure 10.2. This diagram illustrates where the basis postures are located in Kismet's three-dimensional affect space. The dimensions correspond to arousal (high or low), valence (good or bad), and stance (advance or withdraw). This space is used to generate Kismet's facial expressions based on the robot's overall affective assessment of the current situation. A sampling of where specific emotion categories map onto this space is shown as well.

relax. However, as valence decreases, the brows furrow, the jaw closes, and the lips turn downward. Along the arousal dimension, the ears perk, the eyes widen, and the mouth opens as arousal increases. Along the stance dimension, the robot leans toward (increasing) or away from (decreasing) the stimulus. The expressions become more intense as the affect state moves to more extreme values in the affect space.

Kismet's face functions as a window by which a person can view the robot's underlying affective state. This transparency plays an important role in providing the human with the necessary feedback to understand and predict the robot's behavior when coupled with biologically inspired emotive responses.

ARCHITECTURAL OVERVIEW

Inspired by models of intelligence in natural systems, the design of our architecture features both a cognitive system and an emotive system (see Fig. 10.3). The two operate in parallel and are deeply intertwined to foster appropriate adaptive functioning of the robot in the environment as it interacts with people.

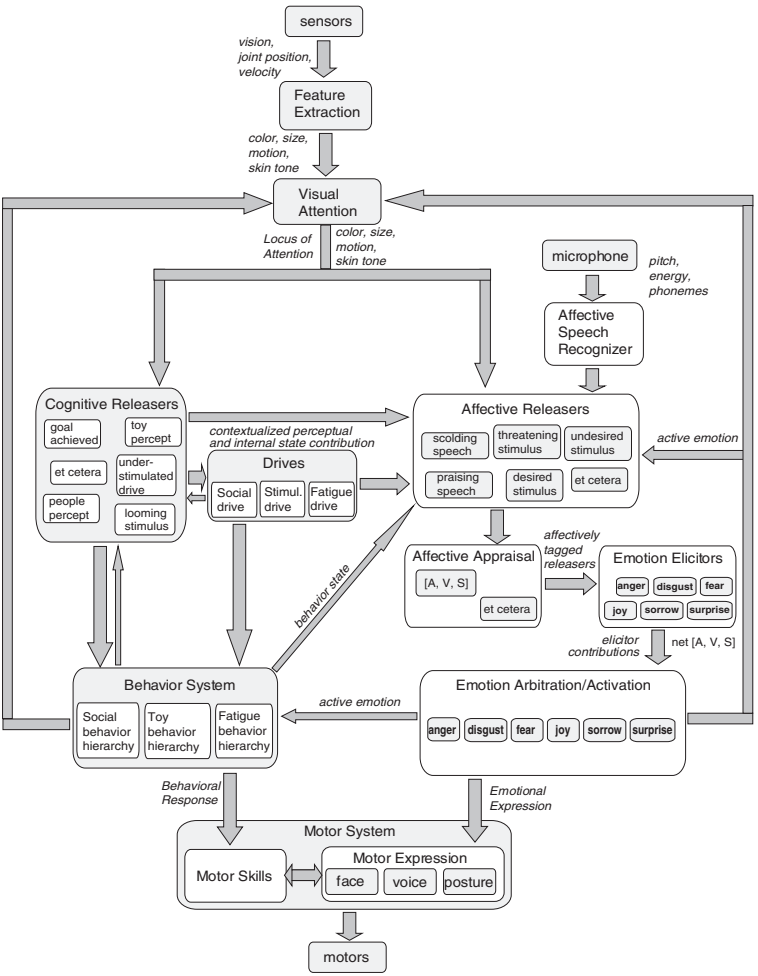


Figure 10.3. An architectural overview showing the tight integration of the cognitive system (light gray), the emotive system (white), and the motor system. The cognitive system is comprised of perception, attention, drives, and behavior systems. The emotive system is composed of affective releasers, appraisals, elicitors, and "gateway" processes that orchestrate emotive responses. A, arousal; V, valence; S, stance.

It is common for biologically inspired architectures to be constructed from a network of interacting elements (e.g., subsumption architecture [Brooks, 1986], neural networks [McCulloch & Pitts, 1943], or agent architectures [Minsky, 1986]). Ours is implemented as an agent architecture where each computational element is conceptualized as a specialist (Minsky 1986). Hence, each drive, behavior, perceptual releaser, motor,

and emotion-related process is modeled as a different type of specialist that is specifically tailored for its role in the overall system architecture.

Each specialist receives messages from those connected to its inputs, performs some sort of specific computation based on these messages, and then sends the results to those elements connected to its outputs. Its activation level, A, is computed by the following equation:

$$A = \left(\sum_{j=1}^{n} w_j i_j \right) + b$$

for integer values of inputs i_j, weights w_j, and bias b over the number of inputs n. The weights can be either positive or negative; a positive weight corresponds to an excitatory connection, and a negative weight corresponds to an inhibitory connection. Each process is responsible for computing its own activation level. The process is active when its activation level exceeds an activation threshold. When active, the process can send activation energy to other nodes to favor their activation. It may also perform some special computation, send output messages to connected processes, and/or express itself through motor acts by sending outputs to actuators. Hence, although the specialists differ in function, they all follow this basic activation scheme.

Units are connected to form networks of interacting processes that allow for more complex computation. This involves connecting the output(s) of one unit to the input(s) of another unit(s). When a unit is active, besides passing messages to the units connected to it, it can also pass some of its activation energy. This is called *spreading activation* and is a mechanism by which units can influence the activation or suppression of other units. This mechanism was originally conceptualized by Lorenz (1973) in his hydraulic model of behavior. Minsky (1986) uses a similar scheme in his ideas of memory formation using K-lines. Popular architectures of Brooks (1986) and Maes (1991) are similar in spirit. However, ours is heavily inspired by ethological models and, hence, is most similar to that of Blumberg, Todd, and Maes (1996).

Ethology, comparative psychology, and neuroscience have shown that observable behavior is influenced by internal factors (i.e., motivations, past experience, etc.) and by external factors (i.e., perception). This demands that different types of systems be able to communicate and influence each other despite their different functions and modes of computation. This has led ethologists such as McFarland and Bosser (1993) and Lorenz (1973) to propose that there must be a common currency to the perceptual, motivational, and behavioral systems. Furthermore, as the system becomes more complex, it is possible that some components conflict with others (e.g., competing for shared resources such as motor or perceptual abilities of the creature). In this

case, the computational processes must have some means for competing for expression.

Based upon the use of common currency, Kismet's architecture is implemented as a *value-based system*. This simply means that each process computes numeric values (in a common currency) from its inputs. These values are passed as messages (or activation energy) throughout the network, either within a system or between systems. Conceptually, the magnitude of the value represents the strength of the contribution in influencing other processes. Using a value-based approach has the effect of allowing influences to be graded in intensity, instead of simply being on or off. Other processes compute their relevance based on the incoming activation energies or messages and use their computed activation level to compete with others for exerting influence upon the other systems.

OVERVIEW OF THE COGNITIVE SYSTEM

The cognitive system is responsible for perceiving and interpreting events and for arbitrating among the robot's goal-achieving behaviors to address competing motivations. There are two kinds of motivation modeled in Kismet. The drives reside in the cognitive system and are modeled as homeostatic processes that represent the robot's "health" related goals. The emotive system also motivates behavior, as described below (see Overview of the Emotive System).

The computational subsystems and mechanisms that comprise the cognitive system work in concert to decide which behavior to activate, at what time, and for how long, to service the appropriate objective. Overall, the robot's behavior must exhibit an appropriate degree of relevance, persistence, flexibility, and robustness. To achieve this, we based the design of the cognitive system on ethological models of animal behavior (Gould, 1982). Below, we discuss how Kismet's emotion-inspired mechanisms further improve upon the basic decision-making functionality provided by the cognitive system (see Integrated Cognitive and Emotive Responses).

Perceptual Elicitors

Sensory inputs arising from the environment are sent to the perceptual system, where key features are extracted from the robot's sensors (cameras, microphones, etc.). These features are fed into an associated releaser process. Each releaser can be thought of as a simple perceptual elicitor of behavior that combines lower-level features into behaviorally significant perceptual categories.

For instance, the visual features of color, size, motion, and proximity are integrated to form a toy percept. Other releasers are designed to characterize important internal events, such as the urgency to tend to a particular motive. There are many different releasers designed for Kismet (too many to list here), each signaling a particular event or object of behavioral significance. If the input features are present and of sufficient intensity, the activation level of the releaser process rises above threshold, signifying that the conditions specified by that releaser hold. Given this, its output is passed to its corresponding behavior process in the behavior system, thereby preferentially contributing to that behavior's activation. Note that Kismet is not a stimulus–response system, given that internal factors (i.e., motivations as defined by drives and "emotions") also contribute to the robot's behavior activation.

Cognitive Motivations: Drives

Within the cognitive system, Kismet's drives implement autopoiesis-related processes for satisfying the robot's "health-related" and time-varying goals (Maturana & Varela, 1980). Analogous to the motivations of thirst, hunger, and fatigue for an animal, Kismet's drives motivate it to receive the right amount of the desired kind of stimulation in a timely manner. Kismet's drives correspond to a "need" to interact with people (the social drive), to be stimulated by toys (the stimulation drive), and to occasionally rest (the fatigue drive).

In living creatures, these autopoietic processes are innate and directly tied to physiology (see Chapter 3, Kelley). The design of each drive in Kismet is heavily inspired by ethological views of the analogous process in animals, where a change in intensity reflects the ongoing needs of the creature and the urgency for tending to them.

Each drive maintains a level of intensity within a bounded range, neither too much nor too little, as defined by a desired operational point and acceptable bounds of operation around that point (called the *homeostatic regime*). A drive remains in its homeostatic regime when it is encountering its satiatory stimulus of appropriate intensity. Given no satiatory stimulation, a drive will tend to increase in (positive) intensity. The degree to which each drive is satiated in a timely fashion contributes to the robot's overall measure of "well-being."

This information is also assigned affective value by the emotive system (described below, see Affective Appraisal). For example, negative value is assigned when the robot's needs are not being met properly, and positive value is assigned when they are. This is a rough analogy to the discussion of rewards and punishments associated with homeostatic need states in Chap-

ter 5 (Rolls). Hence, the affective contributions of the drives do not directly evoke emotive responses, but they do bias the robot's net affective state. In this sense, the drives contribute to the robot's "mood" over time, which makes the corresponding emotive responses easier to elicit.

Drives shape the internal agenda of the robot (in concert with perceptual and emotive factors) and play an important role in determining which behavior to next engage. To keep its activation level within the homeostatic regime, each drive can preferentially spread activation to behaviors at the top level of the behavior hierarchy that help to restore that drive (described in detail in the following section).

Behaviors, in turn, encode specific task-achieving goals that serve to maintain the robot's internal state (as defined by the state of the drives and "emotions"). To remain in balance (near the center of the spectrum), it is not sufficient that the satiatory stimulus merely be present; it must also be of good quality. For instance, in the absence of the satiatory stimulus (or if the intensity is too low), a drive increases in intensity to the positive end of the spectrum and preferentially biases the activation of those behaviors that seek out that stimulus. In addition, the affective contribution of the drive (negative valence and low arousal) contributes to a net affective state that makes it easier for the sorrow emotive response to become active. Sorrow represents a different strategy to help the robot come into contact with a desired stimulus by signaling to people that it needs attention.

Alternatively, if the satiatory stimulus is too intense (e.g., a visual stimulus moving too fast or too close to the robot's face), a drive tends toward the extreme negative end of the spectrum. In this circumstance, the drive biases the activation of avoidance behaviors to limit the robot's exposure to the intense stimulus. Also, the affective contribution of the drive (negative valence, high arousal) contributes to a net affective state that makes it easier for Kismet's fear response to become active. Once active, the fearful expression on Kismet's face signals people to back off a bit.

Hence, the drives work in concert with behaviors and contribute to an affective state that helps Kismet keep its level of interaction with the world and people in balance, neither too much nor too little.

Behavior Arbitration as Decision Making

Within the behavior system, the behavior processes are organized into loosely layered, heterogeneous hierarchies of behavior groups (see Fig. 10.4), much in the spirit of those ethological models proposed by Tinbergen (1951) and Lorenz (1973). Implicit in this model is that a decision is being made among several alternatives at every level of the hierarchy, of which one is chosen.

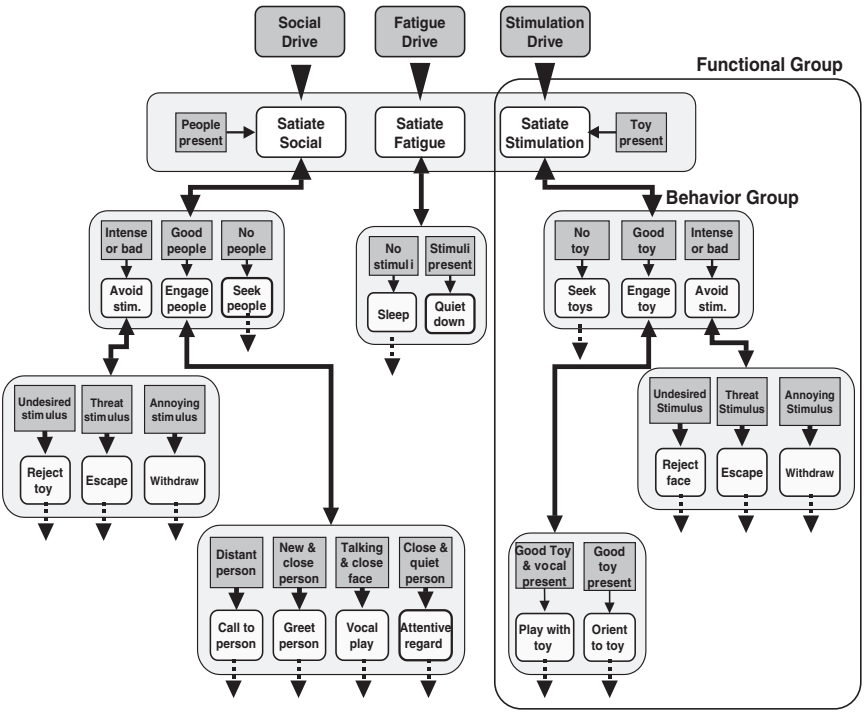


Figure 10.4. Schematic of Kismet's behavior hierarchy that resides within the behavior system of the cognitive–affective architecture. The affective factors for each behavior (both inputs and outputs) are not shown. Dashed arrows represent connections to the motor system. Input from the drives feeds into the top level of the hierarchy, as shown by the dark gray boxes with rounded edges. Functional groups are represented as the major branches of the hierarchy. For instance, the functional group for satiating Kismet's drive to interact with toys is highlighted. Behavior groups are shown as light gray boxes with rounded edges, containing competing behaviors (white boxes with rounded edges) and their perceptual elicitors (dark gray boxes).

Functional Groups

Decisions are very general at the top of the hierarchy (which drive to satiate) and become increasingly more specific as one moves down the hierarchy. At the topmost level, behaviors are organized into competing functional groups (i.e., the primary branches of the hierarchy, of which there are three in Kismet). Each functional group is responsible for maintaining one of the three homeostatic functions, and only one functional group can be active at a time. This property is inspired by animal behavior, where an animal en-

gages in one class of activities at a time: foraging for food, sleeping, defending territory, mating, etc. Thus, the influence of the drives is strongest at the top level of the hierarchy, biasing which functional group should be active.

Behavior Groups

Each functional group consists of an organized hierarchy of behavior groups that are akin to Tinbergen's (1951) behavioral centers. At each level within a functional group hierarchy, each behavior group represents a competing strategy (a collection of behaviors) for satisfying the goal of its parent behavior.

Behaviors

Each behavior within a behavior group is viewed as a task-achieving entity, the particular goal of which contributes to the strategy of its behavior group. Each behavior process within a group competes with the others in a winner-take-all fashion for expression based on its measure of relevance to the current situation. A behavior's measure of relevance takes into account several factors, including the perceived environment through its releaser inputs, internal motives through its drive and emotion inputs, and internally computed progress measures, such as level of interest (how long this behavior has been active) and level of frustration (a measure of progress toward its goal over time). When active, a behavior coordinates sensorimotor patterns to achieve a particular task, such as search behaviors, approach behaviors, avoidance behaviors, and interaction behaviors. The perceptual, homeostatic, and affective factors that contribute to behavioral relevance allow the robot to act in a manner that is persistent (e.g., trying new strategies to attain a blocked goal) but also suitably opportunistic (described in more detail below, see Integrated Cognitive and Emotive Responses).

Therefore, the observed behavior of the robot is the result of competition at the functional, strategic, and task levels. At the behavioral level, the functional groups compete to determine which drive is to be met (for Kismet, this corresponds to socializing, playing, or sleeping). At the strategy level, behavior groups of the winning functional group compete for expression. For instance, two of the behavior groups at the second level contain several strategies for acquiring a desired stimulus of an appropriate

intensity: seek or acquire the desired stimulus if it is too weak or not present, avoid or reduce the intensity of the stimulus if it is too overwhelming, or engage the stimulus if it is of appropriate intensity. Finally, on the task level, the behaviors of the winning group compete for expression to determine which subgoal the robot pursues (i.e., ways of acquiring the desired stimulus, ways of reducing its intensity if it is overwhelming, and ways of engaging the stimulus when it is appropriate). As one moves down in depth, the behaviors more finely tune the relation between the robot and its environment, in particular the relation between the robot and the human (Breazeal, 2002a).

OVERVIEW OF THE EMOTIVE SYSTEM

The emotive system is responsible for perceiving and recognizing internal and external events with affective value, assessing and signaling this value to other systems, regulating and biasing the cognitive system to promote appropriate and flexible decision making, and communicating the robot's internal state. Kismet communicates its emotive state in a transparent and familiar manner through facial expression, body posture, and tone of voice. This makes the robot's behavior more predictable and understandable to the person who is interacting with it. These expressive behaviors allow Kismet to socially regulate people's behavior toward it in a natural way that is beneficial to the robot.

Thus, in concert with the robot's cognitive system, the emotive system is designed to be a flexible and complementary system that mediates between environmental, social, and internal events to elicit an adaptive behavioral response that serves either social or self-maintenance functions. In humans and other animals, each specific emotion motivates and coordinates cognitive systems and patterns of behavior responses to facilitate development, adaptation, and coping in a particular way. The remainder of this section outlines how Kismet's emotive responses are implemented. Each emotive response consists of the following:

- A precipitating event (see Affective Releasers)
- An affective appraisal of that event (see Affective Appraisal)
- A characteristic display that can be expressed through facial expression, vocal quality, or body posture (see above, Expression of Affective State)
- Modulation of the cognitive and motor systems to motivate a behavioral response (see Emotion Elicitors and Arbitration, followed by Integrated Cognitive and Emotive Responses)

According to Rolls (see Chapter 5), emotions can be pragmatically defined as states elicited by rewards and punishments, where a reward is something for which an animal will work and a punishment is something it will work to escape or avoid. Kismet's affective appraisal process involves assessing whether something is rewarding or punishing and tagging it with a value that reflects its expected benefit or harm (see also Chapter 7, Ortony et al.). The emotive system combines the myriad affectively tagged inputs to compute the net affective state of the robot. Similar to Rolls, the affective tags serve as the common currency for the inputs to the response-selection mechanism. Thus, they are used to determine the most relevant emotive response for the given situation. Once a particular emotive response is active, Kismet engages in a process of behavioral homeostasis, where the active emotive response maintains behavioral activity in its particular manner through external and internal feedback until the correct relation of robot to environment is established (Plutchik, 1991).

Kismet's desired internal and external relationship is comprised of two factors: whether the robot's homeostatic needs are being met in a timely manner and whether its net affective state corresponds to a mildly positive, aroused, and open state. When Kismet's internal state diverges from this desired internal relationship, the robot will work to restore the balance—to acquire desired stimuli, to avoid undesired stimuli, and to escape dangerous stimuli. Each emotive response carries this out in a distinct fashion by interacting with the cognitive system to evoke a characteristic behavioral pattern and to socially cue others as to whether the interaction is appropriate or not (and how they might respond to correct the problem). (A detailed description of the implementation of each emotive response can be found in Breazeal, 2002a.)

Functions of Basic Emotions

The organization and operation of Kismet's emotive system is strongly inspired by various theories of basic emotions in humans and other animals (Ekman, 1992). These few select emotions are endowed by evolution because of their proven ability to facilitate adaptive responses that promote a creature's daily survival in a complex and often hostile environment. Basic emotions are "independent" emotions because their emergence does not require or reduce to cognitive processes (Ackerman, Abe, & Izard, 1998). As shown in Table 10.2, a number of basic emotive responses have been implemented in Kismet. This table is derived from the cross-species and social functions hypothesized by Izard and Ackerman (2000) and Plutchik (1991).

Table 10.2. Summary of the Antecedents and Behavioral Responses that Comprise Kismet's Emotive Responses

Antecedent Conditions	Emotion	Behavior	Function
Delay, difficulty in achieving goal of active behavior	Anger, frustration	Display agitation, energize	Show displeasure to modify human's behavior; try new behavior to surmount blocked goal
Presence of an undesired stimulus	Disgust	Withdraw	Signal rejection of presented stimulus
Presence of a threatening or overwhelming stimulus	Fear, distress	Display fear, escape	Move away from a potentially dangerous stimulus
Prolonged presence of a desired stimulus	Calm	Engage	Continued interaction with a desired stimulus
Success in achieving goal of active behavior or praise	Joy	Display pleasure	Reallocate resources to the next relevant behavior
Prolonged absence of a desired stimulus or scolding	Sorrow	Display sorrow	Evoke sympathy and attention from human
A sudden, close stimulus	Surprise	Startle response	Alert
Appearance of a desired stimulus	Interest	Orient, explore	Attend to new, salient object, engage
Absence of stimulus	Boredom	Seek	Explore environment for desired stimulus

Antecedents are the eliciting perceptual conditions for each emotion. The behavior column denotes the observable response that becomes active with the emotion. For some, this is simply a facial expression. For others, it is a behavior such as escape. The column to the right describes the function each emotive response serves for Kismet.

Anger

In living systems, anger serves to mobilize and sustain energy and vigorous motor activity at high levels (Tomkins, 1963). It is often elicited when progress toward a goal is hindered or blocked. Similarly, in Kismet, a frustrated state (increasing in intensity to anger) arises when progress toward the current goal is slow. This mobilizes the robot to try alternate strategies.

Disgust

Tomkins (1963) describes disgust as a reaction to unwanted intimacy with a repellent entity. Generally speaking, disgust is manifested as a distancing from some object, event, or situation and can be characterized as rejection

(Izard, 1997). It is in this sense that disgust is incorporated into Kismet's emotive repertoire. Kismet's disgust response signals rejection of an unwanted stimulus.

Fear

The unique function of fear is to motivate avoidance or escape from a dangerous situation. For Kismet, the fear response protects it from possible harm when faced with a threatening stimulus that could cause damage (e.g., large stimuli moving fast and close to the robot's face). Kismet's fearful expression is a communicative cue that signals to a person that he or she should back off a bit (Breazeal & Scassellati, 2000). If they persist, the robot will evoke a protective escape response (e.g., close its eyes and turn its head away from the offending stimulus).

Joy

The emotion of joy is believed to heighten openness to experience. It often arises upon the success of achieving a goal or the pleasure of mastery, exhibited even by very young children (Meltzoff & Moore, 1997). In humans, openness in social situations contributes to affiliative behavior and the strengthening of social bonds (see also Chapter 9, Arkin). The expression of joy operates as a universally recognizable signal of readiness for friendly interaction. For Kismet, it serves a social function, to encourage people to interact with it. It also arises when the robot has achieved a pursued goal, accompanied by a reallocation of cognitive/behavioral resources to the next relevant task.

Sorrow

Izard and Ackerman (2000) argues that sadness is unique in its capacity to slow the cognitive and motor systems. Tomkins (1963) suggests that slowing down enables one to reflect upon a disappointing performance and gain a new perspective that will help improve future performance. Sadness can also strengthen social bonds. The expression of sorrow communicates to others that one is in trouble and increases the likelihood that the others will feel sympathy and lend assistance (Moore, Underwood, & Rosenhan, 1984). Similarly, Kismet's expression of sorrow serves a communicative function that encourages people to pay attention to it and to try to cheer

it up (Breazeal, 2002b). In HRI studies, we have found that Kismet's expression elicits sympathy responses in people (Breazeal, 2002b).

Surprise

Children show surprise when there are violations of expected events or as a response to discovery, such as an "Aha!" experience. Hence, cognitive processes play an important role in the emergence of this early emotion. Given that simple cognitions elicit surprise, some emotion theorists do not consider surprise to be a basic emotion, even though it appears early in infancy (around 6 months of age). For Kismet, surprise is a startle response elicited by a sudden and unexpected event, such as a quickly looming stimulus.

Interest

In humans and other animals, interest motivates exploration, learning, and creativity. It mobilizes the creature for engagement and interaction. It serves as a mechanism of selective attention that keeps the creature's attention focused on a particular object, person, or situation and away from other distractions that impinge upon its senses. For Kismet, it serves a similar function with respect to focusing attention and motivating exploration and interaction.

Boredom

In Kismet, boredom is treated as a basic emotion that arises when the robot is not stimulated for a while. Over the long term, this prolonged absence will elicit sorrow. In the shorter term, boredom motivates searching behaviors similar to interest; however, its function is to come into contact with a desired stimulus, rather than to engage one that is already present.

Affective Releasers

The affective releasers assess the value of perceptual inputs arising from the environment. They are similar to the perceptual releasers of the cognitive system, but rather than being only a perceptual interpretation of stimuli into objects and events, they are also cognitively appraised in relation to the motivational state of the robot and its current goals. Beyond simple percep-

tual features, the affective releasers go through a more detailed cognitive appraisal to judge their expected benefit to the robot: the quality of the stimulus (e.g., the intensity is too low, too high, or just right) or whether it is desired or not (e.g., it relates to the active goals or motivations). For instance, if the stimulation drive is being tended to and a nearby toy is moving neither too fast nor too close to the robot, then the desired toy releaser is active. However, if the social drive is being tended to instead, then the undesired toy releaser is active. If the toy has an aggressive motion (i.e., too close and moving too fast), then the threatening toy releaser is active. This evaluation is converted into an activation level for that affective releaser. If the activation level is above threshold, then its output is passed to the affective appraisal stage, where it can influence the net affective state and emotive response of the robot.

Recognition of Communicated Affect

Objects with which Kismet interacts can have affective value, such as a toy that is moving in a threatening manner. However, people can communicate affect directly to Kismet through tone of voice. Developmental psycholinguists can tell us much about how preverbal infants achieve this. Based on a series of cross-linguistic analyses, there appear to be at least four different pitch contours that prelinguistic infants can recognize affectively (approval, prohibition, comfort, and attention), each associated with a different emotional state (Fernald, 1989). Characteristic prosody curves for each are shown in Figure 10.5.

Inspired by these theories, we have implemented a recognizer for distinguishing these four distinct prosodic patterns from Kismet-directed speech. The implemented classifier consists of several miniclassifiers, which execute in stages (see Fig. 10.6). A detailed presentation of the recognizer and its performance assessment can be found in Breazeal and Aryananda (2002).

Based on our recordings, the preprocessed pitch contours from the training set resemble Fernald's prototypical prosodic contours for approval, attention, prohibition, comfort/soothing, and neutral. Hence, we used Fernald's insights to select those features that would prove useful in distinguishing these five classes. For the first classifier stage, global pitch and energy features (i.e., pitch mean and energy variance) partitioned the samples into useful intermediate classes (see Fig. 10.7).

For instance, the prohibition samples are clustered in the low-pitch mean and high-energy variance region. The approval and attention classes form a cluster at the high-pitch mean and high-energy variance region. The soothing samples are clustered in the low-pitch mean and low-energy variance

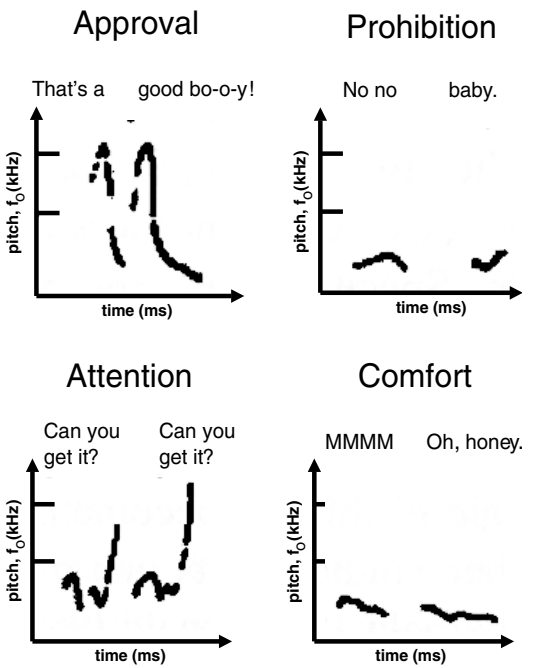


Figure 10.5. Fernald's prototypical contours for approval, prohibition, attention, and soothing, which are matched to saliency measures hardwired into an infant's auditory processing system (Fernald, 1989). F_o, pitch.

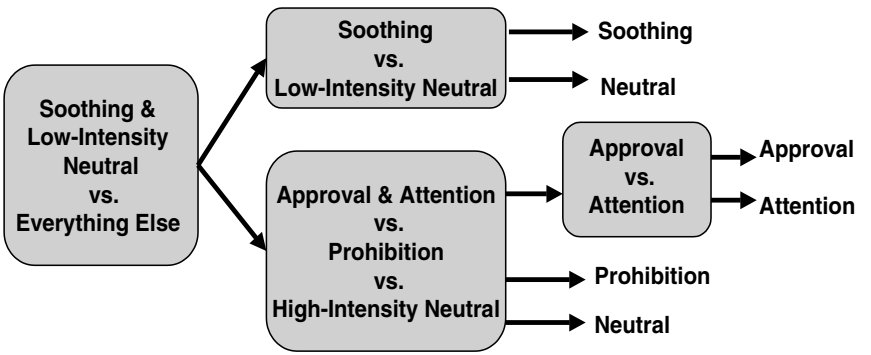


Figure 10.6. The spoken affective intent recognizer. Each stage is a mini-classifier that uses acoustic features identified by Fernald (1989) to recognize each of the four affective intents.

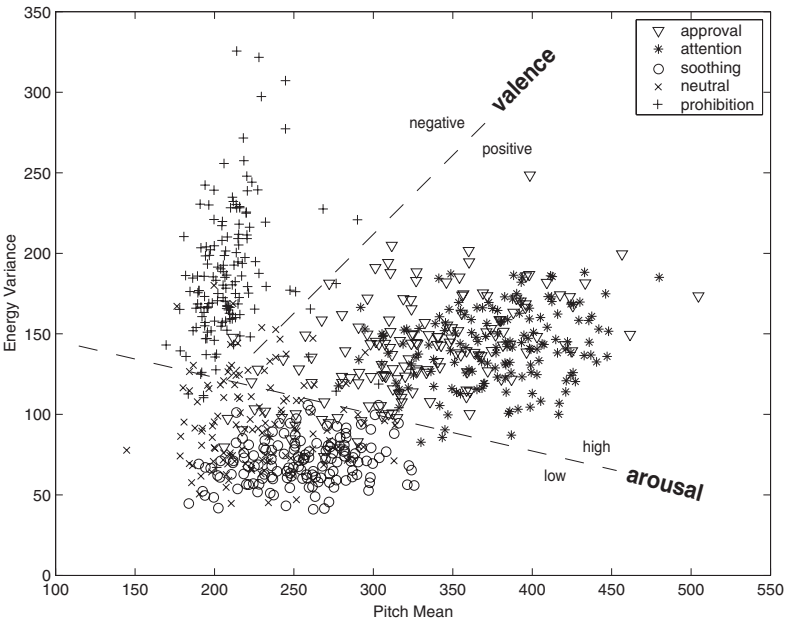


Figure 10.7. Feature space of all five classes with respect to energy variance and pitch mean. There are three distinguishable clusters (roughly partitioned by arousal and valence) for prohibition, for soothing and neutral, and for approval and attention.

region. Finally, the neutral samples have low-pitch mean but are divided into two regions in terms of their energy variance values. The structure of each of the mini-classifiers follows logically from these observations. Table 10.3 shows the resulting classification performance using a Gaussian mixture model, updated with the Expectation-Maximization algorithm, to represent the distribution of data. The output of each affective intent classifier is treated as an affective releaser and sent with the others to the affective appraisal phase.

Affective Appraisal

In Kismet's implementation, there is an explicit assessment phase for each active releaser, of which there are a number of factors that contribute to the assessment made. The assessment consists of labeling the releaser with affective tags, a mechanism inspired by Damasio's (1994) somatic marker hypothesis, where incoming perceptual, behavioral, or motivational information is "tagged" with affective information. The tagged value reflects whether the releaser is expected to be rewarding or punishing to the robot.

Table 10.3. Overall Classification Performance Evaluated Using a New Test Set of 371 Utterances from Five Adult Female Speakers Ranging in Age from 23 to 54 Years

Category	Test Size	Approval	Attention	Prohibition	Comfort	Neutral	% Correct
Approval	84	64	15	0	5	0	76.2
Attention	77	21	55	0	0	1	74.3
Prohibition	80	0	1	78	0	1	97.5
Comfort	68	0	0	0	55	13	80.9
Neutral	62	3	4	0	3	52	83.9
All	371						81.9

For example, there are three classes of tags used by Kismet to affectively characterize a given releaser. Each tag has an associated intensity that scales its contribution to the overall affective state. The arousal tag, *A*, specifies how arousing this factor is to the emotional system. It very roughly corresponds to the activity of the autonomic nervous system. Positive values correspond to a high arousal stimulus, whereas negative values correspond to a low arousal stimulus. The valence tag, *V*, specifies how favorable or unfavorable this percept is to the robot. Positive values correspond to a beneficial stimulus, whereas negative values correspond to a stimulus that is not beneficial. The stance tag, *S*, specifies how approachable the percept is. Positive values correspond to advance, whereas negative values correspond to retreat.

There are three factors that contribute to an appraisal of an active releaser. The first is the intensity of the stimulus, which generally maps to arousal. Threatening or very intense stimuli are tagged with high arousal. Absent or low intensity stimuli are tagged with low arousal. The second is the relevance of the stimulus to whether it addresses the current goals of the robot. This influences the valence and stance values. Stimuli that are relevant are desirable. They are tagged with positive valence and approaching stance. Stimuli that are not relevant are undesirable. They are tagged with negative arousal and withdrawing stance. The third factor is intrinsic pleasantness. Some stimuli are hardwired to influence the robot's affective state in a specific manner. For instance, praising speech is tagged with positive valence and slightly high arousal, whereas scolding speech is tagged with negative valence and low arousal, which tends to elicit a sorrowful response. In Kismet, there is a fixed mapping from each of these factors to how much they contribute to arousal, valence, or stance.

In addition to the perceptual contribution of the releasers, other internal factors can influence the robot's emotive state. For instance, the drives contribute according to how well they are being satiated. The homeostatic

regime is marked with positive valence and balanced arousal, contributing to a contented affective state. The understimulated regime (large positive values) is marked with negative valence and low arousal, contributing to a bored affective state that can eventually decline to sorrow. The overstimulated regime (large negative values) is marked with negative valence and high arousal, contributing to an affective state of distress. Another factor is progress toward achieving the desired goal of the active behavior. Success in achieving a goal promotes joy and is tagged with positive valence. Prolonged delay in achieving a goal results in frustration and is tagged with negative valence and withdrawn stance. It is also possible for the active emotion to either contribute to or inhibit the activation of other emotions, making it difficult for a creature to be both happy and angry simultaneously, for instance.

Because there are potentially many different kinds of factor that modulate the robot's affective state (e.g., behaviors, motivations, perceptions), this tagging process converts the myriad factors into a common currency that can be combined to determine the net affective state. For Kismet, the *A*, *V*, *S* trio is the currency the emotive system uses to determine which response should be active. In the current implementation, the values of the affective tags for the releasers are specified by the designer. These may be fixed constants or linearly varying quantities.

Emotion Elicitors and Arbitration

All somatically marked inputs (e.g., releasers, the state of each drive, etc.) are passed to the emotion elicitors. There is an elicitor associated with each basic emotion "gateway" process (e.g., anger, fear, disgust). The elicitor determines the relevance of its emotive response based on the myriad factors contributing to it. In a living creature, this might include neural factors, sensorimotor factors, motivational factors, and cognitive factors (Izard, 1993). Each elicitor computes the relevance of its affiliated emotion process and contributes to its activation. Each elicitor can thus be modeled as a process that computes its activation energy, $E_{emot}(i)$, for emotion, i, according to the following function:

$$E_{emot}(i) = R_{emot}(i) + Dr_{emot}(i) + Em_{emot}^{excite}(i) - Em_{emot}^{inhibit}(i) + Bh_{emot}(i)$$

where $R_{emot}(i)$ is the weighted contribution of the active releasers, $Dr_{emot}(i)$ is the weighted contribution of the active drive, $Em_{emot}^{excite}(i)$ is the weighted contribution of the other active emotions that excite this process, $Em_{emot}^{inhibit}(i)$ is the weighted contribution of the other active emotions that inhibit this process, and $Bh_{emot}(i)$ is the weighted contribution of the behavioral progress toward the current goal.

Each emotion gateway process competes for control in a winner-take-all manner based on its activation level. Although these processes are always active, their intensity must exceed a threshold level before they are expressed externally. The activation of each process is computed by the following equation:

$$A_{emot}(i) = \sum_{i} \left[E_{emot}(i) + b_{emot}(i) + p_{emot}(i) \right] - \delta_t(i)$$

where $E_{emot}(i)$ is the activation level computed by the affiliated emotive elicitor process described above, $b_{emot}(i)$ is a constant offset that can be used to make the emotion processes easier or harder to activate in relation to the activation threshold, and $p_{emot}(i)$ adds a level of persistence to the active emotion. This introduces a form of inertia so that different emotion processes do not rapidly switch back and forth. Finally, $\delta_t(i)$ is a decay term that restores an emotion to its bias value once it becomes active.

When active, each emotion acts as a gateway, such that when "open" it can spread activation to a number of different cognitive systems (i.e., behavior, attention, expression). As a result, the emotive state of the robot is distributed throughout the overall architecture; it does not reside in the gateway process itself.

Each emotion gateway process plays a distinct regulatory role, modulating the cognitive and expressive systems in a characteristic manner. In a process of behavioral homeostasis, the emotive response maintains activity through external and internal feedback until the correct relation of robot to environment is established (Plutchik, 1991). Concurrently, the affective state of the robot, as specified by the net A, V, S of the active process is sent to the expressive components of the motor system, causing a distinct facial expression, vocal quality, and body posture to arise. (A detailed description of the implementation of each emotive response can be found in Breazeal, 2002a.)

INTEGRATED COGNITIVE AND EMOTIVE RESPONSES

In this section, we illustrate how Kismet's emotive system works in concert with its cognitive system to address its competing goals and motives given its limited resources and the ever-changing demands of interacting with people (Breazeal, 2002a). The emotive system achieves this by assessing and signaling the value of immediate events in order to appropriately regulate and bias the cognitive system to help focus attention, prioritize goals, and pursue the current goal with an appropriate degree of persistence and opportunism. The emotive responses protect the robot from intense interactions that may be potentially harmful and help the robot to sustain interactions

that are beneficial. The emotive system improves the performance of the robot over that provided by the cognitive system alone.

Communicative Expression

Each emotive response entry of Table 10.2 is composed of a goal-achieving behavioral component and an accompanying expressive display. For some of the emotive responses, the expressive display addresses both aspects when it serves a communicative function that is designed to elicit a desired behavioral response from the human that satisfies the robot's goal.

Kismet's expressive abilities successfully serve its goals when interacting with a person for two main reasons. First, we have found that people[1] enjoy playing with Kismet and want to sustain a pleasurable interaction with it (Breazeal & Scassellati, 2000; Breazeal, 2002a, 2003a). People find the robot to be lively and to have an appealing personality and convincing social presence. This is a result of the way Kismet's emotive system is designed to interact with its cognitive system (as argued above, see Why Social/Sociable Robots?). Thus, both Kismet and the person have the shared goal of establishing and maintaining a beneficial interaction. The interaction is beneficial to the human if it is enjoyable, and it is beneficial to the robot if its motivations and goals are satisfied. Second, Kismet's expressive behavior is effective because it is readily understandable and predictable to the person who interacts with it. This follows from the fact that Kismet's emotive responses are modeled after basic emotions that are universally understood by people (Ekman, 1992). As a result, people readily infer how they must adapt their behavior to obtain a desired response from Kismet—to keep it happy and interested and to avoid causing it distress.

For instance, Kismet exhibits sorrow upon the prolonged absence of a desired stimulus. This may occur if the robot has not been engaged with a toy for a long time. The sorrowful expression is intended to elicit attentive acts from the human analogous to how an infant's cries elicit nurturing responses from its caregiver. Kismet uses other expressive displays, such as fearful expression to encourage people to slow down or back off a bit if they are crowding its cameras or moving too fast for it to perceive them. This allows the robot to tune the human's behavior so that it is appropriate for it. When the interaction is beneficial to Kismet, the robot conveys a state of interest and joy that encourages people to sustain the interaction. In a number of HRI studies with Kismet, we have found this to be quite effective as people find pleasure in cheering up the robot and keeping it engaged without being instructed to do so (Breazeal, 2003a).

Biasing Attention

Kismet's level of interest improves the robot's attention, biasing it toward desired stimuli (e.g., those relevant to the current goal) and away from irrelevant stimuli. For instance, Kismet's exploratory responses include visual searching for a desired stimulus and/or maintaining visual engagement of a relevant stimulus. Kismet's visual attention system directs the robot's gaze to the most salient object in its field of view, where the overall salience measure is a combination of the object's raw perceptual salience (e.g. size, motion, color) and its relevance to the current goal. It is important to note that Kismet's level of interest biases it to focus its attention on a goal-relevant stimulus that is beneficial, even when that object may have less perceptual salience over another "flashy" yet less goal-relevant stimulus. Without the influence of interest on Kismet's attention, the robot would end up looking at the flashy stimulus even if it has less behavioral benefit to the robot.

In addition, Kismet's disgust response allows it to reject and look away from an undesired stimulus. This directs the robot's gaze to another point in the visual field, where it might find a more desirable object to attend. It also provides an expressive cue that tells the human that the robot wants to look at something else. The person often responds by trying to engage Kismet with a different toy, for example. This increases the robot's chances that it might be presented with a stimulus that is more appropriate to its goal. We have found that people are quick to determine which stimulus the robot is after and readily present it to Kismet (Breazeal, 2002b, 2003a; Breazeal & Scassellati, 2000). This allows the robot to cooperate with the human to obtain a desired stimulus faster than it would if it had to discover one on its own.

Goal Prioritization, Persistence, and Opportunism

Emotion-inspired processes play an important role in helping Kismet to prioritize goals and to decide when to switch among them. They contribute to this process through a variety of mechanisms to make Kismet's goal-pursuing behavior flexible, opportunistic, and appropriately persistent.

Emotive Influences

For instance, Kismet's fear response allows it to quickly switch from engagement behaviors to avoidance behaviors once an interaction becomes too intense or turns potentially harmful. This is an example of a rapid repriori-

tization of goals. The fear response accomplishes this by effectively "hijacking" the behavior and motor systems to rapidly respond to the situation. For instance, the fear response may evoke Kismet's escape behavior, causing the robot to close its eyes and turn its head away from the offending stimulus.

Affective Drive Influences

In addition, affective signals arising from the drives bias which behaviors become active to satiate a particular motive. These affective influences contribute to activating behaviors that are the most relevant to the robot's "health"-related needs. When the drives are reasonably well satiated, the perceptual contributions play the dominant role in determining which goals to pursue. Hence, the presence of a person will tend to elicit social behaviors and the presence of a toy will tend to elicit toy-directed behaviors. As a result, Kismet's behavior appears strongly opportunistic, taking advantage of whatever stimulus presents itself.

However, if a particular drive is not satiated for a while, its influence on behavior selection will grow in intensity. When this occurs, the robot becomes less opportunistic and grows more persistent about pursing those goals that are relevant to that particular drive. For instance, the robot's behavior becomes more "finicky" as it grows more prone to give a disgust response to stimuli that do not satiate that specific drive. The robot will also start to exhibit a stronger-looking preference to stimuli that satiate that drive over those that do not. These aspects of persistent behavior continue until the drive is reasonably satiated again.

Affective Behavior Influences

Another class of affective responses influences arbitration between competing behavioral strategies to achieve the same goal. Delayed progress of a particular behavior results in a state of growing frustration, reflected by a stern expression on the robot's face. As Kismet grows more frustrated, it lowers the activation level of the active behavior within the behavior group. This makes it more likely to switch to another behavior within the same group, which could have a greater chance of achieving the current goal.

For instance, if Kismet's goal is to socialize with a person, it will try to get a person to interact with it in a suitable manner (e.g., arousing but not too aggressive). If the perceptual system detects the presence of a person but the person is ignoring Kismet, the robot will engage in behaviors to attract the person's attention. For instance, the robot's initial strategy might

be to vocalize to the person to get his or her attention. If this strategy fails over a few attempts, the level of frustration associated with this behavior increases as its activation level decreases. This gives other competing behaviors within the same behavior group a chance to win the competition and become active instead. For instance, the next active behavior strategy might be one where Kismet leans forward and wiggles its ears in an attention-grabbing display. If this also fails, the prolonged absence of social interaction will eventually elicit sorrow, which encourages sympathy responses from people, a third strategy to get attention from people to satiate the social drive.

CONCLUSION

In this chapter, we have explored the benefits that emotive and cognitive aspects bring to the design of autonomous robots that operate in complex and uncertain environments and perform in cooperation with people. Our examples highlight how Kismet's emotive system works intimately with its cognitive system to improve its overall performance. Although the cognitive system is designed with a variety of mechanisms to support attention, behavior arbitration, and motor expression (see "Overview of the Cognitive System"), these cognitive mechanisms are enhanced by emotion-inspired mechanisms that further improve Kismet's communicative effectiveness, its ability to focus its attention on relevant stimuli despite distractions, and its ability to prioritize goals to promote flexible behavior that is suitably opportunistic when it can afford to be persistent when it needs to be.

What about the external expression of emotion? Even if one were to accept the internal regulatory and biasing benefits of emotion-inspired mechanisms, do these need to be accompanied by social–emotive expression? Granted, it is certainly possible to use other information-based displays to reveal the internal state of robots: flashing lights, laser pointers, graphics, etc. However, people would have to learn how to decipher such displays to understand what they mean. Furthermore, information-based displays fail to leverage from the socio-affective impact and intuitive meaning that biological signals have for people.

Kismet's emotive system implements the style and personality of the robot, encoding and conveying its attitudes and behavioral inclinations toward the events it encounters. People constantly observe Kismet's behavior and its manner of expression to infer its internal state as they interact with it. They use these expressive cues as feedback to infer whether the robot understood them, its attitude about the interaction, whether they are engaging the robot appropriately, whether the robot is responding appropriately to them, etc. This helps the person form a useful mental model for the

robot, making the robot's behavior more understandable and predictable. As a result, the person can respond appropriately to suit the robot's needs and shape the interaction as he or she desires. It also makes the interaction more intuitive, natural, and enjoyable for the person and sustains his or her interest in the encounter.

In sum, although one could always argue that a robot does not need emotion-based mechanisms to address these issues, our point is that such mechanisms can be used to address these issues notably better than with cognitive mechanisms alone. Furthermore, robots today are far from behaving and learning as intelligently, flexibly, and robustly as people and other animals do with emotions. In our own work, we have shown that insights from emotion theory can be used to improve the performance of an autonomous robot that must pursue and achieve multiple goals in uncertain environments with people. With both cognitive and emotive systems working in concert, Kismet functions more adeptly—both from a decision-making and task-achieving perspective as well as from a social interaction and communication perspective.

As robot builders, we shall continue to design integrated systems for robots with internal mechanisms that complement and modulate its cognitive capabilities to improve the robot's overall performance. Several of these mechanisms may be close analogs to those regulatory, signaling, biasing, and other useful attention, value-assessment, and prioritization mechanisms associated with emotions in living creatures. As a consequence, we will effectively be giving robots a system that serves the same useful functions that emotions serve in us—no matter what we call it. Kismet is an early exploration of these ideas and a promising first step. Much more work has yet to be done to more deeply explore, demonstrate, and understand the benefit of emotion-inspired mechanisms on intelligent decision-making, reasoning, memory, and learning strategies of autonomous robots. Improvement of these abilities will be critical for autonomous robots that will one day play a rich and rewarding part in our daily lives.

Note

The principal author gratefully acknowledges the MIT Media Lab corporate sponsors of the Things That Think and Digital Life consortia for supporting her work and that of her students. Kismet was developed at the MIT Artificial Intelligence Lab while working in the Humanoids Robotics Group of the second author. The development of Kismet was funded by Nippon Telegraph and Telephone and Defense Advanced Research Projects Agency contract DABT 63-99-1-0012.

1. In collaboration with sociologist Dr. Sherry Turkle, human subjects of different ages were brought in to participate in HRI studies with Kismet.

References

Ackerman, B., Abe, J., & Izard, C. (1998). Differential emotions theory and emotional development: Mindful of modularity. In M. Mascolo & S. Griffen (Eds.), *What develops in emotional development?* (pp. 85–106). New York: Plenum.

Baron-Cohen, S. (1995). *Mindblindness*. Cambridge, MA: MIT Press.

Blumberg, B., Todd, P., & Maes, M. (1996). No bad dogs: Ethological lessons for learning. In *Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior* (*SAB96*) (pp. 295–304). Cambridge, MA: MIT Press.

Breazeal, C. (2002a). *Designing sociable robots*. Cambridge, MA: MIT Press.

Breazeal, C. (2002b). Regulation and entrainment for human–robot interaction. *International Journal of Experimental Robotics*, *21*, 883–902.

Breazeal, C. (2003a). Emotion and sociable humanoid robots. *International Journal of Human–Computer Studies*, *59*, 119–155.

Breazeal, C. (2003b). Emotive qualities in lip synchronized robot speech. *Advanced Robotics*, *17*, 97–113.

Breazeal, C. (2004). Social interactions in HRI: The robot view. *IEEE Transactions on Systems, Man, and Cybernetics*, *34* (Part C. Applications and Reviews), 181–186.

Breazeal, C. (2004). Function meets style: Insights from emotion theory applied to HRI. *IEEE Transactions on Systems, Man, and Cybernetics*, *34* (Part C: Applications and Reviews), 187–194.

Breazeal, C., & Aryananda, L. (2002). Recognition of affective communicative intent in robot-directed speech. *Autonomous Robots*, *12*, 83–104.

Breazeal, C., Fitzpatrick, P., & Scassellati, B. (2001). Active vision systems for sociable robots. *IEEE Transactions on Systems, Man, and Cybernetics: Part A*, *31*, 443–453.

Breazeal, C., & Scassellati, B. (2000). Infant-like social interactions between a robot and a human caretaker. *Adaptive Behavior*, *8*, 47–72.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, *RA-2*, 253–262.

Damasio, A. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Putnam.

Darwin, C. (1872). *The expression of the emotions in man and animals*. London: John Murray.

Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Ekman, P. (1992). Are there basic emotions? *Psychological Review*, *99*, 550–553.

Estrada, C., Isen, A., & Young, M. (1994). Positive affect influences creative problem solving and reported source of practice satisfaction in physicians. *Motivation and Emotion*, *18*, 285–299.

Fernald, A. (1989). Intonation and communicative intent in mother's speech to infants: Is the melody the message? *Child Development*, *60*, 1497–1510.

Gould, J. (1982). *Ethology*. New York: Norton.

Grosz, B. (1996). Collaborative systems: AAAI-94 presidential address. *AI Magazine*, *17*, 67–85.

Isen, A. (1999). Positive affect and creativity. In S. Russ (Ed.), *Affect, creative experience, and psychological adjustment* (pp. 3–17). Philadelphia: Brunner-Mazel.

Isen, A. (2000). Positive affect and decision making. In M. Lewis & J. Haviland-Jones (Eds.), *Handbook of emotions* (2nd ed., pp. 417–435). New York: Guilford.

Isen, A., Rosenzweig, A., & Young, M. (1991). The influence of positive affect on clinical problem solving. *Medical Decision Making, 11*, 221–227.

Isen, A., Shalker, T., Clark, M., & Karp, L. (1978). Affect, accessibility of material and behavior: A cognitive loop? *Journal of Personality and Social Psychology, 36*, 1–12.

Izard, C. (1993). Four systems for emotion activation: Cognitive and noncognitive processes. *Psychological Review, 100*, 68–90.

Izard, C. (1997). Emotions and facial expressions: A perspective from differential emotions theory. In J. Russell & J. Fernandez-Dols (Eds.), *The psychology of facial expression* (pp. 57–77). Cambridge: Cambridge University Press.

Izard, C., & Ackerman, B. (2000). Motivational, organizational and regulatory functions of discrete emotions. In M. Lewis & J. Haviland-Jones (Eds.), *Handbook of emotions* (2nd ed., pp. 253–264). New York: Guilford.

Lorenz, K. (1950). Part and parcel in animal and human societies. In K. Lorenz (Ed.), *Studies in animal and human behavior* (Vol. 2, pp. 115–195). London: Methuen.

Lorenz, K. (1973). *Foundations of ethology*. New York: Springer-Verlag.

Maes, P. (1991). Learning behavior networks from experience. In *Proceedings of the First European Conference on Artificial Life* (*ECAL90*), Paris. Cambridge, MA: MIT Press.

Maturana, H., & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living*. Boston: Reidel.

McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115–133.

McFarland, D., & Bosser, T. (1993). *Intelligent behavior in animals and robots*. Cambridge, MA: MIT Press.

Meltzoff, A., & Moore, M. K. (1997). Explaining facial imitation: A theoretical model. *Early Development and Parenting, 6*, 179–192.

Minsky, M. (1986). *The society of mind*. New York: Simon & Schuster.

Moore, B., Underwood, B., & Rosenhan, D. (1984). Emotion, self, and others. In C. Izard, J. Kagen, & R. Zajonc (Eds.), *Emotions, cognition, and behavior* (pp. 464–483). New York: Cambridge University Press.

Norman, D. (2001). How might humans interact with robots? Keynote address to the Defense Advanced Research Projects Agency/National Science Foundation Workshop on Human–Robot Interaction, San Luis Obispo, CA. http://www.dnd.org/du.mss/Humans_and_Robots.html

Picard, R. (1997). *Affective computation*. Cambridge, MA: MIT Press.

Plutchik, R. (1991). *The emotions*. Lanham, MD: University Press of America.

Premack, D., & Premack, A. (1995). Origins of human social competence. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 205–218). New York: Bradford.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*, p. 320. Distributed for the

Center for the Study of Language and Information. Chicago: University of Chicago Press.

Smith, C., & Scott, H. (1997). A componential approach to the meaning of facial expressions. In J. Russell & J. Fernandez-Dols (Eds.), *The psychology of facial expression* (pp. 229–254). Cambridge: Cambridge University Press.

Termine, N., & Izard, C. (1988). Infants' responses to their mothers' expressions of joy and sadness. *Developmental Psychology*, *24*, 223–229.

Tinbergen, N. (1951). *The study of instinct*. New York: Oxford University Press.

Tomkins, S. (1963). *Affect, imagery, consciousness: The negative affects* (Vol. 2). New York: Springer.

Trevarthen, C. (1979). Communication and cooperation in early infancy: A description of primary intersubjectivity. In M. Bullowa (Ed.), *Before speech: The beginning of interpersonal communication* (pp. 321–348). Cambridge: Cambridge University Press.

Tronick, E., Als, H., & Adamson, L. (1979). Structure of early face-to-face communicative interactions. In M. Bullowa (Ed.), *Before speech: The beginning of interpersonal Communication* (pp. 349–370). Cambridge: Cambridge University Press.

11 The Role of Emotions in Multiagent Teamwork

RANJIT NAIR, MILIND TAMBE, AND STACY MARSELLA

Emotions play a significant role in human teamwork. However, despite the significant progress in AI work on multiagent architectures, as well as progress in computational models of emotions, there have been very few investigations of the role of emotions in multiagent teamwork. This chapter begins to address this shortcoming. We provide a short survey of the state of the art in multiagent teamwork and in computational models of emotions. We then consider three cases of teamwork—teams of simulated humans, agent-human teams, and pure agent teams—and examine the effects of introducing emotions in each. Finally, we provide experimental results illustrating the impact of emotions on multiagent teamwork.

The advantages of teamwork among humans have been widely endorsed by experts in sports (Jennings, 1990) and business organizations (Katzenbach & Smith, 1994). Andrew Carnegie, one of America's most successful businessmen, highlighted the crucial role of teamwork in any organization:

> Teamwork is the ability to work together toward a common vision. The ability to direct individual accomplishments toward organizational objectives. It is the fuel that allows common people to attain uncommon results.

When team members align their personal goals with the goals of the team, they can achieve more than any of them individually.

Moving away from human organizations to organizations of artificial intelligence entities called "agents," we find similar advantages for teamwork. An *agent* is defined as "a computer system that is *situated* in some *environment*, and is capable of *autonomous action* in this environment in order to meet its design objectives" (Wooldridge, 2000). This computer system could be either a software agent that exists in a virtual environment or a hardware entity like a robot that operates in a real environment. The design objectives of the system can be thought of as the goals of the agent. The study of multiple agents working collaboratively or competitively in an environment is a subfield of distributed artificial intelligence called *multiagent systems*. In this chapter, we will focus on *collaborative* multiagent systems, where agents can benefit by working as a team.

In today's multiagent applications, such as simulated or robotic soccer (Kitano et al., 1997), urban search-and-rescue simulations (Kitano, Tadokoro, & Noda, 1999), battlefield simulations (Tambe, 1997), and artificial personal assistants (Scerri, Pynadath, & Tambe, 2002), agents have to work together in order to complete some task. For instance, ambulance and fire-engine agents need to work together to save as many civilians as possible in an urban search-and-rescue simulation (Kitano, Tadokoro, & Noda, 1999), and personal-assistant agents representing different humans need to work together to schedule a meeting between these humans (Scerri, Pynadath, & Tambe, 2002). This involves choosing individual goals that are aligned with the overall team goal. To that end, several teamwork theories and models (Cohen & Levesque, 1991; Grosz & Kraus, 1996; Tambe, 1997; Jennings, 1995) have been proposed that help in the coordination of teams, deciding, for instance, when and what they should communicate (Pynadath & Tambe, 2002) and how they should form and reform these teams (Hunsberger and Grosz, 2000; Nair, Tambe, & Marsella, 2003). Through the use of these models of teamwork, large-scale multiagent teams have been deployed successfully in a variety of complex domains (Kitano et al., 1997, 1999; Tambe, 1997; Scerri, Pynadath, & Tambe, 2002).

Despite the practical success of multiagent teamwork, the role of emotions in such teamwork remains to be investigated. In human teams, much emphasis is placed on the emotional state of the members and on methods of making sure that the members understand each others' emotions and help keep each other motivated about the team's goal (Katzenbach & Smith, 1994; Jennings, 1990). Behavioral work in humans and other animals (Lazarus, 1991; Darwin, 1872/1998; Oatley, 1992; Goleman, 1995) suggests several roles for emotions and emotional expression in teamwork. First, emotions act like a value system, allowing each individual to perceive its situation and

then arrive at a decision rapidly. This can be very beneficial in situations where the individual needs to think and act quickly. Second, the emotional expressions of an individual can act as a cue to others, communicating to them something about the situation that it is in and about its likely behavior. For instance, if we detect fear in someone's behavior, we are alerted that something dangerous might be present. Also, a person displaying an emotion like fear may behave in an irrational way. Being receptive to the emotional cues of this fearful team member allows us to collaborate with that person or compensate for that person's behavior.

In spite of these advantages to human teams, the role of emotions has not been studied adequately for multiagent teams. In this chapter, we will speculate on how multiagent teams stand to gain through the introduction of emotions. The following section describes briefly the state of the art in multiagent teamwork and in agent emotions. We then describe how multiagent teamwork and emotions can be intermixed and the benefits of such a synthesis. In particular, we will consider three types of team: simulated human teams, mixed agent–human teams, and pure agent teams (see also Chapter 10, Breazeal & Brooks). Finally, we will demonstrate empirically the effect of introducing emotions in a team of helicopters involved in a mission rehearsal.

STATE OF THE ART IN MULTIAGENT TEAMWORK AND AGENT EMOTIONS: A QUICK SURVEY

There is an emerging consensus among researchers in multiagent systems that teamwork can enable flexible coordination among multiple heterogeneous entities and allow them to achieve their shared goals (Cohen & Levesque, 1991; Tambe, 1997; Grosz & Kraus, 1996; Jennings, 1995). Furthermore, such work has also illustrated that effective teamwork can be achieved through team-coordination algorithms (sometimes called "teamwork models") that are independent of the domain in which the agents are situated. Given that each agent is empowered with teamwork capabilities via teamwork models, it is feasible to write a high-level team-oriented program (TOP) (Tambe, 1997; Tidhar, 1993), and the teamwork models then automatically generate the required coordination. In particular, TOPs omit details of coordination, thus enabling developers to provide high-level specifications to the team to perform team actions rather than invest effort in writing code for low-level detailed coordination. The teamwork models that govern coordination are based on a *belief–desire–intention* (*BDI*) *architecture*, where *beliefs* are information about the world that an agent believes to be true, *desires* are world states that the agent would like to see happen, and *intentions* are

effects that the agent has committed itself to achieving. The beliefs of an agent need not be true; for instance, an agent may receive incorrect observations owing to faulty sensors, and the different agents in the team could have different beliefs owing to differences in how and what they observe. Also, different agents in the team could have desires that conflict with each other. The goal of the team-coordination algorithms is to achieve *mutual belief* among the team members and to form joint intentions to allow the agents to work toward the same goal (see also Chapter 8, Sloman et al.).

We illustrate the use of TOPs through several example domains: mission rehearsal (Tambe, 1997), RoboCupRescue (Nair, Tambe, & Marsella, 2003), and Electric Elves (Scerri, Pynadath, & Tambe, 2002). A description of these domains is also helpful for our discussion of emotions below. For expository purposes, we have intentionally simplified the mission rehearsal domain: a helicopter team is executing a mission of transporting valuable cargo from point X to point Y through enemy terrain (see Fig. 11.1). There are three paths from X to Y of different lengths and different risks due to enemy fire. One or more scouting subteams must be sent out on different routes (some routes may not have a scouting team); and the larger the size of a scouting subteam, the safer it is. When scouts clear up any one path from X to Y, the transports can move more safely along that path. However, the scouts may fail along a path and may need to be replaced by a transport at the cost of not transporting cargo. Of course, we wish for the largest amount of cargo to be transported in the quickest possible manner within the mission deadline.

The TOPs for domains such as these consist of three key aspects of a team: (1) a team organization hierarchy consisting of roles, (2) a team (reac-

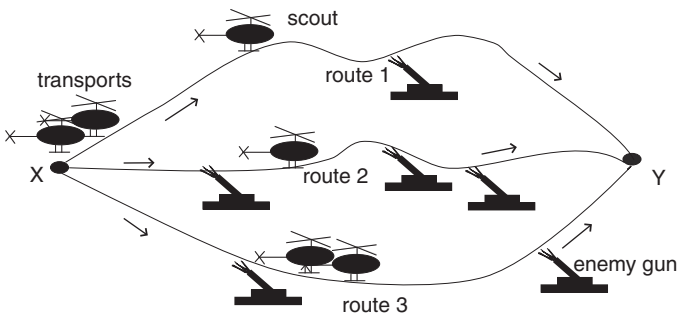


Figure 11.1. Helicopter mission rehearsal domain. Helicopters take on the role of either scouting a route or transporting cargo along a scouted route. Helicopters may be shot down by enemy guns on unscouted routes. The goal is to determine what roles each of the helicopters should take so as to get as much cargo as possible from point X to point Y within the mission deadline.

tive) plan hierarchy, and (3) an assignment of roles to execute plans. Thus, the developer need not specify low-level coordination details. Instead, the TOP interpreter (the underlying coordination infrastructure) automatically enables agents to decide when and with whom to communicate and how to reallocate roles upon failure. In the TOP for this example, we first specify the team organization hierarchy (see Fig. 11.2*a*). *Task Force* is the highest-level team in this organization and consists of two subteams, scouting and transport, where the scouting subteam has roles for each of three sub-subteams. Next, we specify a hierarchy of reactive team plans (see Fig. 11.2*b*). Reactive team plans explicitly express joint activities of the relevant team and consist of (1) initiation conditions under which the plan is to be proposed, (2) termination conditions under which the plan is to be ended, and (3) team-level actions to be executed as part of the plan. In Figure 11.2*b*, the highest-level plan, *Execute Mission*, has three subplans: *DoScouting* to make one path from X to Y safe for the transports, *DoTransport* to move the transports along a scouted path, and *RemainingScouts* for the scouts which have not reached the destination.

Figure 11.2*b* also shows coordination relationships: an *AND* relationship (depicted with a solid arc) indicates subplans that need to be completed successfully for the parent plan to succeed, while an *OR* relationship (depicted with a dashed arc) indicates that success of any one of the sub-plans will result in the parent subplan succeeding. Thus, *DoScouting*, *DoTransport*, and *RemainingScouts* must all be successful, while at least one of *UseRoute1*, *UseRoute2*, *and UseRoute3* must be performed successfully. There is also a temporal dependence relationship among the subplans (depicted with a

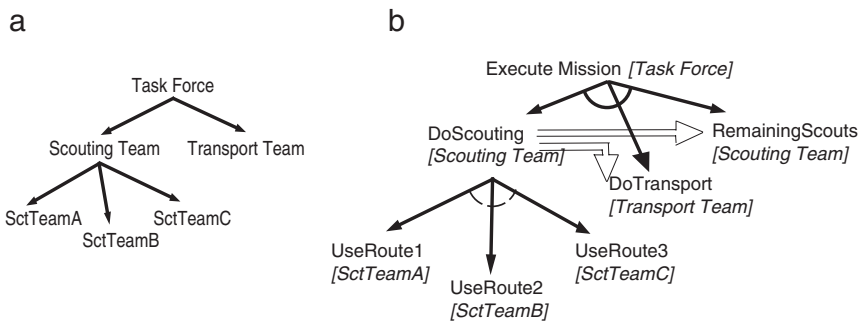


Figure 11.2. Team-oriented program for the helicopter domain. (*a*) Organization hierarchy. (*b*) Plan hierarchy. An AND relationship (depicted with a solid arc) indicates subplans that need to be completed successfully for the parent plan to succeed; an OR relationship (depicted with a dashed arc) indicates that success of any one of the subplans will result in the parent subplan succeeding.

double arrow), which implies that succeeding subplans cannot begin until the preceding subplan has been successfully completed. Thus, the subteams assigned to perform *DoTransport* or *RemainingScouts* cannot do so until the *DoScouting* plan has succeeded. However, *DoTransport* and *RemainingScouts* execute in parallel. Finally, we assign roles to plans. Figure 11.2*b* shows the assignment in brackets adjacent to the plans. For instance, The *Task Force* team is assigned to jointly perform *Execute Mission*.

New techniques for combining BDI approaches like TOP with decision-theoretic approaches based on partially observable Markov decision processes (POMDPs) have recently emerged (Schut, Wooldridge, & Parsons, 2001; Nair, Tambe, & Marsella, 2003). The advantage of the BDI approach is that it allows specification of large plans for complex domains. Unfortunately, such complex domains generally contain uncertainty. A Markov decision process (MDP) (Howard, 1960) is a formal representation of a domain with a single agent where there is uncertainty because the agent's actions have probabilistic outcomes. However, MDPs make an unrealistic assumption that the agent can sense the world state precisely. A POMDP is a generalization of an MDP, where the single agent may not observe the entire world state but only some observations drawn from some probability distribution. However, both MDPs and POMDPs are for single agents. Distributed POMDP models (Bernstein, Zilberstein, & Immerman, 2000; Boutilier, 1996; Pynadath & Tambe, 2002; Nair, Tambe, & Marsella, 2003) are generalizations of POMDPs to the case where there are multiple agents, each with a possibly different partial view of the world state. Both POMDPs and distributed POMDPs are computationally expensive to use to find the optimal plan for very large domains. However, they are very useful for analyzing existing team plans and coordination algorithms. For instance, Schut, Wooldridge, and Parsons (2001) compare various strategies for intention reconsideration (deciding when to deliberate about its intentions) by modeling a BDI system using a POMDP, but their work is confined to a single agent.

Nair, Tambe, and Marsella (2003) use role-based Markov team decision problem (RMTDP), a distributed POMDP model, for analysis of TOPs, where the results of RMTDP analysis are fed back into the BDI-based TOPs (see Fig. 11.3). The RMTDP for a team of n agents is defined as a tuple $<S, A, P, \Omega, O, R, RL>$. It consists of a finite set of states, S. $P(s, <a_1 \ldots a_n> s')$ gives the probability of transitioning from state s to state s' given that the agents perform the actions $<a_1 \ldots a_n> \in A$ jointly. Each agent i receives an observation $\omega_i \in \Omega_i$ based on the function $O(s, <a_1 \ldots a_n>, \omega_1 \ldots \omega_n)$, which gives the probability that the agents receive the observations, $\omega_1 \ldots \omega_n$, given that the world state is s and they perform $<a_1 \ldots a_n>$ jointly. $RL = \{r_1 \ldots r_s\}$ is a set of all roles that the agents can undertake. Each instance of role r_j may be assigned some agent i to fulfill it. Each agent's actions are distinguishable

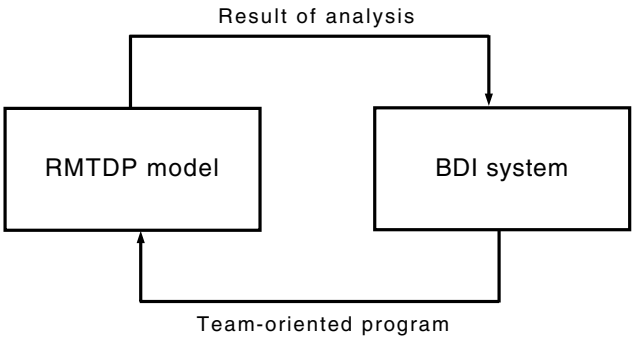


Figure 11.3. Integration of the belief–desire–intention (BDI)–based team–oriented programming approach with the role-based multiagent team decision problem (RMTDP) model, a distributed partially observable Markov decision process model. Analysis of the team-oriented program using the RMTDP model is fed back to improve the team-oriented program.

into role-taking and role-execution actions. The agents receive a single joint reward, $R(s, a_1 \ldots a_n)$. This model is used for evaluating various role-allocation and reallocation strategies for TOPs. Below (see "Experimental Illustration"), we show how the emotional state of the agents in the team can be modeled using RMTDP and empirically how the emotional state of the agents can affect how roles should be allocated.

Another example domain where TOPs have been applied is RoboCupRescue (Nair, Tambe, & Marsella, 2003; Kitano, Tadokoro, & Noda, 1999). This is a large-scale simulation of a city that has just undergone an earthquake. Here, teams of ambulance and fire-engine agents have to be formed and dispatched to various buildings to put out fires and rescue civilians trapped within them. TOPs allow us to flexibly coordinate the activities of these agents, monitoring the situations and reforming the teams if necessary.

The third example domain we discuss is the Electric Elves (E-Elves) project, which deploys an agent organization in support of the daily activities of a human organization (Scerri, Pynadath, & Tambe, 2002; Chalupsky et al., 2001). We believe this application to be fairly typical of future generations of applications involving teams of agents and humans. The operation of a human organization requires the performance of many everyday tasks to ensure coherence in organizational activities, for example, monitoring the status of activities, gathering information, and keeping everyone informed of changes in activities. Teams of software agents (proxy agents) can aid organizations in accomplishing these tasks, facilitating coherent functioning and rapid response to crises. The notion of an agent proxy is similar to

that of "robot as avatar" (see Chapter 10, Brezeal and Brooks). While the goal of both is to reduce the cognitive load on humans, the key difference is that an agent proxy exists only in software and needs to interact with several other agent proxies in addition to the human it represents. Each agent proxy is called "Friday" (after Robinson Crusoe's manservant Friday) and acts on behalf of its user in the agent team. Currently, each Friday can perform several tasks for its user. If a user is unable to attend a meeting, Friday can reschedule the meeting, informing other Fridays, who in turn inform their users. If there is a research presentation slot open, Friday may respond to the invitation to present on behalf of its user. Friday can also order its user's meals and track the user's location, posting it on a web page. Friday communicates with users using wireless devices, such as personal digital assistants. Each Friday's team behavior is based on a teamwork model, called "*S*hell for *TEAM*work" (STEAM) (Tambe, 1997). The STEAM model encodes and enforces the constraints between roles that are required for the success of the joint activity, for instance, meeting attendees should arrive at a meeting simultaneously. When a role within the team needs to be filled, STEAM requires that a team member be assigned responsibility. To find the best-suited person, the team auctions off the role, allowing it to consider a combination of factors and assign the best-suited user.

Computational Models for Emotion

Interest in computational models of emotion and emotional behavior has been steadily growing in the agent and artificial intelligence research communities. Although the creation of general computational models is of potential interest in understanding human behavior, much of the interest in the agent community has been fueled by the application areas for such models. For example, there has been a growing body of work in the design of *virtual humans*, software artifacts that act like people but exist in virtual worlds, interacting with immersed humans and other virtual humans. Virtual human technology is being applied to training applications (Rickel et al., 2002), health interventions (Marsella, Johnson, & LaBore, 2000), marketing (André, Rist, Mulken, & Klesen, 2000), and entertainment (Cavazza, Charles, & Mead, 2002). Emotion models are a critical component of this technology, enabling virtual humans that are better facsimiles of humans as well as providing a more engaging experience. Emotion models have also been proposed as a critical component of more effective human–computer interaction that factors in the emotional state of the user (Lisetti & Schiano, 2000; Picard, 1997).

Much of the work on computational models of emotion has been strongly influenced by cognitive appraisal theories of emotion (Ortony, Clore, & Collins, 1988; Frijda, 1987; Lazarus, 1991), although some computational research (Velásquez, 1998) has also been influenced by theories that posit noncognitive sources of emotion (Izard, 1993). Appraisal theories argue that emotion stems from a person's assessment of his or her relationship to the environment in terms of a set of appraisal variables or dimensions covering such factors as whether an event facilitates or inhibits the individual's goals, how desirable the impacted goals are, who deserves blame or credit, etc. Among the various cognitive appraisal theories, there is broad agreement on the set of appraisal variables, and these have provided a detailed framework for building computational models of the causes of emotion (see also Chapter 7, Ortony et al.). Of course, emotions also impact behavior in myriad ways. In particular, appraisal has been related to action readiness (or tendencies) and facial expressions (Smith & Scott, 1997).

The work of Richard Lazarus (1991) also tightly integrates appraisal with human *coping* behavior the process of managing one's emotions by either acting externally on the world (*problem-focused coping*) or acting internally to change one's beliefs or attention (*emotion-focused coping*). Specifically, coping manages emotions by attempting to alter the person's appraisal through a combination of altering the factors in the environment that are leading the emotion, altering beliefs about those factors, or altering the attention to the factors. For example, a person can focus effort and attention on countering a threat to a goal, decide that the goal is not so important, or avoid thinking about the threat. Which of these will be an effective coping response depends on the seriousness/likelihood of the threat, the relevance of the goal, and a person's potential to deal with the threat, among other factors. A person may or may not make an effective response, and thus, emotional stress may lead to adaptive or maladaptive coping behavior (see Arbib's Chapter 12). The interaction of appraisal and coping unfolds over time, modeling the temporal character of emotion noted by several researchers (Lazarus, 1991; Scherer, 1984): an agent may "feel" distress for an event (appraisal), which motivates the shifting of blame (coping) to another person, leading to anger at the now blameworthy other person (reappraisal).

One of the appeals of cognitive appraisal models as the basis of a computational model is the ease with which appraisal can be tied to the BDI framework often used in agent systems. In fact, computer science researchers have realized a range of appraisal-based approaches to modeling how emotions arise. For example, Elliott's (1992) affective reasoner is a computational realization of Ortony, Clore, and Collins' (1988) appraisal model (see also Chapter 7, Ortony et al.). Elliott's model characterizes events in

terms of specific appraisal variables and has the capability to appraise the same event from multiple perspectives (from the agent's own perspective and the supposed perspective of other agents), clearly a useful capability from a social interaction and teamwork perspective. However, the model requires domain-specific rules to appraise events.

Recent approaches have increasingly incorporated emotions into general artificial intelligence architectures that fit well with the planning frameworks typically used in multiagent teams. For example, Neil Reilly (1996) uses a reactive planning framework that associates desirability with probability of goal attainment but with domain-specific rules to derive the likelihood of threats or goal attainment. El Nasr, Yen, and Ioerger (2000) use MDP to provide a very general framework for characterizing the desirability of actions and events. A key advance of this method is that it can represent indirect consequences of actions by examining their impact on future reward (as encoded in the MDP). The Will architecture (Moffat & Frijda, 1995) ties appraisal variables to an explicit model of plans (which capture the causal relationships between actions and effects).

One aspect of emotional behavior that is missing from most computational models is a detailed account of the many ways that humans cope with emotion. Rather, the emphasis of the models has been on simple action selection and facial expression. Marsella and Gratch (2002) address this limitation by providing a domain-independent model of coping that attempts to capture the full range of human coping behavior, including not only action selection but also more sophisticated problem-focused and emotion-focused strategies. Among these strategies are planful problem solving, *positive re-interpretation* (finding positive meaning in an otherwise negative event such as a loved one's illness), acceptance that a future negative event is inevitable, shifting blame, and denial/wishful thinking (Marsella & Gratch, 2003). In their model, coping is essentially the inverse of appraisal, changing one or more of the appraisal factors that contributed to the emotion. Both appraisal and coping are integrated within a general BDI planning framework that employs a uniform causal interpretation of the world from the agent's perspective. The causal interpretation incorporates both past events (an episodic memory) and the agent's plans concerning future events. Thus, appraisal of past events can lead to coping responses that operate on the beliefs about that event; an agent may, for example, begin to believe a bad accident was someone else's fault. Similarly, intentions about future events can impact present emotions; forming the intent to redress a wrong will make an agent feel better even before the intent is enacted. Coping is modeled as a set of basic operations that manipulate appraisal factors and can be combined to create a range of coping strategies. For example, an agent may plan a response to a threat while engaging in wishful thinking about the likelihood of suc-

cess or potential negative consequences. This mirrors how coping processes are understood to operate in human behavior, whereby, for example, people may employ a mix of problem-focused and emotion-focused coping to deal with stress. This work tightly integrates appraisal and coping in an effort to model their unfolding temporal dynamics.

HOW COULD EMOTIONS AFFECT MULTIAGENT TEAMWORK?

In this section, we will discuss the implication of introducing emotions into different kinds of multiagent team. In particular, we consider three types of team: teams of simulated humans, mixed agent–human teams, and pure agent teams. The need for including emotions in multiagent teams is different depending on the nature of the team. For instance, in the case of teams of simulated humans, emotions need to be modeled for an accurate simulation. The implications of introducing emotions will vary depending on the constituents of the team. In the case of mixed human–agent teams, the introduction of emotions may actually improve performance. In this section, we discuss the role that emotions play in each type of team and the issues involved in adding emotions to the team. We conclude that, at least in the first two cases, i.e., in teams of simulated humans and in mixed agent–human teams, computational models of emotions based on appraisal theories can play a critical role.

Teams of Simulated Humans

The central role of emotion in human teamwork becomes apparent when working through the wide-ranging impacts it has on decision making, goal prioritization, perception, belief changes, action selection, etc. Hence, in teams where each agent is a facsimile of a human being, one would have to introduce emotions in order to represent human behavior faithfully. For example, domains such as the mission-rehearsal domain are focused on simulation-based training, to provide the right training environment for human participants by requiring each agent to simulate human behavior. In order to analyze or predict the behavior of humans in adverse scenarios, it is important to study the influence of emotions like fear that such scenarios bring about in humans. For example, in a demonstration of the helicopter agents that did not model emotions, it was found that even after all its teammates were shot down the sole remaining helicopter continued to execute its mission completely unaffectedly, much to the consternation of the military experts. In particular, fear for

their own self-survival might motivate the members of a human team to abandon the team's goals in the face of the high number of fatalities. Further, an individual's fear would tend to spread across members of the team, influencing the decision making of the surviving members.

Introducing emotions like fear could result in the team's performance worsening, but as this example clearly highlights, for an accurate portrayal of human organizational behavior, it is important to include such emotions. In particular, within such teams and organizations, emotions play a major role in choosing between a human's private and team goals. In the helicopter scenario, each helicopter should have a private goal of self-preservation and a public team goal to accomplish the mission. As the scenario changes, the emotional state of the agent should change as well. Appraisal-based models can express the impact of such survival fear by providing survival as one of the agent's goals. Threats to that survival would lead to fear, with fear increasing with the expectation that the threat was more certain. At some juncture, this fear for individual survival would override the agent's desire to achieve the team's mission goals. Further, the contagion-like process (Hatfield, Cacioppo, & Rapson, 1994) of one agent's fear affecting another could be modeled as a process whereby one agent's fear would affect another agent's appraisal of the seriousness of the threat.

Thus, depending on the current emotional state, the helicopter agent would choose between saving itself by retreating or continuing with the mission. Even after the agent chooses which goal to perform, emotions play a role in action selection. Thus, emotions could act as a value system that the agent uses in assessing a situation.

At the interagent level, emotions could act as cues that allow individuals to synchronize their goals/plans, to synchronize perceptual biases, to compensate for another's emotional state, or to establish a shared mental model. Again, we can look at natural systems for inspiration. For example, the expression of emotion on a mother's face has been shown to bias how a baby acts in ambiguous situations (Campos & Sternberg, 1981). This communicative role for emotions in social situations is also well recognized in psychological theories of emotion (Oatley, 1992; see also Adolphs' Chapter 2). In the helicopter domain, the pilots could perceive fear in the voices of the other pilots and, hence, conclude that the situation is dangerous even though the danger may not be visible to them. In addition, humans can, in essence, appraise events from others' perspectives and, thus, know how they will feel (see Jeannerod's Chapter 6). In the absence of more detailed information that can form the basis of more accurate threat assessment, these emotional signals can be very useful in threat detection.

Based on this discussion, we conclude that in teams where agents simulate humans we may need to introduce emotions because humans use emo-

tions for goal prioritization and action selection. We also need to build into agents the ability to perceive other agents' emotions and use these emotional cues to conclude something about the state and decision making of the other agents.

Mixed Agent–Human Teams: Virtual Organizations

In mixed agent–human teams, it is not important for agents to simulate humans and human emotions. However, it would be very beneficial if the agents model the emotions of the human teammates in order to get a better understanding of their needs and expected behaviors (Lisetti & Schiano, 2000; Picard, 1997). For example, in E-Elves, it would be useful if the agents could perceive the emotional state or mood of humans in order to know how to behave with them. For instance, if the "elf" could sense that the human was upset, it could decide not to disturb him or her with trivial questions. If the elf knew something about the emotional state of the human, it could anticipate the human's actions and be prepared ahead of time with information that the human would find useful.

Although agents need not have emotions themselves, they could display emotions in order to achieve a better rapport with the humans, achieve the cooperation of the humans, and make the human feel less like he or she was interacting with an inanimate entity. Thus, in agent–human teams, it maybe useful for the agents to not only model the humans' emotions but also display emotion itself (see Chapter 10, Breazeal and Brooks).

Pure Agent Teams

The argument for including emotions in pure agent or robotic teams is more challenging to make. If we focus only on the role of emotion as a signal to other members of a team, a key question is whether emotion, as a signal, provides some capability not subsumed by existing models of communication in agent teams. In human teamwork, emotional signals like facial expressions inform each agent about the state of the world and about the internal state of its teammates. This communication differs in many respects from how agent teams communicate. In particular, the content of this communication is not simply specific statements about the world, as it often is in agent teams, but rather the individual's attitudes and emotional reactions to the world. Further, emotional signals can be intentional, unintentional, or a mixture of the two, as when people try to suppress their emotional expressions (Ekman, 2001). In contrast, pure agent teams communicate via

explicit, intended communication or by the intended actions they take in the world. Further emotional signals are communicated across a variety of channels, verbally and nonverbally. These channels vary in capacity, the specificity of the information effectively communicated, and the cognitive overhead in using them. A person can smile at a cute baby without much thought but may need more resources to verbally express happiness. Agent teams typically have two channels: communication and action. These differences suggest potential benefits for using emotions in pure agent teams. For instance, there might be an advantage to having agent teams communicate attitudinal or emotional information as well as an advantage to exposing this information to teammates automatically, through low-cost channels. Consider building agents so that they could not only communicate and act deliberately after an accurate and possibly computationally intensive assessment of the state, but also emit some low-cost emotional signal based on an approximate state assessment. For example, a robot could have hardwired circuitry that triggers light-emitting diodes that represent emotional cues like fear to indicate a state where the robot is in danger, worry to indicate low likelihood of success, and helplessness to indicate that it needs to help. These emotional cues can be computed and transmitted quickly and could result in the team being able to coordinate itself without having to wait for the accurate state estimation to be performed. If, for example, agents could use these emotional cues to determine action selection of the other agents in the team, it could result in greater synchronization and, consequently, better teamwork.

EXPERIMENTAL ILLUSTRATION

In this section, as an illustration of the effect of emotions on multiagent teamwork, we demonstrate how the allocation of roles in a team is affected by emotions like fear. Our approach is to introduce an RMTDP (Nair, Tambe, & Marsella, 2003) for the team of agents, then to model the agents such that their emotional states are included.

We now demonstrate how emotions can affect decision making in a team of helicopters. To this end, recall the RMTDP analysis of TOPs mentioned above. The emotional state of the agent could skew how the agent sees the world. This could result in the agent applying different transition, observation, or reward functions. In this discussion, we will focus on how fear may affect the reward function used in the RMTDP. For instance, in a fearful state, agents may consider the risk of failure to be much higher than in a nonfearful state. In the helicopter domain, such agents might

penalize heavily those states where a helicopter crashes. We now demonstrate how such a change in the emotional state of the agents would affect the best role allocation.

We consider a team of six helicopters and vary the number of agents that fear losing a helicopter to enemy fire. These agents would place a heavy penalty on those states where one or more helicopter crashed. Figure 11.4*a*,*b* shows the number of scouts allocated to each route (X-axis) as we vary the number of fearful agents in the team (Y-axis) from none to all six for two different penalties for helicopter crashes. In Figure 11.4*a*, when all the agents were fearless, the number of scouts sent out was three, all on route 2; however, when fearful agents were introduced, the number of scouts sent out changed to four, also on route 2, because the team was now prepared to lose out on the chance of a higher reward if they could ensure that each scout that was sent out would be safer. In Figure 11.4*b*, we reduced the amount of penalty the agents ascribed to a helicopter crash. When fearful agents were introduced, the number of scouts remained unchanged but the scouts now used route 1, a safer albeit longer route, instead of route 2, which was more dangerous but allowed the mission to be completed more quickly. Thus, with the introduction of fear, we found that the team's decision-making behavior changed such that the members either deployed more scouts or assigned the scouts to a safer route.

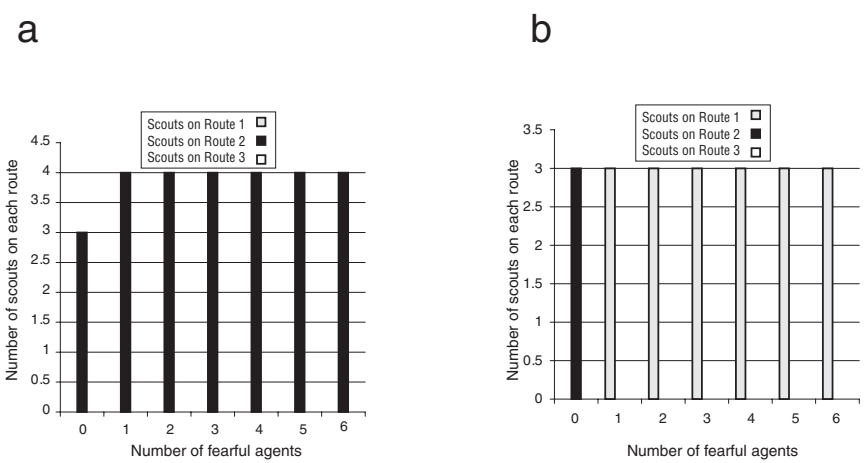


Figure 11.4. Role allocations in fearful teams with different reward functions. Role allocations for reward function. (*a*) Increasing the number of fearful agents results in more scouts being sent together to increase the safety of the scouting team. (*b*) Increasing the number of fearful agents results in moving scouts from a shorter but more risky route to a longer but safer route.

Although, the emotion "fear" was modeled simply as a penalty for states where a helicopter crashes, the purpose of the experiment was simply to show that emotional response affects what the team perceives is its best allocation. In order to evaluate teams where emotions are represented more realistically , we would need the following:

- A more realistic model of how an agent's emotional state would change based on new percepts. This model of how the emotional state transitions can be incorporated as part of the transition function in the RMTDP model in order to evaluate the team's performance in the presence of emotion.
- A more realistic model of how humans (which the agents are simulating) would respond based on their emotional state. This would form part of the TOP where the individual agent's action selection is specified.

Both the model of how emotional state changes as well as the model of human behavior in the presence of emotion should ideally be informed by human behavior in such task domains.

CONCLUSION

This chapter represents the first step in introducing emotions in multiagent teamwork. We examined the role of emotions in three different kinds of team: first, in teams of simulated humans, introducing emotions results in more believable agent behavior and consequently better simulations; second, in virtual organizations, where agents could simulate emotions to be more believable and engaging to the human and anticipate the human's needs by modeling the human; and third, in pure agent teams, where the introduction of emotions could help bring in the same advantages that emotions bring to human teams.

Teams of simulated agents and mixed human–agent teams can greatly benefit with computational models of emotion. In particular, to evaluate and improve such teams, we would need the following:

- A model of how an agent's emotional state would change based on new percepts
- A model of how humans would respond based on their emotional state

Acknowledgment This research was supported by grant 0208580 from the National Science Foundation.

References

André, E., Rist, T., Mulken, S. V., & Klesen, M. (2000). The automated design of believable dialogues for animated presentation teams. In J. Cassell, J. Sullivan, S. Prevost, & E. Churchill (Eds.), *Embodied conversational agents* (pp. 220–255). Cambridge, MA: MIT Press.

Bernstein, D. S., Zilberstein, S., & Immerman, N. (2000). The complexity of decentralized control of MDPs. In C. Boutilier & M. Goldszmidt (Eds.), *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence* (pp. 32–37), Stanford University. Stanford, CA: Morgan Kaufmann.

Boutilier, C. (1996). Planning, learning and coordination in multiagent decision processes. In Y. Shoham (Ed.), *Proceedings of the Sixth Conference on Theoretical Aspects of Rationality and Knowledge* (pp. 195–210). De Zeeuwse Stromen, The Netherlands: Morgan Kaufmann.

Campos, J., & Sternberg, C. (1981). Perception, appraisal and emotion: The onset of social referencing. In M. Lamb & L. Sherrod (Eds.), *Infant social cognition* (pp. 273–314). Hillsdale, NJ: Erlbaum.

Cavazza, M., Charles, F., & Mead, S. J., (2002). Interacting with virtual characters in interactive storytelling, In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems* (pp. 15–19). Bologna, Italy: ACM.

Chalupsky, H., Gil, Y., Knoblock, C., Lerman, K., Oh, J., Pynadath, D., Russ, T., & Tambe, M. (2001). Electric elves: Applying agent technology to support human organizations. In H. Hirsh & S. Chien (Eds.), *Proceedings of the Thirteenth Innovative Applications of Artificial Intelligence Conference*. Seattle, WA: AAAI.

Cohen, P. R., & Levesque, H. J., (1991). Teamwork. In *Noûs*, *25*(4), 487–512.

Darwin, C. (1998). *The expression of emotions in man and animals* (3rd ed.). New York: Oxford University Press. (Original work published 1872)

Ekman, P. (2001). *Telling lies: Clues to deceit in the marketplace, politics and marriage*. New York: Norton.

Elliott, C. (1992). The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System. Evanston, IL, Northwestern University Institute for the Learning Sciences. Dissertation.

El Nasr, M. S., Yen, J., & Ioerger, T. (2000). Flame: Fuzzy logic adaptive model of emotions. *Journal of Autonomous Agents and Multiagent Systems*, *3*, 219–257.

Frijda, N. (1987). Emotion, cognitive structure, and action tendency. *Cognition and Emotion*, *1*, 115–143.

Goleman, D. (1995). *Emotional intelligence*. New York: Oxford University Press.

Grosz, B., & Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*, *86*, 269–357.

Hatfield, E., Cacioppo, J., & Rapson, R. (1994). *Emotional contagion*, Cambridge: Cambridge University Press.

Howard, R. A. (1960). *Dynamic programming and Markov processes*. Cambridge, MA: MIT Press.

Hunsberger, L., & Grosz, B. (2000). A combinatorial auction for collaborative

planning. In *Proceedings of Fourth International Conference on Multiagent Systems* (*ICMAS-2000*) (pp. 151–158). Boston: IEEE Computer Society.

Izard, C. (1993). Four systems for emotion activation: Cognitive and noncognitive processes. *Psychological Review, 100*, 68–90.

Jennings, J. (1990). *Teamwork: United in victory*. Englewood Cliffs, NJ: Silver Burdett Press.

Jennings, N. (1995). Controlling cooperative problem solving in industrial multiagent systems using joint intentions. *Artificial Intelligence, 75*, 195–240.

Katzenbach, J., & Smith, D. K. (1994). *The wisdom of teams*. New York: Harper Business.

Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., & Osawa, E. (1997). RoboCup: The robot world cup initiative. In *Proceedings of First International Conference on Autonomous Agents* (*Agents '97*) (pp. 340–347), Marina del Rey, CA, Feb 5–8. New York: ACM Press.

Kitano, H., Tadokoro, S., & Noda, I. (1999). RoboCup-Rescue: Search and rescue for large scale disasters as a domain for multiagent research. In *Proceedings of IEEE Conference on Systems, Men, and Cybernetics (SMC-99)*. Tokyo, Japan: IEEE System, Man, and Cybernetics Society.

Lazarus, R. (1991). *Emotion and adaptation*. New York: Oxford University Press.

Lisetti, C. L., & Schiano, D. (2000). Facial expression recognition: Where human–computer interaction, artificial intelligence, and cognitive science intersect. *Pragmatics and Cognition, 8* 185–235.

Marsella, S., & Gratch, J. (2002). A step toward irrationality: Using emotion to change belief. In *Proceedings of First International Joint Conference on Autonomous Agents and Multi-agent Systems* (*AAMAS-02*) (pp. 334–341). Bologna, Italy: ACM.

Marsella, S., & Gratch, J. (2003). Modeling coping behavior in virtual humans: Don't worry, be happy. In *Proceedings of Second International Joint Conference on Autonomous Agents and Multi-agent Systems* (*AAMAS-03*) (pp. 313–320). Melbourne: ACM.

Marsella, S., Johnson, W. L., & LaBore, C. (2000). Interactive pedagogical drama. In *Proceedings of Fourth International Conference on Autonomous Agents* (*ICMAS-2000*) (301–308). Barcelona, Spain: ACM.

Moffat, D., & Frijda, N. (1994). Where there's a will there's an agent. In *Proceedings of Workshop on Agent Theories, Architectures and Languages* (*ATAL-95*) (pp. 245–260). Amsterdam: Springer.

Nair, R., Tambe, M., & Marsella, S. (2003). Role allocation and reallocation in multiagent teams: Towards a practical analysis. In *Proceedings of Second International Joint Conference on Autonomous Agents and Multi-agent Systems* (*AAMAS-03*) (pp. 552–559). Melbourne: ACM.

Oatley, K. (1992). *Best laid schemes: The psychology of emotions*. Cambridge: Cambridge University Press.

Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge: Cambridge University Press.

Picard, R.W. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Pynadath, D. V., & Tambe, M. (2002). The communicative multiagent team decision problem: Analyzing teamwork theories and models. *Journal of Artificial Intelligence Research*, *16*, 389–423.

Reilly, N. (1996). *Believable Social and Emotional Agents*. Pittsburgh: Carnegie Mellon University. Dissertation.

Rickel, J., Marsella, S., Gratch, J., Hill, R., Traum, D., & Swartout, W. (2002). Toward a new generation of virtual humans for interactive experiences. *IEEE Intelligent Systems*, *17*(4), 32–38.

Scerri, P., Pynadath, D. V., & Tambe, M. (2002). Towards adjustable autonomy for the real world. *Journal of Artificial Intelligence Research*, *17*, 171–228.

Scherer, K. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 293–317). Hillsdale, NJ: Erlbaum.

Schut, M. C., Wooldridge, M., & Parsons, S. (2001). Reasoning about intentions in uncertain domains. In Proceedings of European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Toulouse, France. *Lecture Notes in Computer Sciences*, *2143*, 84–85.

Smith, C. A., & Scott, H. S. (1997). A componential approach to the meaning of facial expressions. In J. A. Russell & J. M. Fernández-Dols, (Eds.), *The psychology of facial expression* (pp. 229–254). Paris: Cambridge University Press.

Tambe, M. (1997). Towards flexible teamwork. *Journal of Artificial Intelligence Research*, *7*, 83–124.

Tidhar, G. (1993). *Team-oriented programming: Social structures*. *Technical report 47*. Melbourne, Australia: Australian A.I. Institute.

Velásquez, J. (1998). When robots weep: Emotional memories and decision-making. In *Proceedings of Fifteenth National Conference on Artificial Intelligence* (*AAAI-98*) (pp. 70–75), Madison, WI: Cambridge, MA: MIT Press.

Wooldridge, M. (2000). Intelligent agents. In G. Weiss (Ed.), *Multiagent systems: A modern approach to distributed AI* (pp. 27–78). Cambridge, MA: MIT Press.

This page intentionally left blank

PART IV

CONCLUSIONS

This page intentionally left blank

12 Beware the Passionate Robot

MICHAEL A. ARBIB

The warning, "Beware the Passionate Robot," comes from the observation that human emotions sometimes have unfortunate effects, raising the concern that robot emotions might not always be optimal. However, the bulk of the chapter is concerned with biology: analyzing brain mechanisms for vision and language to ground an evolutionary account relating motivational systems to emotions and the cortical systems which elaborate them. Finally, I address the issue of whether and how to characterize emotions in such a way that one might say that a robot has emotions even if they are not empathically linked to human emotions.

A CAUTIONARY TALE

On Tuesday, I had an appointment with N at 3 P.M., but when I phoned his secretary at 2:45 to check the place of the meeting, I learned that she had forgotten to confirm the meeting with N. I was not particularly upset, we rescheduled the meeting for 4 P.M. the next day, and I proceeded to make contented use of the unexpected free time to catch up on my correspondence.

On Wednesday, I decided in midafternoon to put together a chart to discuss with N at our meeting; but the printer gave me some problems, and it was already after 4 when I left my office for the meeting, feeling somewhat flustered but glad that I had a useful

handout and pleased that I was only a few minutes late. I was looking forward to the meeting and experienced no regret or frustration that it had been postponed the day before. However, when I arrived at about 4:06, N's secretary wasn't there and neither was N. Another secretary explained that N's secretary had left a message for me earlier in the day to move the meeting forward to 2:30. (I had not received the message because my secretary had taken the day off to tend to her ill mother. When I heard the news that morning, I felt slightly frustrated that some correspondence would be delayed but equally concerned for the mother's health and had no question about the appropriateness of my secretary's action. The matter, having been accepted, had no further effect upon my mood, at least until I learned of the loss of the message. At midday, I had been transiently and mildly upset by the cancellation of a luncheon appointment.) N was not available. Would I care to make an appointment? With a curt "No," I turned on my heels, and stormed out of the office and back to my own. As I walked, I simultaneously felt fury at the double cancellation and shame at my own rude behavior, as well as realizing that another appointment had to be made. I felt tight and constricted. After a minute or two in my office, unable to concentrate and with my thoughts dominated by this stew of emotions, I decided to return to N's office. Once there, I apologized to the secretary for my discourtesy, explained the annoyance of the double cancellation, set a new appointment, made a feeble joke in the form of a threat about the dire consequences of another cancellation, and then returned to my office. The physiological effects I had felt just a few minutes earlier quickly dissipated. I felt pleased that I had "done the right thing."

In this particular instance, no real harm was done. In other cases, my shows of temper have triggered unfortunate effects in others slower to anger and slower to forgive, which have had long-lasting consequences. I say this not to confess my faults or engender curiosity about my autobiography but simply to make the point, almost entirely neglected elsewhere in this book, that emotions can have negative consequences (cf. Chapter 10, "Emotional disturbances," of Hebb, 1949).[1] Thus my warning, "Beware the Passionate Robot." If we create a computer tutor, it may be useful to provide it with a human-like voice and perhaps even a face that can provide emotional inflections to its delivery of information, thus adding a human-like emollient that may aid the human student's learning. However, if the computer were to become so exasperated with a poor student that it could lose its temper, screaming out the student's shortcomings in emotional tones laced with in-

vective and carrying out the electronic equivalent of tearing up the student's papers in a rage, then where would the benefit lie? One might argue that even though such outbursts are harmful to many children, they may be the only way to "get through" to others; but if this is so, and the production of emotional behavior is carefully computed within an optimal tutoring strategy, it may be debated whether the computer tutor really has emotions or is simply "simulating the appearance of emotional behavior"—a key distinction for the discussion of robot emotions. We will return to these questions later (see Emotion without Biology, below).

To complement the above account of my own tangled emotions on one occasion, I turn to a fictional account of the mental life of a chimpanzee under stress, an excerpt from a lecture by the fictional Australian writer Elizabeth Costello as imagined by J. M. Coetzee (2003):

> In 1912 the Prussian Academy of Sciences established on the island of Tenerife a station devoted to experimentation into the mental capacities of apes, particularly chimpanzees. . . . In 1917 Köhler published a monograph entitled *The Mentality of Apes* describing his experiments. Sultan, the best of his pupils . . . is alone in his pen. He is hungry: the food that used to arrive regularly has unaccountably ceased coming. The man who used to feed him and has now stopped feeding him stretches a wire over the pen three metres above ground level, and hangs a bunch of bananas from it. Into the pen he drags three wooden crates. . . . One thinks: Why is he starving me? One thinks: What have I done? Why has he stopped liking me? One thinks: Why does he not want these crates any more? But none of these is the right thought. . . . The right thought to think is: How does one use the crates to reach the bananas? Sultan drags the crates under the bananas, piles them one on top of the other, climbs the tower he has built, and pulls down the bananas. He thinks: Now will he stop punishing me? . . . At every turn Sultan is driven to think the less interesting thought. From the purity of speculation (Why do men behave like this?) he is relentlessly propelled towards lower, practical, instrumental reason (How does one use this to get that?) and thus towards acceptance of himself as primarily an organism with an appetite that needs to be satisfied. (pp. 71–73)

This may or may not be a realistic account of what Sultan was thinking (see de Waal, 2001, for the views of a primatologist who supports such "anthropomorphism"), but my point here is to stress a "two-way reductionism" (Arbib, 1985; Arbib & Hesse, 1986) which understands the need to establish a dialog between the formal concepts of scientific reductionism and the richness of personal experience that drives our interest in cognition and

emotion. How can we integrate the imagination of the novelist with the rigor of the neurobiologist?

In the opening chapter of this book, our fictitious "Russell" argued for the utility of definitions in the analysis of emotion, only to be rebuffed by "Edison" with his emphasis on inventions. Being somewhat Russellian, let me provide here some definitions based on those in the O*xford English Dictionary* (OED). While some biologists stress that everyday language often lacks the rigor needed to advance scientific research, I believe that—in the spirit of Elizabeth Costello—we have much to gain by challenging our scientific concepts in cognitive science and artificial intelligence by confronting them with the human experience that normal usage enshrines. As those familiar with the OED know, each word comes with its etymology and with a range of definitions and relevant historical quotations. What follows is an edited and no doubt biased sampling that may be useful for systematizing what has been learned in this volume.[2]

> **Emotion: 4b**. *Psychology*. A mental "feeling" or "affection" (*e.g.* of pleasure or pain, desire or aversion, surprise, hope or fear, etc.), as distinguished from cognitive or volitional states of consciousness.
>
> **Motivation: b.** orig. *Psychol.* The (conscious or unconscious) stimulus for action towards a desired goal, esp. as resulting from psychological or social factors; the factors giving purpose or direction to human or animal behaviour.
>
> **Affect: I.** Mental. **1. a.** The way in which one is affected or disposed; mental state, mood, feeling, desire, intention. *esp*. **b.** Inward disposition, feeling, as contrasted with external manifestation or action; intent, intention, earnest, reality. **c.** Feeling, desire, or appetite, as opposed to *reason*; passion, lust, evil-desire.

On the basis of these definitions, I see a spectrum from *motivation* and *affect*, which dispose one to act in a certain way, to *emotion*, which is linked to conscious feelings of "pleasure or pain, desire or aversion, surprise, hope or fear, etc." Thus, where Fellous and LeDoux (Chapter 4), for example, are comfortable speaking of "emotional behavior" that may be unaccompanied by "emotional feelings," I usually regard this as "motivated behavior" and reserve the term *emotion* for cases in which "feelings" are involved. However, I think most authors in this volume would agree that emotional feelings cannot so easily be "distinguished from cognitive or volitional states of consciousness," as the above definition assumes. Alas, the above is the beginning of clarity, not its achievement. One can have emotion without feeling the emotion—as in "I didn't know I was angry until I over-reacted like that"—and one can certainly have feelings—as in "I feel that the color does

not suit you"—that do not seem emotional. My suggestion, however, is that the potential for feeling is essential to the concept of emotion, as I will try to make more clear when I set the stage for a new look at biological evolution below. This is clearly a work in progress.

THE NARRATIVE CHRONOLOGICALLY REARRANGED AND ANNOTATED

In this section, I rearrange the narrative of the previous section, annotating it to make clear the hugely cognitive content of most of the emotional states reported, thus setting down the gauntlet to any theory that jumps too quickly from a basic system for "motivation" to the human brain's capacity to integrate rich cognitions with subtle emotions. How do we get from a basic system of neuromodulators (Kelley, Chapter 3), reward and punishment (Rolls, Chapter 5) or behavioral fear (Fellous & LeDoux, Chapter 4) to these nuanced emotions that humans experience? This question is addressed below, under An Evolutionary Approach to Heated Appraisals. The following section will present evolutionary stories (or essays in comparative cognitive neuroscience) for the diversity of vision and for the expansion of communication to include language in the perspective offered later, under From Drives to Feelings. These insights will then serve to anchor a fresh look at the issue of robot emotions under Emotion without Biology.

Following each excerpt from the narrative, I offer a sequence of mental states and events, without teasing out the overlapping of various segments. The notion $\boxed{x}$ will denote the experience of the emotional state x, by which I mean a reportable experience in which emotional feelings play an important role. The challenge (to be only partially met below) is to understand the iceberg of which such experiences are but the tip. Each sequence is followed by a few comments relevant to any later attempt to explain the underlying neural dynamics. These comments constitute an *a posteriori* reconstruction of what went on in my head at the time: there is no claim that the suggested interpretations are complete, and similar emotional behaviors and experiences might well have quite different causes in different circumstances.

1. *I phoned N's secretary and learned that she had forgotten to confirm the meeting with N. I was not particularly upset, we rescheduled the meeting for 4 P.M. the next day, and I proceeded to make contented use of the unexpected free time to catch up on my correspondence.*

$$\boxed{\text{Hope for success of the meeting}} \rightarrow \text{Expected meeting not confirmed} \rightarrow \boxed{\text{mild annoyance}} \rightarrow \text{meeting rescheduled; good use of free time} \rightarrow \boxed{\text{annoyance dissipated}};\ \boxed{\text{contentment}} \quad (1)$$

Presumably, the fact that the annoyance is mild rests on cognitive computations showing that the meeting is not urgent; this mild reaction was further defused when it was found that the meeting could be rescheduled prior to any pressing deadline.

2. *On Wednesday, my secretary took the day off because her mother was ill. When I heard the news that morning, I felt slightly frustrated that some correspondence would be delayed, but equally concerned for the mother's health and had no question about the appropriateness of my secretary's action. The matter, having been accepted, had—as far as the next few hours were concerned—no further effect upon my mood.*

Absence of secretary → realization of delayed work → [mild annoyance] → mother's ill health; absence is appropriate → [concern for mother]; [annoyance dissipated] (2)

This diagrams the transition in emotions as serial, when it was probably a parallel process in which annoyance and concern were simultaneously activated. What seems to unite the sequences in (1) and (2) is that the blocking of the performance of a plan yields [annoyance]. What determines the intensity of affect is a point to which I return in (5). What is important here is that the emotional state can continue, coloring and modifying a variety of cognitive states until it is in some way resolved. The resolution in (1) can be put down to the formulation of an alternative plan (a new appointment); the resolution in (2) is more complex, accepting that circumstances require a delay. The experience of [concern] is separate but helps to mitigate the [annoyance] by offering an acceptable reason for this particular change of plan. Moreover, I assumed my secretary would care for her mother, so this concern dissipated in turn.[3]

3. *At midday, I was transiently and mildly upset by the cancellation of a luncheon appointment.*

Luncheon cancelled → [mild disappointment] → other activity → [disappointment dissipated] (3)

Why disappointment and not annoyance? The former shades toward resignation; the latter shades toward anger and the possibility of impulsive action.

Why did this new cancellation not have the aggravative effect of the second disappointment with N, to be recounted in (5)? The notion is that an emotional state may be terminated

when a plan is completed which addresses the source of that state; or it may simply dissipate when alternative plans are made. However, if some major goal has been rendered unattainable by some event, the negative emotion associated with the event may not be dissipated by embarking on other plans since they do not approach the goal. Another issue is that our memory of an episode may be more or less charged with the emotional state which was part of the event. On one occasion, one might recall the death of a loved one with real grief; on another occasion, that death is recalled without any trace of sadness. Thus, our emotional state on any one occasion can be modulated by the emotional states evoked by the more or less conscious recall of associated episodes. (I do not claim to have any theory of how memory and emotion interact or of the variability of these interactions; but see, e.g., Christianson, 1992.)

4. *I decided in mid-afternoon to put together a chart to discuss with N at our meeting, but the printer gave me some problems, and it was already after 4 when I left my office for the meeting, feeling somewhat flustered, but glad that I had a useful handout and pleased that I was only a few minutes late. I was looking forward to the meeting, and experienced no regret or frustration that it had been postponed the day before.*

Putting together a chart → using printer and having problems → running late for meeting → feeling flustered → but only mildly late and with a good handout → glad; pleasant anticipation (4)

Presumably, the extra time to prepare the chart provides another reason to dispel residual annoyance (if any) from Tuesday's cancellation; but this opportunity was not realized in time to affect my mental state on Tuesday. Consequences of a situation may not be realized until long afterward. Note that the lack of regret or frustration is not part of my mental state at this point. Rather, it is part of the narrative composed later and was designed to highlight what happened next.

5. *However, when I arrived at about 4:06, N's secretary wasn't there and neither was N. Another secretary explained that N's secretary had left a message for me earlier in the day to move the meeting forward to 2:30. (I had not received the message because of my secretary's absence that day.) N was not available. Would I care to make an appointment? With a curt "No," I turned on my heels, and stormed out of the office and back to my own.*

Absence of N and N's secretary → [mild disappointment] → news that message had been left that meeting had been cancelled → [fury] → curt response to offer to set new appointment; abrupt return to my office (5)

It is perhaps worth considering to what extent the "over-the-top" level of annoyance here labeled "fury" was targeted rather than diffuse. It was *not* targeted at my secretary—the recollection of her absence served to explain why I had not received the message (it had been left on the voicemail, which she would normally relay to me), not to blame her for being away. It was the cancellation of the meeting, not the loss of the message, that annoyed me; and this fury was directed at N and his secretary. However, in their absence, the "bearer of bad tidings" received the immediate brunt of my anger. I do not think anyone unconnected with this news would have received an overt action beyond seeing the facial expression of this strong negative emotion.

Note the immense difference from (1). This dramatic overreaction is not a response to the cancellation alone but seems to be a case of "state dependence" (Blaney, 1986). In this case, the cumulative effect of earlier negative emotional states was "to blame" (recall the earlier comment that the present emotional state can be modulated by the emotional states evoked by the recall of associated episodes).

6. *As I walked, I simultaneously felt fury at the double cancellation and shame at my own rude behavior, as well as realizing that another appointment had to be made. I felt tight and constricted. After a minute or two in my office, unable to concentrate and with my thoughts dominated by this stew of emotions, I decided to return to N's office.*

[fury] → realization that behavior was inappropriate → [fury mixed with shame] → recognition that an apology is due to the secretary and that another appointment must be made (6)

The emotional state of fury provides a strong drive for a set of violent behaviors. Here, we see the internal battle between the acting out of this aggression and the social imperatives of "correct" behavior. This provides another example of the competition and cooperation that is so distinctively the computing style of the brain (Arbib, 1989). Note the role here of social norms in judging the behavior to be inappropriate with the concomitant emotion of shame, providing the motivation to take a

course of action (apology) that will make amends for the inappropriate behavior; note that the situation was simplified because this course of action was consistent with the (nonemotional?) recognition that another appointment had to be made. Indeed, many of our action plans are social and require interaction with others to enlist aid, invite sympathy, bargain, intimidate, threaten, provide rewards, etc.

The analysis can continue like this until the end of the narrative. This section has provided a (very restricted) data set on the interaction between perception, emotion, and action that has brought out the interaction between cognitive processing that stresses both the "heat" and the state dependence involved in emotions. Below, I will attempt to make sense of this data set within an evolutionary framework (see An Evolutionary Approach to Heated Appraisals). To set the stage for this, in the following section we present a general evolutionary framework, then an analysis of vision and language within that framework, and finally a look at the motivational systems which ground the emotions.

HUGHLINGS JACKSON: AN EVOLUTIONARY FRAMEWORK

I now offer a general framework for the study of the evolution of brain mechanisms which will inform the following two sections. Hughlings Jackson was a 19th century British neurologist who viewed the brain in terms of levels of increasing evolutionary complexity (Jackson, 1878–79). Influenced by the then still novel Darwinian concepts of evolution, he argued that damage to a "higher" level of the brain disinhibited "older" brain regions from controls evolved later, to reveal evolutionarily more primitive behaviors. My arguments in this chapter will be structured by my attempt (Arbib, 1989) to extract computational lessons from Jackson's views on the evolution of a system that exhibits hierarchical levels.

- The process starts with one or more basic systems to extract useful information from a particular type of sensory input.
- These basic systems make data available which can provide the substrate for the evolution of higher-level systems to extract new properties of the sensory input.
- The higher-level systems then enrich the information environment of the basic systems by return pathways.
- The basic systems can then be adjusted to exploit the new sources of information.

Thus, evolution yields not only new brain regions connected to the old but also reciprocal connections which modify those older regions. The resulting system is not unidirectional in its operation, with lower levels simply providing input to higher levels. Rather, there is a dynamic equilibrium of multiple subsystems of the evolved system that continually adjust to significant changes in each other and (more or less directly) in the world.

The following section will offer a "Jacksonian" analysis of the evolution of brain mechanisms for vision and—via mechanisms for the visual control and recognition of hand movements—language, rooted in a brief comparison of frogs, rats, monkeys, and humans.

The usual caveats: (a) Frog → Rat → Monkey → Human is not an evolutionary sequence; rather, the first three species are used as reference points for stages in human evolution within the mammalian lineage. (b) There is no claim that human vision is inherently better than that of other species. The question is, rather, how it has adapted to our ecological niche. Human vision (to say nothing of human bodies) is ill-suited to, for instance, making a living in a pond by catching flies. We will turn to (a suitable abstraction of) the notion of "ecological niche" when we return to our discussion of in what senses may robots have emotions.

The relevance of vision and language to our account of the evolution of emotion and its underlying brain mechanisms (see below, From Drives to Feelings) is as follows:

1. We apply the term *vision* for the processing of retinal signals in all these creatures. However, vision in the frog is midbrain-dominated and specially adapted to a limited repertoire suitable for species survival, whereas mammals augment these midbrain mechanisms with a rich set of cortical mechanisms that make possible a visual repertoire which becomes increasingly open-ended as we pass from rat to monkey to human. In other words, we see an evolutionary change in vision which is qualitative in nature. Yet, the ancestral mechanisms remain an integral part of the human visual system.
2. All these creatures have communication in the sense of vocal or other motor signals that can coordinate behavior between conspecifics. Yet, none of these communication systems forms a language in the human sense of an open-ended system for expressing novel as well as familiar meanings. The closest we can come, perhaps, is the "language" of bees, but this is limited to messages whose novelty lies in the variation of three parameters which express the quality, heading, and distance of a food source. We again see in human evolution a qualitative change

but this time with a special terminology to express it—from *communication* as a general phenomenon among animals to *language* as a specific form of communication among humans.

Note that it is not my point to say that English has made the right absolute choices. I am simply observing that we do use the term *vision* for the processing of focused patterns of variations in light intensity whether in fly, octopus, bird, or human. In each creature, vision is served by a network of interacting mechanisms, but this network varies from species to species in ways related to the animal's ecological niche. What humans lack in visual acuity they may make up for in visual processes for face recognition and manual control. By contrast, although some people will use the term *language* loosely for any form of communication, most people will understand the sense of the claim that "Humans are the only creatures that have language;" yet, none would accept the claim that "Only humans have vision" unless *vision* were used in the metaphorical sense of "the ability to envision explicitly alternative futures and plan accordingly."

3. Again, all these creatures are endowed with motivational systems—hunger, thirst, fear, sex, etc.—and we may trace the linkage of these systems with developing cortical mechanisms as we trace our quasi-evolutionary sequence. However, are we to follow the analogy with vision and refer to the processes involved as "emotion" throughout, or should we instead follow the example of communication and suggest that emotion provides in some sense a special form of motivational system? Here, I do not think there is the same consensus as there is for *vision* or *language*. I choose the latter path, suggesting the following:

$$\text{Motivation} \approx \text{Communication}$$
$$\text{Emotion} \approx \text{Language}$$

There is this difference: whereas language seems to be restricted to humans alone, emotion seems to be less clear-cut. Following Darwin, we see expressions of emotion in dogs and monkeys, even if most people would not credit them with the capability for the emotional nuances of a Jane Austen heroine.

4. Much of the biological discussion of emotion has turned on the distinction between "emotional behavior" and "emotional feelings." "Emotional expression" adds another dimension, where mammalian (especially primate) facial expressions are understood to signal the emotional state of the animal but distinguished from the emotional behavior itself (e.g., a fearful expression is

different from such fear behavior as fleeing or freezing). Emotional feelings are tied up with notions of consciousness, but it is well known that one may be conscious of the possible emotional overtones of a situation yet not feel emotionally involved oneself (and brain damage may leave a person incapable of emotional feelings; cf. the Chapter 3 section "A Modern Phineas Gage" in Damasio, 1994).

Below, we will discuss the notion of emotion as suitable for characterizing aspects of the behavior and inner workings of robots that share with humans neither an evolutionary history as flesh-and-blood organisms nor the facial or vocal expressions which can ground empathy. In particular, we will return to the question of ecological niches for robots and the issue of to what extent emotions may contribute to, or detract from, the success of a "species" of robots in filling their ecological niche.

Elsewhere (e.g., Arbib, 1989), I have developed a theory of schemas as functional (as distinct from structural) units in a hierarchical analysis of the brain. Extant schemas may be combined to form new schemas as coordinated control programs linking simpler (perceptual and motor) schemas to more abstract schemas which underlie thought and language more generally. The behavioral phenotype of an organism need not be linked to a localized structure of the brain but may involve subtle patterns of *cooperative computation* between brain regions which form a schema. Selection may thus act as much on schemas as it does on localized neural structures. Developing this view, Arbib and Liaw (1995) argued that evolution yields not only new *schemas* connected to the old but also reciprocal connections which modify those older schemas, linking the above Jacksonian analysis to the language of schema theory.

EVOLUTION OF THE BRAIN MECHANISMS SUPPORTING VISION AND LANGUAGE

Over the years, I have attempted to create a comparative computational neuroethology (i.e., a comparative computational analysis of neural mechanisms underlying animal behavior) in which the brains of humans and other creatures come to be better understood by seeing homologous mechanisms as computational variants which may be related to the different evolutionary history or ecological niche of the creatures that contain them. Arbib (2003) stresses the notion of "*conceptual* neural evolution" as a way of understanding complex neural mechanisms through incremental modeling. Although somewhat *ad hoc*, this process of adding features to a model "to see

what happens" is constrained by biological data linking behavior to anatomy and neurophysiology, though without a necessary analysis of the underlying genes. The aim is to discover relations between modules (neural circuits at some grain of resolution) that implement basic schemas (functions, as distinct from structures) in simpler species with those that underlie more elaborate schemas in other species. Clearly, the evolutionary path described in this way is not necessarily substantiated as the actual path of evolution by natural selection that shaped the brains of the species we study today but has two benefits: (1) making very complex systems more comprehensible and (2) developing hypotheses on biological evolution for genetic analysis. In 2003 I offered a conceptual evolutionary perspective on brain models for frog, rat, monkey, and human. For rat, I showed how a frog-like taxon-affordance model (Guazzelli, Corbacho, Bota, & Arbib, 1998) provides a basis for the spatial navigation mechanisms that involve the hippocampus and other brain regions. (As in Chapters by Rolls and Kelley, *taxis* [plural *taxes*] are simple movements in response to a set of key stimuli. *Affordances* (Gibson, 1966) are parameters for motor interactions signaled by sensory cues without the necessary intervention of "high-level processes" of object recognition.) For monkey, I recalled two models of neural mechanisms for visuomotor coordination. The first, for saccades, showed how interactions between the parietal and frontal cortex augment the superior colliculus, seen as the homolog of the frog tectum (Dominey & Arbib, 1992). The second, for grasping, continued the theme of parietofrontal interactions, linking parietal affordances to motor schemas in the premotor cortex (Fagg & Arbib, 1998). This further emphasized the mirror system for grasping, in which neurons are active both when the monkey executes a specific grasp and when it observes a similar grasp executed by others. The model of human brain mechanisms is based on the mirror-system hypothesis of the evolution of the language-ready brain, which sees the human Broca's area as an evolved extension of the mirror system for grasping. In the next section, I will offer a related account for vision and next note how dexterity involves the emergence of new types of visual system, carrying forward the mirror-system hypothesis of the evolution of the language-ready brain. The section ends with a brief presentation of a theory of how human consciousness may have evolved to have greater linkages to language than animal awareness more generally. However, these sections say nothing about motivation, let alone emotion. Thus, my challenge in the section From Drives to Feelings is to use these insights to both apply and critique the evolutionary frameworks offered in Chapters 3–5 by Kelley, Rolls, and Fellous & LeDoux and thus to try to gain fresh insight into the relations between emotion and motivation and between feelings and behavior. The mirror-system hypothesis, with its emphasis on communication, provides one example of how we may link this brain-in-the-individual

approach to the social interactions stressed by Adolphs (Chapter 2). Indeed, Jeannerod (Chapter 6) explores the possible role of mirror systems in empathy and our ability to understand the emotions of others. However, I must confess here that the current chapter will place most emphasis on the brain-in-the-individual approach and will conclude by giving a theory of robot emotions grounded in the analysis of a robot going about its tasks in some ecological niche, rather than emphasizing social interactions.

Vision Evolving

The year 1959 saw the publication of two great papers on the neurophysiology of vertebrate vision: the study by Lettvin, Maturana, McCulloch, and Pitts (1959) of feature detectors in the frog's retina and that by Hubel and Wiesel (1959) of receptive fields of neurons in the cat primary visual cortex. We will analyze the first work in relation to later studies of frog behavior (postponing a brief look at the role of motivation; we will then look at the more generic coding in the cat visual system and ponder its implications.

Action-Oriented Feature Detectors in Frog Retina

Lettvin, Maturana, McCulloch, and Pitts (1959) studied "what the frog's eye tells the frog's brain" and reported that frog ganglion cells (the output cells of the retina) come in four varieties, each providing a retinotopic map of a different feature to the tectum, the key visual region of the midbrain (the homolog, or "evolutionary cousin," of what in mammals is often referred to as the "superior colliculus"):

1. The boundary detectors
2. The movement-gated, dark convex boundary detectors
3. The moving or changing contrast detectors
4. The dimming detectors

Indeed, axons of the cells of each group end in a separate layer of the tectum but are in registration: points in different layers which are stacked atop each other in the tectum correspond to the same small region of the retina. All this shows that the function of the frog retina is not to transmit information about the point-to-point pattern distribution of light upon it but rather to analyze this image at every point in terms of boundaries, moving curvatures, changing contrasts, and local dimming. Lettvin's group argues that the convexity detectors (operation 2 above) serve as "bug perceivers," while operation 4 could be thought of as providing "predator detectors."

However, this is only the first approximation in unraveling the circuits which enable the frog to tell predator from prey. Where Lettvin's group emphasized retinal fly and enemy detectors, later work emphasized tectal integration (Grüsser-Cornehls & Grüsser, 1976) and interactive processes involving the optic tectum and the thalamic pretectal region (Ewert, 1987). Cobas and Arbib (1992) defined the perceptual and motor schemas involved in prey catching and predator avoidance in frog and toad, charting how differential activity in the tectum and pretectum could play upon midbrain mechanisms to activate the appropriate motor schemas:

Prey capture: orient toward prey, advance, snap, consume
Predator avoidance: orient away from predator, advance

Note that the former includes "special-purpose" motor pattern generators, those for snapping and ingestion, while the latter uses only "general-purpose" motor pattern generators for turning and locomotion.

Generic Feature Detectors in Cat Primary Visual Cortex

In 1959, Hubel and Wiesel published "Receptive fields of single neurones in the cat's striate cortex." A whole string of further papers (such as Hubel & Wiesel, 1962, 1965, 1968; Wiesel & Hubel, 1963; Hubel, Wiesel, & LeVay, 1977) extended the story from cat to monkey, placed the neurophysiology in an anatomical and developmental framework, and introduced the crucial notions of orientation and ocular dominance columns in visual cortex—a cumulative achievement honored with a Nobel Prize in 1981. Where Kuffler (1953) had characterized retinal ganglion cells in cat as on-center off-surround and off-center on-surround, Hubel and Wiesel showed that cells in the primary visual cortex of cat (and monkey) could be classified as "simple" cortical cells, responsive to edges at a specific orientation in a specific place, and "complex" cells, which respond to edges of a given orientation in varying locations. Paralleling the work of Mountcastle and Powell (1959) on somatosensory cortex, Hubel and Wiesel found that the basic unit of mammalian visual cortex is the hypercolumn, 1 mm^2 × 2 mm deep. Each such hypercolumn contains columns responsive to specific orientations. The columns form an overarching retinotopic map, with fine-grained details such as orientation available as a "local tag" at each point of the map. Overlaid on this is the pattern of ocular dominance "columns" (really more like zebra stripes when viewed across the cortical surface), alternate bands each dominated by input from a single eye.

How are we to reconcile the "ecologically significant" features extracted by the frog retina with the far more generic features seen in cats and primates

at the much higher level of visual cortex? Different animals live in different environments, have different behaviors, and have different capabilities for motor behavior. As a result, the information that they need about their world varies greatly. On this basis, we may hope to better understand the problem of vision if we can come to see which aspects of visual system design converge and which differences are correlated with the differing behavioral needs of different species. The frog will snap at, or orient toward, an object moving in prey-like fashion and will avoid a large moving object. It responds to localized features of the environment—information from a large region of its visual field only affects its action when determining a barrier it must avoid when seeking prey or escaping an enemy, and this is mediated elsewhere in the brain. Thus, preprocessing at the ganglion cell level in the frog is already action-oriented. In the cat (and monkeys and humans), processing in the primary visual cortex is "action-neutral," providing efficient encoding of natural stimuli and serving as a precursor to processes as diverse as face recognition and manual dexterity. Specializations appropriate to certain crucial tasks do occur but only further along the visual pathway.

The Where, What, and How of Vision

Until the late 1960s, the study of the visual system of mammals emphasized the contributions of the visual cortex, with little attention paid to midbrain mechanisms. An important move toward a more subtle understanding came with the symposium contributed to by Ingle, Schneider, Trevarthen, and Held (1967), who suggested that we should think of vision not in terms of a single pathway running through the lateral geniculate nucleus to the visual cortex (the geniculostriate pathway) but rather in terms of the interaction of two pathways: the geniculostriate system for identifying and a midbrain system, the superior colliculus or tectum, for locating (see Schneider, 1969, for relevant data on the hamster). It thus became fashionable to talk about the "two visual systems" in mammals, one for *what* and one for *where*.

However, analysis of the frog (e.g., Arbib, 1987, for a review) showed that there could be more than two visual systems even subcortically, with different parts of the brain serving different visual mechanisms. For example, prey catching by the frog seems to rely on the tectum for processing of visual cues. The pretectum seems necessary for the tectum to play its role in the avoidance of visual threat, as well as in mediating the recognition of barriers. The role of the tectum in directing whole-body movements in the frog is analogous to the role of the superior colliculus in directing eye movements in the cat and monkey. When humans without primary visual cortex are asked "Am I moving my left or right hand?" they say "I can't see" but,

asked to make a guess, will point in the direction of the moving hand. They can catch a ball even though they believe they cannot see it. This phenomenon is referred to as *blindsight* (Weiskrantz, Warrington, Sanders, & Marshall, 1974; see Stoerig, 2001, for a review and Humphrey, 1970, for a study linking frog and monkey). The midbrain visual system is thus quite powerful but not connected to consciousness. Indeed, when a normal person catches a ball, he or she is usually aware of seeing the ball and of reaching out to catch it but certainly not of the processes which translate retinal stimulation into muscle contraction, so most neural net activity is clearly unconscious. The lesson is that even schemas that we think of as normally under conscious control can in fact proceed without our being conscious of their activity.

Recent research has extended the *what* and *where* dichotomy to a variety of cortical systems. Studies of the visual system of monkeys led Ungerleider and Mishkin (1982) to distinguish inferotemporal mechanisms for object recognition (*what*) from parietal mechanisms for localizing objects (*where*). Goodale, Milner, Jakobson, and Carey (1991) studied a human patient (D. F.) who had developed a profound visual form of agnosia following a bilateral lesion of the occipito-temporal cortex. The pathways from the occipital lobe toward the parietal lobe appeared to be intact. When the patient was asked to indicate the width of any one of a set of blocks either verbally or by means of her index finger and thumb, her finger separation bore no relationship to the dimensions of the object and showed considerable trial-to-trial variability. Yet, when she was asked simply to reach out and pick up the block, the peak aperture between her index finger and thumb (prior to contact with the object) changed systematically with the width of the object, as in normal controls. A similar dissociation was seen in her responses to the orientation of stimuli. In other words, D. F. could preshape her hand accurately, even though she appeared to have no conscious appreciation (either verbal or by pantomime) of the visual parameters that guided the preshape. With Goodale and Milner (1992), then, we may rename the *where* pathway as the *how* pathway, stressing that it extracts a variety of affordances relevant to action (recall that affordances are parameters for motor interactions extracted from sensory cues), not just object location.

The Many Systems of Vision

This brief tour of the neural mechanisms of vertebrate vision, and a great body of related modeling and empirical data, supports the enunciation of a general property of vertebrate neural control: a multiplicity of different representations must be linked into an integrated whole. However, this may be

mediated by distributed processes of competition and cooperation. There need be no one place in the brain where an integrated representation of space plays the sole executive role in linking perception of the current environment to action.

Dean, Redgrave, and Westby (1989; see also Dean & Redgrave, 1989) used a study of the rat informed by findings from the study of the frog to provide an important bridge between frog and monkey. Where most research on the superior colliculus of cat and monkey focuses on its role in saccadic eye movements—an approach behavior for the eyes—Dean et al. looked at the rat's own movements and found two response systems in the superior colliculus which were comparable with the approach and avoidance systems studied in the frog and toad. We thus see the transition from having the superior colliculus itself commit the animal to a course of action (frog and rat) to having it more often (but not always) relinquish that role and instead direct attention to information for use by cortical mechanisms in committing the organism to action (e.g., cat, monkey, and human). We now turn to one system for committing the organism to action, that for grasping, and then present an evolutionary hypothesis which links cerebral mechanisms for grasping to those that support language.

The Mirror System and the Evolution of Language

Having looked at vision from a very general perspective, I now focus on two very specific visual systems that are especially well developed in primates: the system that recognizes visual affordances for grasping and the system that recognizes grasping actions made by others. I shall then argue that these systems provide the key to a system that seems specifically human: the brain mechanisms that support language.

Brain Mechanisms for Grasping

In macaque monkeys, parietal area AIP (the anterior region of the intraparietal sulcus; Taira et al., 1990) and ventral premotor area F5 (Rizzolatti et al., 1988) anchor the cortical circuit which transforms visual information on intrinsic properties of an object into hand movements for grasping it. The AIP processes visual information on objects to extract affordances (grasp parameters) relevant to the control of hand movements and is reciprocally connected with the so-called canonical neurons of F5. Discharge in most grasp-related F5 neurons correlates with an action rather than with the indi-

vidual movements that form it so that one may relate F5 neurons to various motor schemas corresponding to the action associated with their discharge.

The FARS model (named for Fagg, Arbib, Rizzolatti & Sakata; Fagg & Arbib, 1998) provides a computational account centered on the pathway:

AIP (object affordances) → ($F5_{canonical}$ (abstract motor schemas)
→ F1 (motor cortex instructions to lower motor areas and motor neurons)

Figure 12.1 gives a view of "FARS Modificato," the FARS model updated on the basis of suggestions by Rizzolatti and Luppino (2003), based on the neuroanatomical data reviewed Rizzolatti and Luppino (2001), so that information on object semantics and the goals of the individual influences AIP rather than F5 neurons, as was the case in Fagg and Arbib (1998). The dorsal stream via the AIP does not know *what* the object is; it can only see the object as a set of possible affordances (it lies on the *how* pathway). The ventral stream (from primary visual cortex to inferotemporal cortex), by contrast, is able to recognize what the object is. This information is passed

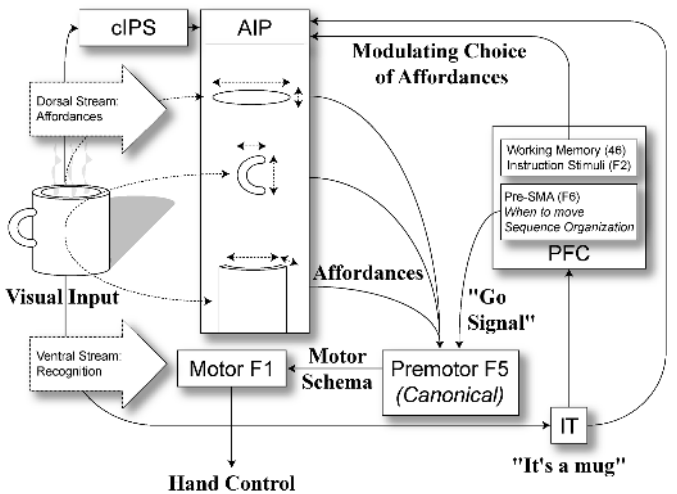


Figure 12.1. A reconceptualization of the FARS model (Fagg & Arbib, 1998), in which the primary influence of the prefrontal cortex (PFC) on the selection of affordances is on the parietal cortex (AIP, anterior intraparietal sulcus) rather than the premotor cortex (hand area F5). This diagram includes neither the circuitry encoding a sequence, possibly the part of the supplementary motor area called the pre-SMA (Rizzolatti, Luppino, & Matelli, 1998), nor the administration of the sequence (inhibiting extraneous actions, while priming imminent actions) by the basal ganglia.

to the prefrontal cortex, which can then, on the basis of the current goals of the organism and the recognition of the nature of the object, bias the AIP to choose the affordance appropriate to the task at hand. Figure 12.1 gives only a partial view of the FARS model, which also provides mechanisms for sequencing actions. It segregates the F5 circuitry, which encodes unit actions from the circuitry encoding a sequence, possibly the part of the supplementary motor area called "pre-SMA" (Rizzolatti, Luppino, & Matelli, 1998). The administration of the sequence (inhibiting extraneous actions, while priming imminent actions) is then carried out by the basal ganglia (Bischoff-Grethe, Crowley, & Arbib, 2003).

Bringing in the Mirror System

Further study revealed a class of F5 neurons that discharged not only when the monkey grasped or manipulated objects but also when the monkey observed the experimenter make a gesture similar to the one that, when actively performed by the monkey, involved activity of the neuron (Rizzolatti, Fadiga, Gallese, & Fogassi, 1995). Neurons with this property are called "mirror neurons." The majority of mirror neurons are selective for one type of action, and for almost all mirror neurons there is a link between the effective observed movement and the effective executed movement.

Two positron emission tomography (PET) experiments (Rizzolatti et al., 1996; Grafton, Arbib, Fadiga, & Rizzolatti, 1996) were then designed to seek mirror systems for grasping in humans. Grasp observation significantly activated the superior temporal sulcus (STS), the inferior parietal lobule, and the inferior frontal gyrus (area 45). All activations were in the left hemisphere. The last area is of especial interest—areas 44 and 45 in the left hemisphere of the human brain constitute Broca's area, a major component of the language mechanisms. Indeed, F5 is generally considered to be the homolog of Broca's area.

And on to Language

The finding that human Broca's area contains a mirror system for grasping led us (Arbib & Rizzolatti, 1997; Rizzolatti and Arbib, 1998) to explore the hypothesis that the mirror system provided the basis for the evolution of human language via seven stages:

1. Grasping.
2. A mirror system for grasping.

3. A "simple" imitation system: we hypothesize that brain mechanisms supporting a simple imitation system—imitation of novel object-directed actions through repeated exposure—for grasping developed in the 15 million-year evolution from the common ancestor of monkeys and apes to the common ancestor of apes and humans.
4. A "complex" imitation system: we hypothesize that brain mechanisms supporting a complex imitation system—acquiring (longer) novel sequences of more abstract actions in a single trial—developed in the 5 million-year evolution from the common ancestor of apes and humans along the hominid line that led, in particular, to *Homo sapiens*.
5. *Protosign*, a manual-based communication system, resulting from the freeing of action from praxis to be used in pantomime and then in manual communication more generally.
6. *Protospeech*, a vocal-based communication system exploiting the brain mechanisms that evolved to support protosign.
7. Language. Arbib (2002) argues that stages 6 and 7 are separate, characterizing protospeech as being the open-ended production and perception of sequences of vocal gestures, without implying that these sequences have the syntax and semantics adequate to constitute a language. But the stages may be interleaved.

Nonhuman primates have a call system and orofacial gestures expressive of a limited range of emotional and related social indicators. However, we do not regard primate calls as the direct precursor of speech. Combinatorial properties for the openness of communication are virtually absent in basic primate calls, even though individual calls may be graded. Moreover, the neural substrate for primate calls is in a region of the cingulate cortex distinct from F5. The mirror-system hypothesis offers detailed reasons why Broca's area—as the homologue of F5—rather than the area already involved in vocalization, provided the evolutionary substrate for language.

Consciousness, Briefly

We have now established that vision is no single faculty but embraces a wide variety of capabilities, some mediated by subcortical systems, others involving cooperation between these and other, more highly evolved systems in the cerebral cortex. The evolution of manual dexterity went hand in hand [!] with the evolution of a dorsal cortical pathway dedicated to extracting the visual affordances appropriate to that dexterity and a ventral cortical

pathway of much more general capability, able to recognize objects and relationships in a fashion that enables the prefrontal cortex to support the planning of action, thus determining which affordances to exploit in the current situation. The mirror-system hypothesis suggests how the recognition of manual actions might have paved the way for cumulative evolutionary changes in body and brain to yield early humans with the capability for complex imitation, protosign, and protospeech.

I would argue that we are conscious in a fully human sense only because we have language—i.e., that as awareness piggybacks on all manner of neural functions, so too must it piggyback on language, thus reaching a subtlety and complexity that would otherwise be impossible. However, I strongly deny that consciousness is merely a function of language. For example, one can be aware of the shape and shading and coloration of a face in great subtlety and be totally unable to put one's vivid, conscious perception of that face into words. Moreover, I view consciousness as a system function that involves networks including, but not necessarily limited to, the cerebral cortex and that as the cerebral cortex evolves, so too does consciousness.

Arbib and Hesse (1986; Arbib, 1985) suggest that the key transition from the limited set of vocalizations used in communication by, say, vervet monkeys to the richness of human language came with a migration in time from an execution/observation matching system, enabling an individual to recognize the action (as distinct from the mere movement) that another individual is making, to the individual becoming able to pantomime "this is the action I am about to take" (see Arbib, 2001, for an exposition of the Arbib-Hesse theory within the mirror-system framework.) Arbib and Hesse emphasize the changes within the individual brain made possible by the availability of a "précis"—a gesturable representation—of intended future movements (as distinct from current movements). They use the term *communication plexus* for the circuits involved in generating this representation. The Jacksonian element of their analysis is that the evolution of the communication plexus provides an environment for the further evolution of older systems. They suggest that once the brain has such a communication plexus, a new process of evolution begins whereby the précis comes to serve not only as a basis for communication between the members of a group but also as a resource for planning and coordination within the brain itself. This communication plexus thus evolves a crucial role in schema coordination. The thesis is that it is the activity of this coevolved process that constitutes consciousness. As such, it will progress in richness along with the increased richness of communication that culminates as language in the human line. Since lower-level schema activity can often proceed successfully without this highest-level coordination, consciousness may sometimes be active, if active at all, as a monitor

rather than as a director of action. In other cases, the précis of schema activity plays the crucial role in determining the future course of schema activity and, thus, of action.

FROM DRIVES TO FEELINGS

Our conceptual evolutionary analysis has allowed us to tease apart a variety of visual mechanisms and relate them to a range of behaviors, from the feeding and fleeing of the frog to the visual control of hand movements in monkeys and humans. We briefly examined accounts of the evolution of language and a particularly human type of consciousness. We saw that this type of consciousness builds upon a more general form of consciousness—awareness of both internal and external states—that we did not explain but for which we made a crucial observation: activities in regions of the cerebral cortex can differ in their access to awareness. However, although we looked at visual processing involved in what some might label two "emotional behaviors" in the frog—feeding and fleeing—we did not explicitly discuss either motivation or emotion, beyond suggesting that nonhumans may be aware of subtle social cues or the difference between feeling maternal and feeling enraged and noting that nonhuman primates have a call system and orofacial gestures expressive of a limited range of emotional and related social indicators. The time has come to put these insights to work. As stated above, some would use the term *emotion* to cover the whole range of motivated behavior, whereas others (myself included) stress the emergent subtlety of emotions. The following section will briefly review an account of motivated behavior in toads and rats, then use this basis together with the insights from the Jacksonian analysis above to offer an integrated perspective on the evolutionary insights provided by Kelley, Rolls, and Fellous & LeDoux in Chapters 3–5.

Basic Models of Motivation

Karl Pribram (1960) has quipped that the limbic system is responsible for the "four Fs:" Feeding, Fighting, Fleeing, and Reproduction. It is interesting that three of the four have a strong social component. In any case, the notion to be developed in this section is that the animal comes with a set of basic drives—for hunger, thirst, sex, self-preservation, etc.—and that these provide the basic motor, or motivation, for behavior. This will then ground our subsequent discussion of motivation.

The Motivated Toad

Earlier, we examined the basic, overlapping circuitry for fleeing and feeding in the toad but did not show how this circuitry could be modulated by motivational systems. Toads can be conditioned to treat objects that normally elicit escape behavior as prey. Toads which are allowed to eat mealworms out of an experimenter's hand can be conditioned to respond to the moving hand alone (Brzoska & Schneider, 1978). Heatwole and Heatwole (1968) showed that the upper size threshold for acceptable prey increases with long-term food deprivation while the lower size threshold remains constant. In spring, prey-catching behavior decreases or fails to occur. Indeed, prey recognition is "exchanged" for female recognition, introducing mating behavior. Moving females release orienting, approaching, and clasping behaviors in the male (Heusser, 1960; Kondrashev, 1976). In many species, males will attempt to clasp practically any moving object including other males during mating season (Wells, 1977).

Betts (1989) carried forward the modeling of tectal–pretectal interactions (reviewed in Arbib, 1987) to include the effects of motivation. The basic idea is that perceptual schemas can be modulated by broadcast signals for drive levels such as those for feeding or mating, thus shifting the balance of behavior. For example, standard models of prey catching address the finding that ablation of the pretectum results in disinhibition of prey catching, with animals snapping at objects much larger than normal prey (Ewert, 1984); modulating the level of pretectal inhibition can thus shift the balance between feeding and fleeing. Betts (1989) further suggested parallels between the effects of pretectal ablation and the conditioning results and changes that occur during the mating season. T5 neurons in the tectum have a variety of responses, including those which we classify as prey recognition. Betts modeled T5 neuron function by distinguishing the T5 base of T5 cells within the tectum from the pretectal inhibition which modulates it. A T5 neuron then has the potential to be, for example, either a prey or mate feature detector depending on this inhibition. Betts suggests that the subclasses of T5 neurons with different detector properties should be regarded as more or less stable states of a modulated system capable of adaptability and changeability.

In summary, the frog has basic systems for feeding, fleeing, and mating; but we can distinguish circuitry that provides basic "subroutines" from circuitry that uses motivational factors to bias their deployment. This separates the motivation from the behavior. From my point of view, fear is not a behavior, such as freezing, but rather a process that biases the system to be more likely to emit such a behavior. Freezing is one of the many possible behaviors that express fear. Which one is chosen depends on learning, species, and all kinds of other bias. Part of the task of a model of emotion and

fear is to offer an account of those biases which may be linked to drive states. To link this to our own experience, we may eat because we are hungry or simply because food is placed in front of us. Hunger is what puts us "on alert" to get food, not the behavior of eating the food, though food itself has an incentive value which can increase the likelihood of eating. It is not the reaction to a stimulus but the bias on the way in which we will react to stimuli, as well as the bias on the stimuli we will seek out, paying attention to cues on how to locate food that might otherwise have been ignored.

Before going further, recall our observation that prey capture in the frog includes special-purpose motor pattern generators, those for snapping and ingestion, while predator avoidance uses only general-purpose motor pattern generators for turning and locomotion. Our bodies have complex systems for chewing, digestion, and excretion specialized for feeding, whereas grasping and manipulation are par excellence general-purpose, playing vital roles in feeding, fighting, and expressions of tenderness, to name just a few. We must thus note an increasing dissociation between motivation and the choice of a specific motor system. This is quite orthogonal to my view of emotion as an evolutionary emergence but serves simply to stress that there is no easy correlation between a motivation system and the class of effectors used for the associated motivated behaviors. In any case, a crucial aspect of primate evolution that may be as intimately linked to distinguishing motivation from emotion is the ability to plan behaviors on the basis of future possibilities rather than only in terms of present contingencies. The frog lives in the present, with very little predictive ability and, therefore, only a short-term action–perception cycle. The long-term (from knowing when to refuel to the day–night cycle to the mating season) is handled for the most part by bodily systems and specialized neural systems closely coupled to them. As we compare frog to rat to cat to monkey, the ability to link current decisions to past experiences and future possibilities becomes more explicit and more diverse as the role of the cortex expands.

The Driven Rat

Arbib and Lieblich (1977; see also Lieblich & Arbib, 1982) posited a set $\{d_1, d_2, \ldots d_k\}$ of discrete drives which control the animal's behavior. Typical drives include *appetitive drives*, like thirst, hunger, and sex, and *aversive drives*, like fear. At time t, each drive d has a value $d(t)$, $0 \leq d(t) \leq d_{max}$.[4] They say a drive increases if it changes toward d_{max} and is reduced if it changes toward 0. Their approach seems consistent with the scheme describing the temporal organization of motivated behavior elaborated by Swanson and Mogenson (1981) and recently reviewed by Watts (2003) but with an

explicitly formal representation of the underlying dynamics to explain how motivation affects the way in which an animal will move around its environment. For Watts, interactions of sensory information, arousal state, and interoceptive information determine the value of various drives; the integration of competing drives, presumably via complex sets of projections from the hypothalamus, then determines which series of actions will generate the most appropriate *procurement* (or appetitive) *phase*, where the goal object which will reduce the drive intensity is actively sought. The motor events expressed during the procurement phase involve foraging behavior, are individualized for the particular situation, and can be quite complex. When the goal object has been located, the subsequent *consummatory phase* involves more stereotypic rhythmic movements—licking, chewing, copulating, etc.—that allow the animal to interact directly with the goal object. Watts also notes that interactions between different drive networks, particularly in the hypothalamus, are of paramount importance. For example, the effects of starvation are not limited to increasing the drive to eat but also include reduced reproductive capacity. Similarly, dehydration leads to severe anorexia as well as increased drive to drink (Watts, 2001). This cross-behavioral coordination is part of the mechanism that selects the drive with the highest priority and most likely involves hormonal modulation acting together with the divergent neuroanatomical outputs from individual drive networks. As Watts notes, the notions of drive and that particular behaviors are selected to reduce the level of specific drive states have been very influential in neuroscience but remain somewhat controversial.

Arbib and Lieblich (1977) represented the animal's knowledge of its world in a structure they called the "world graph" (WG), a set of nodes connected by a set of edges, where the nodes represent places or situations recognized by the animal, and the links represent ways of moving from one situation to another. A crucial notion is that a place encountered in different circumstances may be represented by multiple nodes but that these nodes may be merged when the similarity between these circumstances is recognized. They model the process whereby the animal decides where to move next, on the basis of its current drive state (hunger, thirst, fear, etc.) and how the WG may itself be updated in the process. The model includes the effects of incentive (e.g., sight or smell of food) as well as drives (e.g., hunger) to show how a route, possibly of many steps, that leads to the desired goal may be chosen and how short cuts may be chosen. Perhaps the most important feature of their model is their description of what drive-related information is appended to the nodes of the WG and how the WG changes over time. They postulate that each node x of WG(t) is labeled with the vector $[R(d_1,x,t) \ldots R(d_k,x,t)]$ of the animal's current expectations at time t about the drive-related properties of the place or situation $P(x)$ represented

by x. The changes in the WG are of two kinds: changes in the R values labeling the nodes (this finds echoes in the theory of reinforcement learning: Sutton & Barto, 1998) and actual structural changes in the graph.

More recently, we have integrated the WG model with a model of how a rat can still exhibit spatially guided behavior when its hippocampus is lesioned (Guazzelli, Corbacho, Bota, & Arbib, 1998; Guazzelli, Bota, & Arbib, 2001). Figure 12.2 can be seen as the interaction of the following subsystems:

1. The path Sensory Inputs → Parietal Affordances → Premotor Action Selection → Motor Outputs is modulated by drive state (hypothalamus, nucleus accumbens). These interactions are an example of the taxon affordance model (TAM) (Guazzelli, Corbacho, Bota, & Arbib, 1998).
2. Motor outputs affect goal objects in ways that have consequences (gaining food, getting injured, etc.) for the organism. These can affect the internal state of brain and body, updating the drive state. This affects the modulation of (1) in 2 ways:

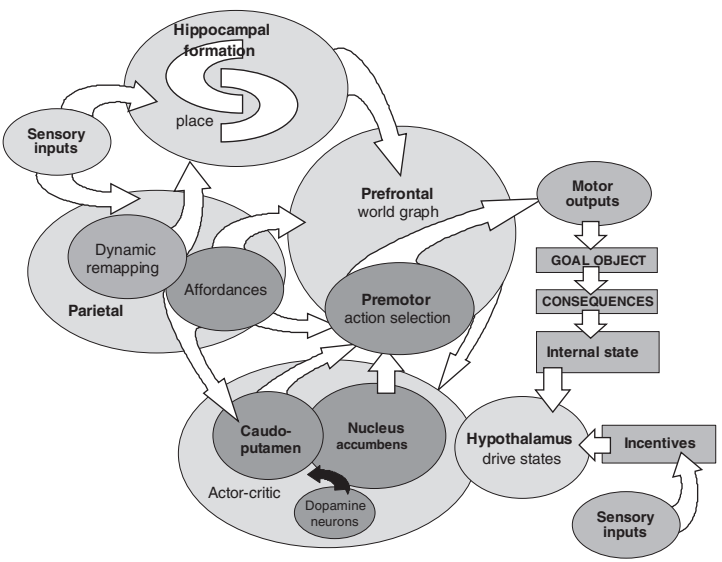


Figure 12.2. The taxon affordance model-world graph (TAM-WG) has as its basis a system, TAM, for exploiting affordances. The path Sensory Inputs → Parietal Affordances → Premotor Action Selection → Motor Outputs is modulated by drive state. This is elaborated by the WG model, which can use a cognitive map, mediated by interactions between hippocampus and prefrontal cortex, to plan paths to targets which are not currently perceptible.

directly and by providing positive or negative reinforcement for learning processes that will affect perceptual, motor, and other schemas. Figure 12.1 shows the latter in terms of the action of dopamine neurons providing the reinforcement for an actor–critic architecture for reinforcement learning in the basal ganglia (Fellous & Suri, 2003; Prescott, Gurney, & Redgrave, 2003).

3. The hippocampus provides a representation of context, most notably by activation of place cells encoding the animal's place in space. Dynamic remapping is the process whereby this representation may be updated on the basis of an efferent copy of the animal's actions even when sensory data on the new context or location may be missing.
4. The hippocampal representation is now seen in greater generality as encoding any situation that is linked to a node in the animal's WG. Thus, the hippocampus is seen as providing the "you are here" function; it must be integrated with the WG to provide a full cognitive map linking current position with current goals to determine a path through the world which will achieve one or more of them. Thus, premotor action selection becomes embedded within the prefrontal planning behavior associated with the WG, planning which depends crucially on representations of goal states and internal states (inputs not shown) as well as a combination of the current situation and the available affordances (Guazzelli, Bota, & Arbib, 2001).

An Evolutionary Approach to Heated Appraisals

In item (3) of the analysis of my emotional narrative (see above, "The Narrative Chronologically Rearranged and Annotated"), I stated the following:

> an emotional state may be "terminated" when a plan is completed which addresses the source of that state; or it may simply dissipate when alternative plans are made. However, if some major goal has been rendered unattainable by some event, the negative emotion associated with the event may not be dissipated by embarking on other plans, since they do not approach the goal.

This analysis, and others elsewhere in the above section, seems at first to be very much in the spirit of appraisal theories of emotion (e.g., Ortony, Clore, & Collins, 1988). However, while Ortony, Clore and Collins admit that visceral sensations and facial expressions set emotions apart from other psychological states, they exclude these from their study (Arbib, 1992). They

insist that the origins of emotional states are based on the cognitive construal of events, with cognition being presupposed by the physiological, behavioral, and expressive aspects of emotion. However, is the addition of a cognitive evaluation, unlinked to a physiological structure, enough to convert information processing into emotion? While I find great value in the attempt of Ortony, Clore, and Collins to see how different cognitive structures may relate to different emotional states, it gives no sense of the "heat" of emotion that I tried to convey in my short narrative. I argue here that the "heat" is added to the appraisal because the cerebral cortex is linked to basic motivational systems, while human emotions are so much more varied and subtle than mere hunger or thirst because basic motivational systems are integrated with cortical systems that can provide varied appraisals.

My approach will be to integrate insights from chapters by Kelley, Rolls, and Fellous & LeDoux with the evolutionary perspective provided above while confronting the personal data set of my emotional narrative.

Behavioral Control Columns

To start, we must note that a full analysis of motivated behavior must include not only the somatomotor behavior (e.g., feeding and fleeing; other forms relevant to the study of motivation-related behavior include orofacial responses and defensive and mating activities) but also autonomic output (e.g., heart rate and blood pressure) and visceroendocrine output (cortisol, adrenaline, release of sex hormones). In general, behavior will combine effects of all three kinds. Kelley (Chapter 3) places special emphasis on Swanson's (2000) notion of the behavioral control column (Fig. 12.3). This is a column of nuclei arrayed along the brain stem. Swanson proposes that very specific and highly interconnected sets of nuclei in the hypothalamus are devoted to the elaboration and control of specific behaviors necessary for survival: spontaneous locomotion, exploration, ingestive, defensive, and reproductive behaviors. Animals with chronic transections above the hypothalamus can more or less eat, drink, reproduce, and show defensive behaviors, whereas if the brain is transected below the hypothalamus, the animal displays only fragments of these behaviors, enabled by motor pattern generators in the brain stem. As Kelley notes, many instances of motivated behavior—eating, drinking, grooming, attacking, sleeping, maternal behavior, hoarding, copulating—have been evoked by direct electrical or chemical stimulation of the hypothalamus.

The behavioral control column contains a rostral and a more caudal segment. The former contains nuclei involved in ingestive and social (reproductive and defensive) behaviors such as sexually dimorphic behaviors, defensive

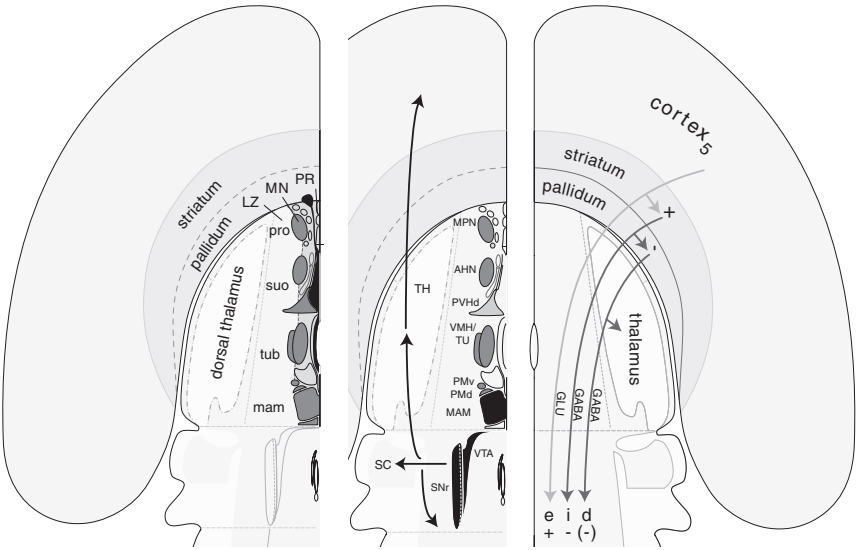


Figure 12.3. Major features of cerebral hemisphere regulation of motivated behavior, according to Swanson, as seen on a flatmap of the rat central nervous system. (*Left*) The neuroendocrine motor zone, shown in black, and three subgroups of hypothalamic nuclei: the periventricular region (PR) most centrally, the medial nuclei (MN,) and the lateral zone (LZ). The PR contains a visceromotor pattern generator network, and the medial nuclei (MN) form the rostral end of the behavior control column. In addition to this longitudinal division, the hypothalamus can be divided into four transverse regions based on the characteristic medial nucleus residing within it: preoptic (pro), supraoptic or anterior (suo), tuberal (tub), and mammillary (mam). (*Center*) An overview of the behavior control column. Almost all nuclei in this column generate a dual, typically branched projection, descending to the motor system and ascending to thalamocortical loops: AHN, anterior hypothalamic nucleus; MAM, mammillary body; MPN, medial preoptic nucleus (lateral part in particular); PMdv, premammillary nuclei, dorsal ventral; PVHd, descending division of paraventricular hypothalamic nucleus; SC, superior colliculus, deeper layers; SNr, reticular substantia nigra; TH, dorsal thalamus; TU, tuberal nucleus; VMH, ventromedial hypothalamic nucleus; VTA, ventral tegmental area. (*Right*) Triple cascading projection from the cerebral hemispheres to the brain-stem motor system. This minimal or prototypical circuit element consists of a glutamatergic (GLU) projection from layer 5 pyramidal neurons of the isocortex (or equivalent pyramidal neurons in allocortex), with a glutamatergic collateral to the striatum. This dual projection appears to be excitatory (e, +). The striatum then generates a γ-aminobutyric acid-ergic (GABAergic) projection to the motor system with a GABAergic collateral to the pallidum. This dual striatal projection appears to be inhibitory (i, –). Finally, the pallidum generates a GABAergic projection to the brain-stem motor system, with a GABAergic collateral to the dorsal thalamus. This dual pallidal projection can be viewed as disinhibitory (d, –) because it is inhibited by the striatal input. (Adapted from Swanson, 2000, Figs. 8, 10, 14, respectively.)

responses, or controls for food and water intake. The more caudal segment of the column is involved in general foraging/exploratory behaviors.

Kelley notes that the lateral hypothalamus is not specifically included in Swanson's behavioral control column scheme but probably plays a critical role in arousal, control of behavioral state, and reward-seeking behavior. It includes what Olds (1977) referred to as the "pleasure center" because rats will press a lever thousands of times per hour to deliver electrical stimulation to this region.

We may distinguish drive signals—energy deficits, osmotic imbalance, visceral cues (including pain, temperature, and heart rate), metabolic and humoral information, etc.—from external cues about objects and other animals in the world. Following Risold, Thompson, & Swanson (1997), Kelley reviews the many paths whereby both kinds of information reach the hypothalamus, with much specificity as to which kinds of information affect which nuclei.

Amygdala, Orbitofrontal Cortex, and their Friends

Kelley notes that the amygdala's role in reward valuation and learning, particularly in its lateral and basolateral aspects (which are intimately connected with the frontotemporal association cortex) can influence and perhaps bias lateral hypothalamic output, citing literature on ingestive behavior which complements the emphasis of Fellous and LeDoux (Chapter 4) of certain nuclei of the amygdala in fear behavior.

Figure 12.4 summarizes the evolutionary perspective of Fellous and LeDoux on fearful behavior. The role of the hippocampus in conditioning to contextual cues can be usefully compared to its "you are here" function in the Figure 12.2 model of motivated spatial behavior. The crucial element from an evolutionary point of view is the set of reciprocal interactions between the amygdala and cerebral cortex: the amygdala can influence cortical areas by way of feedback either from proprioceptive or visceral signals or hormones, via projections to various arousal networks (these are discussed extensively in Kelley's Chapter 3), and through interaction with the medial prefrontal cortex. This area has widespread influences on cognition and behavior and sends connections to several amygdala regions, allowing cognitive functions organized in prefrontal regions to regulate the amygdala and its fear reactions.

This, for fear behavior, provides but one example of the major principle for organization of the behavioral control columns—namely, that they project massively back to the cerebral cortex/voluntary control system directly or indirectly via the dorsal thalamus (Risold, Thompson, & Swanson, 1997;

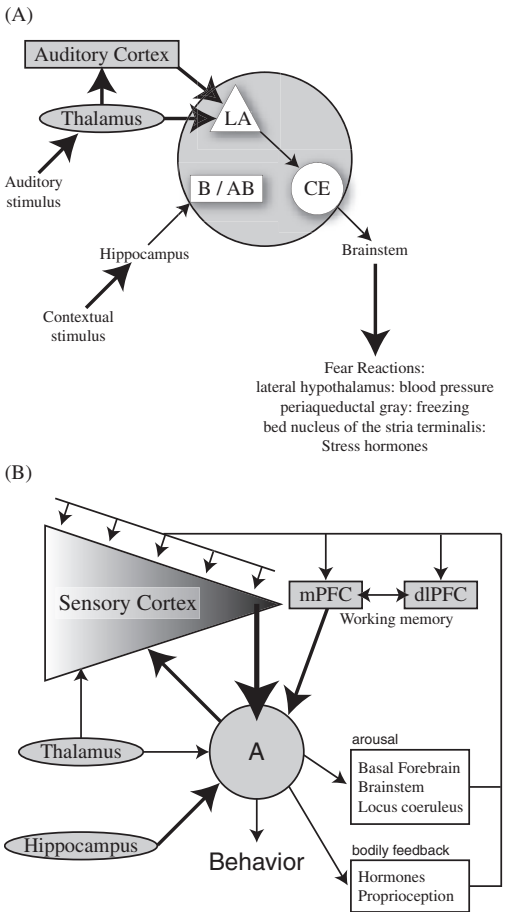


Figure 12.4. (*a*) Differential paths through the amygdala for fear conditioning to an auditory stimulus and to contextual cues. (*b*) Interaction of the amygdala with cortical areas allows cognitive functions organized in prefrontal regions to regulate the amygdala and its fear reactions. LA, lateral nucleus of amygdala; CE, central nucleus of amygdala; B/AB, basal/accessory basal nuclei of amygdala. (Adapted from Fellous & LeDoux, Chapter 4, Fig. 4.2).

Swanson, 2000). Kelley stresses that this feed-forward hypothalamic projection to the cerebral hemispheres provides the anatomical substrate for the

> intimate access of associative and cognitive cortical areas to basic motivational networks [which] enables the generation of emotions, or the manifestation of "motivational potential." Thus, in the primate brain, this substantial reciprocal interaction between . . . behavioral control columns and . . . cortex subserving higher order processes

such as language and cognition has enabled a two-way street for emotion.

Rolls (Chapter 5) emphasizes diverse roles of the amygdala in the monkey. Monkey amygdala receives information about primary reinforcers (e.g., taste and touch) and about visual and auditory stimuli from higher cortical areas (e.g., inferior temporal cortex) that can be associated by learning with primary reinforcers (Fig. 12.5). Monkeys will work in order to obtain electrical stimulation of the amygdala; single neurons in the amygdala are activated by brain-stimulation reward of a number of different sites, and some amygdala neurons respond mainly to rewarding stimuli and others to punishing stimuli. There are neurons in the amygdala (e.g., in the basal accessory nucleus) which respond primarily to faces; they may be related to inferring the emotional content of facial expressions. Moreover, the human amygdala can be activated in neuroimaging studies by observing facial expressions, and lesions of the human amygdala may cause difficulty in the identification of some such expressions (see Rolls, 2000).

Figure 12.5 also suggests the crucial role of the orbitofrontal cortex in linking the frontal cortex to the emotional system. It receives inputs from the inferior temporal visual cortex and superior temporal auditory cortex;

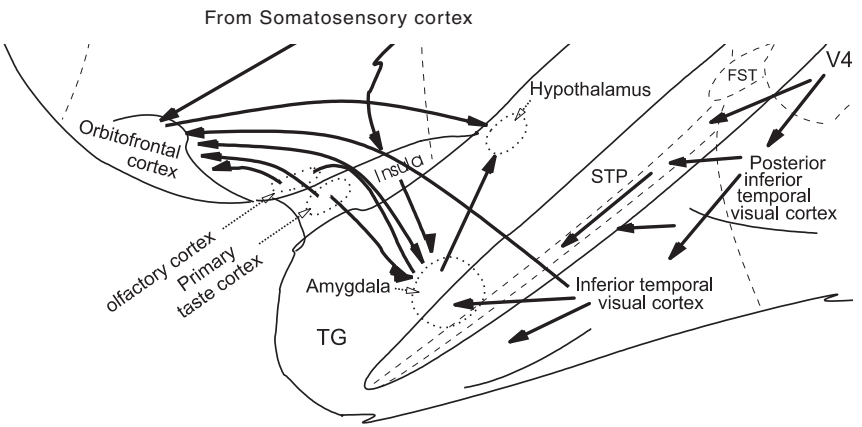


Figure 12.5. Some of the pathways involved in emotion shown on a lateral view of the brain of the macaque monkey, emphasizing connections from the primary taste and olfactory cortices and from the inferior temporal cortex to the orbitofrontal cortex and amygdala. The secondary taste cortex and the secondary olfactory cortex are within the orbitofrontal cortex. Connections from the somatosensory cortex reach the orbitofrontal cortex directly and via the insular cortex, as well as the amygdala via the insular cortex. TG, architectonic area in the temporal pole; V4 is visual area 4. (Adapted from Rolls, Chapter 5, Fig. 5.4.)

from the primary taste cortex and the primary olfactory (pyriform) cortex; from the amygdala; and from midbrain dopamine neurons. As Rolls documents, damage to the caudal orbitofrontal cortex produces emotional changes, which include the tendency to respond when responses are inappropriate (i.e., the tendency of monkeys not to withhold responses to nonrewarded stimuli). Rolls sees orbitofrontal neurons as part of a mechanism which evaluates whether a reward is expected and generates a mismatch (evident as a firing of the nonreward neurons) if the reward is not obtained when it is expected.

As Fellous and LeDoux note, decision-making ability in emotional situations is also impaired in humans with damage to the medial prefrontal cortex and abnormalities in the prefrontal cortex may predispose people to develop fear and anxiety disorders. They suggest that the medial prefrontal cortex allows cognitive information processing in the prefrontal cortex to regulate emotional processing by the amygdala, while emotional processing by the amygdala may influence the decision-making and other cognitive functions of the prefrontal cortex. They then suggest that the prefrontal–amygdala interactions may be involved in the conscious feelings of fear. However, this neat division between the cognitive cortex and emotional amygdala strikes me as too glib—both because not all parts of the cortex give rise to conscious feelings and because human emotions seem to be inextricably bound up with "cortical subtleties."

Neuromodulation

We have now seen something of the crucial roles of the hypothalamus, amygdala, and orbitofrontal cortex in the motivational system. In a similar vein, Fellous (1999) reviewed the involvement of these three areas in emotion and argued that the neural basis for emotion involves both computations in these structures and their neuromodulation. It is thus a useful feature of the present volume that Kelley's analysis of motivation and emotion emphasizes three widely distributed chemical signaling systems and their related functions across different phyla. Following are some key points from her richly detailed survey.

We first discuss dopamine, reward, and plasticity. In mammals, dopamine is proposed to play a major role in motor activation, appetitive motivation, reward processing, and cellular plasticity and may well play a major role in emotion. In the mammalian brain, dopamine is contained in specific pathways, which have their origins in the substantia nigra pars compacta and the ventral tegmental area of the midbrain and ascend to innervate widespread regions of striatal, limbic, and cortical regions such as the striatum, prefrontal cortex, amygdala, and other forebrain regions. Studies in the

awake, behaving monkey show dopamine neurons which fire to predicted rewards and track expected and unexpected environmental events, thereby encoding prediction errors (Schultz, 2000; Fellous & Suri, 2003). Moreover, prefrontal networks are equipped with the ability to hold neural representations in memory and to use them to guide adaptive behavior; dopamine receptors are essential for this ability. Thus, dopamine plays essential roles all the way from basic motivational systems to the working memory systems seen to be essential to the linkage of emotion and consciousness (see below for a critique).

Next, consider serotonin, aggression and depression. Kelley (Chapter 3) shows that serotonin has been widely implicated in many behavioral functions, including behavioral state regulation and arousal, motor pattern generation, sleep, learning and plasticity, food intake, mood, and social behavior. The cell bodies of serotonergic systems are found in midbrain and pontine regions in the mammalian brain and have extensive descending and ascending projections. Serotonin plays a critical role in the modulation of aggression and agonistic social interactions in many animals—in crustaceans, serotonin plays a specific role in social status and aggression; in primates, with the system's expansive development and innervation of the cerebral cortex, serotonin has come to play a much broader role in cognitive and emotional regulation, particularly control of negative mood or affect.

Finally, we look at opioid peptides and their role in pain and pleasure. Kelley shows that opioids, which include the endorphins, enkephalins, and dynorphins, are found particularly within regions involved in emotional regulation, responses to pain and stress, endocrine regulation, and food intake. Increased opioid function is clearly associated with positive affective states such as relief of pain and feelings of euphoria, well-being, and relaxation. Activation of opioid receptors promotes maternal behavior in mothers and attachment behavior and social play in juveniles. Separation distress, exhibited by archetypal behaviors and calls in most mammals and birds, is reduced by opiate agonists and increased by opiate antagonists in many species (Panksepp, 1998). Opiates can also effect the reduction or elimination of the physical sensation induced by a painful stimulus as well as the negative emotional state it induces.

What is striking here is the way in which these three great neuromodulatory systems seem to be distinct from each other in their overall functionalities, while exhibiting immense diversity of behavioral consequences within each family. The different effects depend on both molecular details (the receptors which determine how a cell will respond to the presence of the neuromodulator) and global arrangements (the circuitry within the modulated brain region and the connections of that region within the brain). Kelley notes that much of the investigation of central opioids has been fueled

by an interest in understanding the nature of addiction—perhaps another aspect of the "beware the passionate robot" theme. As she documents in her section Addictive Drugs and Artificial Stimulation of Emotions, these systems may have great adaptive value in certain contexts yet may be maladaptive in others. This raises the intriguing question of whether the effects of neuromodulation could be more adaptive for the animal if the rather large-scale "broadcast" of a few neuromodulators were replaced by a targeted and more information-rich distribution of a far more diverse set of neuromodulators. This makes little sense in terms of the conservatism of biological evolution but may have implications both for the design of drugs which modify neuromodulators to target only cells with specific molecular markers and in future research on robot emotions which seeks to determine useful computational and technological analogs for neuromodulation.

Emotion and Consciousness with a Nod to Empathy

With this, let us turn to notions of the linkage between emotion and consciousness. Fellous and LeDoux (Chapter 4) endorse theories of consciousness built around the concept of working memory. They say

> the feeling of being afraid would be a state of consciousness in which working memory integrates the following disparate kinds of information: (1) an immediately present stimulus (say, a snake on the path in front of you); (2) long-term memories about that stimulus (facts you know about snakes and experiences you have had with them); and (3) emotional arousal by the amygdala.

However, we saw that activity in the parietal cortex may have no access to consciousness (patient D. F.), even though (Fig. 12.1) it is coupled to prefrontal working memory. Thus, working memory is not the key to consciousness; but if we agree to simply accept that some cortical circuits support conscious states while others do not, then we can still agree with Fellous and LeDoux as to the importance of emotional feelings of connections from the amygdala to the medial (anterior cingulate) and ventral (orbital) prefrontal cortex. As they (and Rolls) note, humans with orbitofrontal cortex damage ignore social and emotional cues and make poor decisions, and some may even exhibit sociopathic behavior. They stress that, in addition to being connected with the amygdala, the anterior cingulate and orbital areas are intimately connected with one another as well as with the lateral prefrontal cortex, and each of the prefrontal areas receives information from sensory processing regions and from areas involved in various aspects of implicit and explicit memory processing.

Where Fellous and LeDoux emphasize working memory in their cortical model and include the orbital cortex as part of this cortical refinement, Rolls (1999) includes the orbitofrontal cortex in both routes of his two-route model. Before we look at this model, though, one caveat about Rolls' general view that emotions are states elicited by rewards and punishers. Rolls states that his approach helps with understanding the functions of emotion, classifying different emotions, and understanding what information processing systems in the brain are involved in emotion and how they are involved. Indeed it does but, where Rolls emphasizes the polarity between reward and punishment, I would rather ground a theory of emotions in the basic drives of Arbib and Lieblich (1977) as seen in the basic hypothalamic and midbrain nuclei of Swanson's (2000) behavioral control column and the basic neuromodulatory systems of Fellous (1999) and Kelley (Chapter 3). Where Rolls argues that brains are designed around reward-and-punishment evaluation systems because this is the way that genes can build a complex system that will produce appropriate but flexible behavior to increase their fitness, I would stress (with Swanson) the diversity of specific motor and perceptual systems that the genes provide, while agreeing that various learning systems, based on various patterns of error feedback as well as positive and negative reinforcement, can provide the organism with adaptability in building upon this basic repertoire that would otherwise be unattainable. (Consider Figure 12.2 to see how much machinery evolution has crafted beyond basic drives and incentives, let alone simple reward and punishment.)

Rolls argues that there are two types of route to action performed in relation to reward or punishment in humans. The first route (see the middle row of Fig. 5.2) includes the amygdala and, particularly well-developed in primates, the orbitofrontal cortex. These systems control behavior in relation to previous associations of stimuli with reinforcement. He notes various properties of this system, such as hysteresis, which prevents an animal that is equally hungry and thirsty from continually switching back and forth between eating and drinking. In other words, despite the emphasis that Rolls lays on reward and punishment, the analysis is in many ways linked to the differential effects of different drive systems. The second route (see the top row of Fig. 5.2) involves a computation with many "if . . . then" statements, to implement a plan to obtain a reward. Rolls argues that syntax is required here because the many symbols that are part of the plan must be correctly linked, as in: "if A does this, then B is likely to do this, and this will cause C to do this." I think Rolls may be mistaken to the extent that he conflates syntax in simple planning with the explicit symbolic expression of syntax involved in language. Nonetheless (as in the Arbib-Hesse theory), I do agree that the full range of emotion in humans involves the interaction of the language system with a range of other systems. Rolls holds that the second route is related to consciousness, which

he sees as the state that arises by virtue of having the ability to think about one's own thoughts and analyze long, multistep syntactic plans. He aligns himself implicitly with Fellous and LeDoux when he says that another building block for such planning operations may be the type of short-term memory (i.e., working memory) provided by the prefrontal cortex. The type of working memory system implemented in the dorsolateral and inferior convexity of the prefrontal cortex of nonhuman primates and humans (Goldman-Rakic, 1996) could provide mechanisms essential to forming a multiple-step plan. However, as I have commented earlier, the prefrontal cortex involves a variety of working memories, some of which have no direct relation either to consciousness or to emotion.

The model of thought that emerges here sees each mental state as combining emotional and cognitive components. While in some cases one component or the other may be almost negligible, it seems more appropriate, on this account, to see the emotional states as being continually present and varying, rather than as intermittent. The model also sees physiological states and emotion as inextricably intertwined—a cognitive state may induce an emotional reaction, but a prior emotional state may yield a subconscious physiological residue that influences the ensuing unfolding of cognitive and emotional states.

Adolphs (Chapter 2) stresses the important role of social interaction in the forming of emotions. Clearly, human emotions are greatly shaped by our reactions to the behavior of other people. Returning once more to the OED,

> **Empathy**: The power of projecting one's personality into (and so fully comprehending) the object of contemplation. (This term was apparently introduced to English in 1909 by E. B. Titchener *Lect. Exper. Psychol. Thought-Processes*: "Not only do I see gravity and modesty and pride . . . but I feel or act them in the mind's muscles. This is, I suppose, a simple case of empathy, if we may coin that term as a rendering of *Einfühlung*."

We find a definition which carries within itself the simulation theory discussed by Jeannerod in Chapter 6, but with "the mind's muscles" transformed into the mirror system, which is a network of neurons active both when the "brain owner" acts in a certain way and when he or she observes another acting in a similar fashion. Earlier, we outlined a number of intermediate stages in the evolution of mechanisms that support language. I suggest that, similarly, a number of stages would have to intervene in the evolution of the brain mechanisms that support emotion and empathy. However, this topic and the related issue of the extent to which there has been a synergy between the evolution of language and the evolution of empathy are beyond the scope of the present chapter.

EMOTION WITHOUT BIOLOGY

Norbert Wiener, whose book *Cybernetics: Or Control and Communication in the Animal and the Machine* introduced the term *cybernetics* in the sense that is at the root of its modern usage, also wrote *The Human Use of Human Beings* (Wiener 1950, 1961). I remember that a professor friend of my parents, John Blatt, was shocked by the latter title; in his moral view, it was improper for one human being to "use" another. I suspect that Wiener would have agreed, while noting that modern society does indeed see humans using others in many disturbing ways and, thus, much needs to be done to improve the morality of human interactions. By contrast, robots and other machines are programmed for specific uses. We must thus distinguish two senses of autonomy relevant to our discussion but which often infect each other.

- When we talk of an "autonomous human," the sense of autonomy is that of a person becoming a member of a society and, while working within certain constraints of that society and respecting many or all of its moral conventions, finding his or her own path in which work, play, personal relations, family, and so on can be chosen and balanced in a way that grows out of the subject's experience rather than being imposed by others.
- When we talk of an "autonomous machine," the sense is of a machine that has considerable control over its sensory inputs and the ability to choose actions based on an adaptive set of criteria rather than too rigidly predesigned a program.[5]

On this account, a human slave is autonomous in the machine sense but not in the human sense. Some researchers on autonomous machines seem to speak as if such machines should be autonomous in the human sense. However, when we use computers and robots, it is with pragmatic human-set goals rather than "finding one's own path in life" that we are truly concerned. In computer science, much effort has been expended on showing that programs meet their specifications without harmful side effects. Surely, with robots, too, our concerns will be the same. What makes this more challenging is that, in the future, the controllers for robots will reflect many generations of machine learning and of tweaking by genetic algorithms and be far removed from clean symbolic specifications. Yet, with all that, we as humans will demand warranties that the robots we buy will perform as stated by the supplier. It might be objected that "as adaptive robots learn new behaviors and new contexts for them, it will be impossible for a supplier to issue such a guarantee." However, this will not be acceptable in the marketplace. If, for example, one were to purchase a car that had adaptive circuitry

to detect possible collisions and to find optimal routes to a chosen destination, one should demand that the car does not take a perverse delight (assuming it has emotions!) in avoiding collisions by stopping so suddenly as to risk injury to its human occupants or in switching the destination from that chosen by the driver to one the car prefers. If we purchase a computer tutor, then the ability to provide the appearance of emotions useful to the student may well be part of the specifications, as will the ability to avoid (with the caveats mentioned at the beginning of this chapter) the temper tantrums to which the more excitable human teacher may occasionally succumb. In summary, machine learning must meet constraints of designed use, while at the same time exploring novel solutions to the problem. This does not rule out the possibility of unintended consequences, but when a machine "goes wrong," there should be maintenance routines to fix it that would be very different from either the medical treatment or penal servitude applied to humans.

Once again, let us look at the OED for definitions and recall the warning "beware the passionate robot." We can then consider some key questions in our attempt to understand the nature of robot emotions, if indeed they do or will exist.

> **Action: I**. Generally. **1.** The process or condition of acting or doing (in the widest sense), the exertion of energy or influence; working, agency, operation. **a.** Of persons. (Distinguished from passion, from thought or contemplation, from speaking or writing.)
>
> **Active:** gen. Characterized by action. Hence **A.** adj. **1. a.** Opposed to contemplative or speculative: Given to outward action rather than inward contemplation or speculation. **2.** Opposed to passive: Originating or communicating action, exerting action upon others; acting of its own accord, spontaneous.
>
> **Passion: III. 6. a**. Any kind of feeling by which the mind is powerfully affected or moved; a vehement, commanding, or overpowering emotion; . . . as ambition, avarice, desire, hope, fear, love, hatred, joy, grief, anger, revenge. **7. a**. *spec*. An outburst of anger or bad temper. **8. a.** Amorous feeling; strong sexual affection; love.
>
> **Passive: A.** adj. **2. a.** Suffering action from without; that is the object, as distinguished from the subject, of action; acted upon, affected, or swayed by external force; produced or brought about by external agency.

I included these definitions because I find it instructive to consider that in everyday parlance we lose active control of our selves when we are in the

grip of passion, that is, of strong emotion. In other words, while our emotions and our reason may usually be brought into accord, there are other times when "ambition, avarice, desire, hope, fear, love, hatred, joy, grief, anger, [or] revenge" may consume us, effectively banishing all alternatives from our thoughts. In noting this, I frame the question "If emotions conveyed an advantage in biological evolution, why can they be so harmful as well?" We have already noted that Kelley (Chapter 3) examined the role of opioids in addiction and discussed how these may have great adaptive value in certain contexts yet may be maladaptive in others.

Such issues raise the prior question "Did emotions convey a selective advantage?" and the subsequent questions "Are emotions a side effect of a certain kind of cognitive complexity?" (which might imply that robots of a certain subtlety will automatically have emotion as a side effect) and "Were emotions the result of separate evolutionary changes, and if so, do their advantages outweigh their disadvantages in a way that might make it appropriate to incorporate them in robots (whether through explicit design or selective pressure)?"

At the beginning of this chapter, we considered a scenario for a computer designed to effectively teach some body of material to a human student and saw that we might include "providing what a human will recognize as a helpful emotional tone" to the list of criteria for successful program design. However, there is no evolutionary sequence here as charted by the neurobiologists—none of the serotonin or dopamine of Kelley, none of the punishment and reward of Rolls, none of the "fear circuits" of Fellous & LeDoux. This is not to deny that there can be an interesting study of "computer evolution" from the switches of the ENIAC, to the punchcards of the PDP11 to the keyboard to the use of the mouse and, perhaps, to the computer that perceives and expresses emotions. My point here is simply that the computer's evolution to emotion will not have the biological grounding of human emotion. The computer may use a model of the student's emotions yet may not be itself subject to, for example, reward or punishment. Intriguingly, this is simulation with a vengeance—yet not simulation in the mirror sense employed by Jeannerod in Chapter 6—the simulation is purely of "the other," not a reflection of the other back onto the self. In the same way, one may have a model of a car to drive it without having an internal combustion engine or wheels. Then we must ask if this is an argument against the simulation theory of human emotion. This also points to a multilevel view. At one level, the computer "just follows the program" and humans "just follow the neural dynamics." It is only a multilevel view that lets us single out certain variables as drives. What does that imply for robot emotions?

Suppose, then, that we have a robot that simulates the appearance of emotional behavior but has none of the "heated feeling" that governed so

much of my behavior in the opening anecdote. If neither biology nor feeling remains, can we say that such a robot has emotions? In a related vein, reinforcement learning (Sutton & Barto, 1998) has established itself as being of great value in the study of machine learning and artificial neural networks. However, when are we justified in seeing positive reinforcement at the psychological/emotional level rather than being simply a mathematical term in a synaptic adjustment rule?

Appraisal theory (as in Ortony, Clore, & Collins, 1988) develops a catalog of human emotions and seeks to provide a computational account of the appraisals which lead to invocation of one emotion over another. I suggest that robot emotions may partake of some aspects of the appraisal approach to emotions but without the "heat" provided by their biological underpinnings in humans. If it is part of the job of a robot to simulate the appearance of human emotional behavior to more effectively serve human needs, then it may (as suggested earlier) incorporate a model of human emotions within its circuitry, and this may well be appraisal-based. It might then be a matter of terminology as to whether or not one would wish to speak of such a robot having emotions. However, I think a truly fruitful theory of robot emotions must address the fact that many robots will not have a human–computer interface in which the expression of human-like emotional gestures plays a role. Can one, then, ascribe emotions to a robot (or for that matter an animal or collective of animals) for which empathy is impossible?

Perhaps a more abstract view of emotion is required if we are to speak of robot emotions.

To this end, I must first deliver on my promise to provide an abstraction of the notion of ecological niche suitable for robots. In the case of animals, we usually refer to the part of the world where the animal is to make a living. However, locale is not enough. Foodstuffs that are indigestible to one species may be the staff of life to another, the size of the creature can determine where it can find a suitable resting place, and different creatures in a given environment may have different predators. A new species in an environment may create new ecological niches there for others. In biology, the four Fs (feeding, fighting, fleeing, and reproduction) are paramount, and it is success in these that defines the animal's relation to its environment. However, none of this applies to most robots. One can certainly imagine scenarios in which the "struggle for fuel" plays a dominant role in a robot economy, but robot design will normally be based on the availability of a reliable supply of electricity. Although the study of self-reproducing machines is well established (von Neumann, 1966; Arbib, 1966), the reproduction of robots will normally be left to factories rather than added to the robot's own workload. Thus, the ecological niche of a robot will not be defined in terms of general life functions as much as in a set of tasks that it is

designed to perform, though certainly environmental considerations will have their effect. The design of a Mars rover and a computer tutor must take into account the difference between an environment of temperature extremes and dust storms and an air-conditioned classroom. Nonetheless, my point holds that whereas an animal's set of tasks can be related more or (as in the case of humans) somewhat less directly to the four Fs, in the case of robots, there will be an immense variety of task sets, and these will constrain the sensors, controllers, and effectors in a way which must be referred to the task sets without any foundation in the biological imperatives that have shaped the evolution of motivational and emotional systems for biological creatures.

Another classic brain model is relevant here. Kilmer, McCulloch, and Blum (1969) implemented McCulloch's idea of extending observations on the involvement of the reticular formation in switching the animal from sleep to waking (Magoun, 1963) to the hypothesis that the reticular formation was responsible for switching the overall mode of feeding or fleeing or whatever, and then the rest of the brain, when set into this mode, could do the more detailed computations. The data of Scheibel and Scheibel (1958) on the dendritic trees of neurons of the reticular formation suggested the idea of modeling the reticular formation as a stack of modules, each with a slightly different selection of input but trying to decide to which mode to commit the organism. They would communicate back and forth, competing and cooperating until finally they reached a consensus on the basis of their diverse input; that consensus would switch the mode of the organism. In this framework, Kilmer, McCulloch, and Blum (1969) simulated a model, called S-RETIC, of a modular system designed to compute modes in this cooperative manner. Computer simulation showed that S-RETIC would converge for every input in fewer than 25 cycles and that, once it had converged, it would stay converged for the given input. When the inputs strongly indicate one mode, the response is fast; but when the indication is weak, initial conditions and circuit characteristics may strongly bias the final decision.

Within any mode of behavior many different acts are possible: if the cat should flee, will it take the mouse or leave it, climb a tree or skirt it, jump a creek or swim it? The notion is that a hierarchical structure that computes modes and then acts within them, might in some sense be "better" (irrespective of the particular structural basis ascribed to these functions) than one that tries to determine successive acts directly.

For robot emotions, then, the issue is to what extent emotions may contribute to or detract from the success of a "species" of robots in filling their ecological niche. I thus suggest that an effort to describe robot emotions requires us to analyze the tasks performed by the robot and the strategies available to perform them.

Consider a robot that has a core set of basic functions, each with appropriate perceptual schemas, F_1, F_2 . . . F_n (generalizing the four Fs!), each of which has access to more or less separate motor controllers, M_1, M_2 . . . M_n, though these may share some motor subschemas, as in the use of orienting and locomotion for prey capture and predator avoidance in the frog. Each F_j evaluates the current state to come up with an urgency level for activating its motor schema M_j, as well as determining appropriate motor parameters (it is not enough just to snap, but the frog must snap at the fly). Under basic operating conditions, a winner-take-all or similar process can adjudicate between these processes (does the frog snap at the fly or escape the predator?). We might want to say, then, that a motivational system is a state-evaluation processes that can adjust the relative weighting of the different functions, raising the urgency level for one system while lowering the motivation system for others, so that a stimulus that might have activated M_k in one context will now instead activate M_i.

What if, as is likely, the set of tasks is not a small set of survival tasks but indeed a very large set? It may help to recall (Watts, 2003) that the procurement phase in animal behavior is individualized for the particular situation and can be quite complex, whereas the subsequent consummatory phase involves more stereotypic movements. What I take from this is not the idea of organizing a variety of behaviors with respect to the consummatory phase they serve but, rather, the idea that action selection may well involve grouping a large set of actions into a small number of groups, each containing many actions or tasks. In other words, we consider the modes of the S-RETIC model as abstract groups of tasks rather than as related to biological drives like the four Fs. I consider the case where there are *m* strategies which can be grouped into *n* groups (with *n* much less than *m*) such that it is in general more efficient, when faced with a problem, to first select an appropriate group of strategies and then to select a strategy from within that group. The catch, of course, is in the caveat "in general." There may be cases in which rapid commitment to one group of strategies may preclude finding the most appropriate strategy—possibly at times with disastrous consequences. Effective robot design would thus have to balance this fast commitment process against more subtle evaluative process that can check the suitability of a chosen strategy before committing to it completely. We might then liken motivation to biases which favor one strategy group over another and emotion to the way in which these biases interact with more subtle computations. On this abstract viewpoint, the "passionate robot" is not one which loses its temper in the human-like fashion of the computer tutor imagined earlier but rather one in which biases favoring rapid commitment to one strategy group overwhelm more cautious analysis of the suitability of strategies selected from that group for the task at hand.

As far as the chapters in Part III, Robots, are concerned, Arkin's "Moving Up the Food Chain: Motivation and Emotion in Behavior-based Robots" comes closest to providing insight from animal brains and behavior, whereas Chapter 7 (Ortony et al.) and Chapter 8 (Sloman et al.) provide multilevel views of artificial intelligence that offer fertile ground for a comparative analysis with the approach to conceptual evolution offered in the present chapter. Nonetheless, my approach does not link directly to the crucial role of social interactions in emotion stressed by Adolphs (Chapter 2), though Jeannerod (Chapter 6) does explore the possible role of mirror systems in our ability to understand the emotions of others, and we saw in the mirror system of primates bridging from an individual skill (dexterity) to a social skill (language).

It is thus up to future research to integrate the preliminary theory of robot emotions given here, grounded as it is in the analysis of a robot going about its tasks in some ecological niche, with an approach emphasizing social interactions. To set directions for such integration in the future, I close by noting the relevance of two chapters in this volume. Breazeal and Brooks (Chapter 10) survey progress in making robots more effective in social interactions with humans, arguing that emotion-inspired mechanisms can improve the way autonomous robots operate in a human environment and can improve the ability of these robots to effectively achieve their own goals. More generally, Nair, Tambe, and Marsella (Chapter 11) use a survey of the state of the art in multi-agent teamwork (an agent may be considered a generalization of robots to include "softbots" as well as embodied robots) and in computational models of emotions to consider the role of emotions not only in agent–human teams but also in pure agent teams. The stage seems well set for dramatic progress in integrating brain and society in theories of emotion in animals and humans and for linking solo tasks, robot–human interaction, and teamwork in the further exploration of the topic "Who needs emotion?" Not only will the study of brains continue to inform our analysis of robots, but the precision required to make explicit the computational strategies of robots will enrich the vocabulary for the study of motivation and emotion in humans and other animals.

Notes

1. A note of relevance to the evolutionary perspective of the present chapter: in his Chapter 7, Hebb (1949) states that humans are the most emotional of all animals because degree of emotionality seems to be correlated with the phylogenic development of sophisticated nervous systems.

2. The definition numbers are those given in the online version of the OED (http://dictionary.oed.com/ © Oxford University Press, 2003; accessed December 2003).

3. As already noted, such interpretations are tentative, designed to stimulate a dialogue between the discussion of realistically complex emotional behavior and the neurobiological analysis of well-constrained aspects of motivation and emotion. Thus, in contrast to the above analysis, one might claim that there are no separate emotions, that all emotions are linked somehow, and that the experience of one emotion depends on the experience of another. I invite the reader to conduct a personal accounting of such alternatives.

4. Arbib and Lieblich (1977) used positive values for appetitive drives and negative values for aversive drives, but I will use positive values here even for negative drives.

5. Of course, at one level of analysis, one could construe, for example an autonomous system governed by an adaptive neural network as following a "program" expressed at the level of neural and synaptic dynamics. Conversely, a present-day personal computer will check for e-mail even as its user is actively engaged in some other activity, such as word processing. Thus, the notion of autonomy here is one of degree.

References

Arbib, M. A. (1966). Simple self-reproducing universal automata. *Information and Control*, 9, 177–189.

Arbib, M. A. (1985). *In search of the person: Philosophical explorations in cognitive science*. Amherst: University of Massachusetts Press.

Arbib, M. A. (1987). Levels of modeling of visually guided behavior (with peer commentary and author's response). *Behavioral and Brain Sciences*, *10*, 407–465.

Arbib, M. A. (1989). *The metaphorical brain 2: Neural networks and beyond*. New York: Wiley-Interscience.

Arbib, M. A. (1992). Book review: Andrew Ortony, Gerald L. Clore, & Allan Collins, *The cognitive structure of emotions*. *Artificial Intelligence*, *54*, 229–240.

Arbib, M. A. (2001). Co-evolution of human consciousness and language. In P. C. Marijuan (Ed.), *Cajal and consciousness: Scientific approaches to consciousness on the centennial of Ramón y Cajal's textura* (pp. 195–220). New York: New York Academy of Sciences.

Arbib, M. A. (2002). The mirror system, imitation, and the evolution of language. In C. Nehaniv & K. Dautenhahn (Eds), *Imitation in animals and artifacts*, The MIT Press, pp. 229–280.

Arbib, M. A. (2003). *Rana computatrix* to human language: Towards a computational neuroethology of language evolution. *Philosophical Transactions of the Royal Society of London. Series A*, *361*, 1–35.

Arbib, M. A., & Hesse, M. B. (1986). *The construction of reality*. Cambridge: Cambridge University Press.

Arbib, M. A., & Lieblich, I. (1977). Motivational learning of spatial behavior. In J. Metzler (Ed.), *Systems Neuroscience* (pp. 221–239). New York: Academic Press.

Arbib, M. A., & Liaw, J.-S. (1995). Sensorimotor transformations in the worlds of frogs and robots. *Artificial Intelligence*, *72*, 53–79.

Arbib, M. A., & Rizzolatti, G., (1997). Neural expectations: A possible evolutionary path from manual skills to language. *Communication and Cognition*, *29*, 393–424.

Betts, B. (1989). The T5 base modulator hypothesis: A dynamic model of T5 neuron function in toads. In J. P. Ewert & M. A. Arbib (Eds.) *Visuomotor coordination: Amphibians, comparisons, models and robots* (pp. 269–307). New York: Plenum.

Bischoff-Grethe, A., Crowley, M. G., & Arbib, M. A. (2003). Movement inhibition and next sensory state predictions in the basal ganglia. In VI (A. M. Graybiel, M. R. Delong, & S. T. Kitai (Eds.), *The basal ganglia* (Vol. VI, pp. 267–277). New York: Kluwer Plenum.

Blaney, P. H. (1986). Affect and memory: A review. *Psychological Bulletin*, 99, 229–246.

Brzoska, J., & Schneider, H. (1978). Modification of prey-catching behavior by learning in the common toad *(Bufo b bufo* L, Anura, Amphibia): Changes in response to visual objects and effects of auditory stimuli. *Behavioural Processes*, *3*, 125–136.

Christianson, S.-Å. (Ed.) (1992). *The handbook of emotion and memory: Research and theory*. Hillsdale, NJ: Erlbaum.

Cobas, A., & Arbib, M. (1992). Prey-catching and predator-avoidance in frog and toad: Defining the schemas. *Journal of Theoretical Biology*, *157*, 271–304.

Coetzee, J. M. (2003). *The Lives of Animals: ONE: The Philosophers and the Animals, being Lesson 3 of the novel Elizabeth Costello*. New York: Viking.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Grosset/Putnam.

Dean, P., Redgrave, P. (1989). Approach and avoidance system in the rat. In M. A. Arbib & J-P. Ewert (Eds.), *Visual structures and integrated functions*. Research Notes in Neural Computing. New York: Springer-Verlag.

Dean, P., Redgrave, P., & Westby, G. W. M. (1989). Event or emergency? Two response systems in the mammalian superior colliculus. *Trends in Neuroscience*, *12*, 138–147.

de Waal, F. (2001). *The ape and the sushi master: Cultural reflections by a primatologist*. New York: Basic Books.

Dominey, P. F., & Arbib, M. A. (1992). A cortico-subcortical model for generation of spatially accurate sequential saccades. *Cerebral Cortex*, 2, 153–175.

Ewert, J.-P. (1984). Tectal mechanisms that underlie prey-catching and avoidance behaviors in toads. In H. Vanegas (Ed.), *Comparative neurology of the optic tectum* (pp. 247–416). New York: Plenum.

Ewert, J.-P. (1987). Neuroethology of releasing mechanisms: Prey-catching in toads. *Behavioral and Brain Sciences*, *10*, 337–405.

Fagg, A. H., & Arbib, M. A. (1998). Modeling parietal–premotor interactions in primate control of grasping. *Neural Networks*, *11*, 1277–1303.

Fellous, J. M. (1999). Neuromodulatory basis of emotion. *Neuroscientist*, *5*, 283–294.

Fellous, J.-M., & Suri, R. (2003). Dopamine, roles of. In M. A. Arbib, (Ed.), *The handbook of brain theory and neural networks* (2nd ed., pp. 361–365). Cambridge, MA: Bradford MIT Press.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. London: Allen & Unwin.

Goldman-Rakic, P. S. (1996). The prefrontal landscape: Implications of functional architecture for understanding human mentation and the central executive. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *351*, 1445–1453.

Goodale, M. A., Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neuroscience*, *15*, 20–25.

Goodale, M. A., Milner, A. D., Jakobson, L. S., & Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature*, *349*, 154–156.

Grafton, S. T., Arbib, M. A., Fadiga, L., & Rizzolatti, G. (1996). Localization of grasp representations in humans by PET: 2. Observation compared with imagination. *Experimental Brain Research*, *112*, 103–111.

Grüsser-Cornehls, U., & Grüsser, O.-J. (1976). Neurophysiology of the anuran visual system. In R. Llinás & W. Precht (Eds.), *Frog neurobiology* (pp. 298–385). Berlin: Springer-Verlag.

Guazzelli, A., Corbacho, F. J., Bota, M., & Arbib, M. A. (1998). Affordances, motivation, and the world graph theory. *Adaptive Behavior*, *6*, 435–471.

Guazzelli, G., Bota, B., & Arbib, M. A. (2001. Competitive Hebbian learning and the hippocampal place cell system: Modeling the interaction of visual and path integration cues. *Hippocampus*, *11*, 216–239.

Heatwole, H., & Heatwole, A. (1968). Motivational aspects of feeding behavior in toads. *Copeia*, *4*, 692–698.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Heusser, H. (1960). Instinkterscheinungen an Kröten unter besonderer Berücksichtigung des Fortpflanzungsinstinktes der Erdkrote *(Bufo bufo* L). *Zeitschrift fur Tierpsychologie*, *17*, 67–81.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, *148*, 574–591.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*, 106–154.

Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, *28*, 229–289.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, *195*, 215–243.

Hubel, D. H., Wiesel, T. N., & LeVay, S. (1977). Plasticity of ocular dominance columns in monkey striate cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *278*, 377–409.

Humphrey, N. K. (1970). What the frog's eye tells the monkey's brain. *Brain, Behavior and Evolution*, *3*, 324–337.

Ingle, D., Schneider, G. E., Trevarthen, C. B., & Held, R. (1967). Locating and identifying: Two modes of visual processing (a symposium). *Psychologische Forschung, 31*, 1–4.

Jackson, J. H. (1878–79). On affections of speech from disease of the brain. *Brain, 1*, 304–330; *2*, 203–222, 323–356.

Kilmer, W. L., McCulloch, W. S., & Blum, J. (1969). A model of the vertebrate central command system. *International Journal of Man–Machine Studies, 1*, 279–309.

Köhler, W. (1927). *The mentality of apes* (E. Winter, Trans.). London: Routledge & Kegan Paul.

Kondrashev, S. L. (1976). Influence of the visual stimulus size on the breeding behavior of anuran males. *Akad Nayk Zool Zh, 55*, 1576–1579.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology, 16*, 37–68.

Lettvin, J. Y., Maturana, H., McCulloch, W. S., & Pitts, W. H. (1959). What the frog's eye tells the frog brain. *Proceedings of the IRE, 47*, 1940–1951.

Lieblich, I., & Arbib, M. A. (1982). Multiple representations of space underlying behavior. *Behavioral and Brain Sciences, 5*, 627–659.

Magoun, H. W. (1963). *The waking brain* (2nd ed.). Springfield, IL.: Thomas.

Mountcastle, V. B., & Powell, T. P. S. (1959). Neural mechanisms subserving cutaneous sensibility, with special reference to the role of afferent inhibition in sensory perception and discrimination. *Bulletin of Johns Hopkins Hospital, 105*, 201–232.

Olds, J. (1977). *Drives and Reinforcements: Behavioral studies of hypothalamic functions*. New York: Raven.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. New York: Cambridge University Press.

Panksepp, J. (1998). *Affective neuroscience*. New York: Oxford University Press.

Prescott, T. J., Gurney, K., & Redgrave, P. (2003). Basal ganglia. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (2nd ed., pp. 147–151). Cambridge, MA: Bradford MIT Press.

Pribram, K. H. (1960). A review of theory in physiological psychology. *Annual Review of Psychology, 11*, 1–40.

Risold, P. Y., Thompson, R. H., & Swanson, L. W. (1997). The structural organization of connections between hypothalamus and cerebral cortex. *Brain Research. Brain Research Reviews, 24*, 197–254.

Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences, 21*, 188–194.

Rizzolatti, G., Camarda, R., Fogassi, L., Gentilucci, M., Luppino, G., & Matelli, M. (1988). Functional organization of inferior area 6 in the macaque monkey II. Area F5 and the control of distal movements. *Experimental Brain Research, 71*, 491–507.

Rizzolatti, G., Fadiga L., Gallese, V., & Fogassi, L. (1995). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*, 131–141.

Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Perani, D., & Fazio, F. (1996).

Localization of grasp representations in humans by positron emission tomography: 1. Observation versus execution. *Experimental Brain Research*, *111*, 246–252.

Rizzolatti, G., & Luppino, G. (2001). The cortical motor system. *Neuron*, *31*, 889–901.

Rizzolatti, G., & Luppino, G. (2003). Grasping movements: Visuomotor transformations. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (2nd ed., pp. 501–504). Cambridge, MA: MIT Press.

Rizzolatti, G., Luppino, G., & Matelli, M. (1998). The organization of the cortical motor system: New concepts. *Electroencephalography and Clinical Neurophysiology*, *106*, 283–296.

Rolls, E. T. (1999). *The brain and emotion*. Oxford: Oxford University Press.

Rolls, E. T. (2000). Neurophysiology and functions of the primate amygdala, and the neural basis of emotion. In J. P. Aggleton (Ed.), *The amygdala: A functional analysis* (2nd ed., pp. 447–478). Oxford: Oxford University Press.

Scheibel, M. E., & Scheibel, A. B. (1958). Structural substrates for integrative patterns in the brain stem reticular core. In H. H. Jasper et al. (Eds.), *Reticular formation of the brain* (pp. 31–68). Boston: Little, Brown.

Schneider, G. E. (1969). Two visual systems. *Science*, *163*, 895–902.

Schultz, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience*, *1*, 199–207.

Stoerig, P. (2001). The neuroanatomy of phenomenal vision: A psychological perspective. *Annals of the New York Academy of Science*, *929*, 176–194.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Swanson, L. W. (2000). Cerebral hemisphere regulation of motivated behavior. *Brain Research*, *886*, 113–164.

Swanson, L. W., & Mogenson, G. J. (1981). Neural mechanisms for the functional coupling of autonomic, endocrine and somatomotor responses in adaptive behavior. *Brain Research*, *228*, 1–34.

Taira, M., Mine, S., Georgopoulos, A. P., Murata, A., & Sakata, H. (1990). Parietal cortex neurons of the monkey related to the visual guidance of hand movement. *Experimental Brain Research*, *83*, 29–36.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behavior*. Cambridge, MA: MIT Press.

von Neumann, J. (1966). *Theory of self-reproducing automata*. (edited and completed by A.W. Burks). Urbana: University of Illinois Press.

Watts, A. G. (2001). Neuropeptides and the integration of motor responses to dehydration. *Annual Review of Neuroscience*, *24*, 357–384.

Watts, A. G. (2003). Motivation, neural substrates. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (2nd ed., pp. 680–683). Cambridge, MA: MIT Press.

Weiskrantz, L., Warrington, E. K., Sanders, M. D., & Marshall, J. (1974). Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain*, *97*, 709–728.

Wells, K. D. (1977). The social behavior of anuran amphibians. *Behavior*, *25*, 666–693.

Wiener, N. (1950). *The human use of human beings*. Boston: Houghton Mifflin.

Wiener, N. (1961). *Cybernetics: Or control and communication in the animal and the machine* (2nd ed.). Cambridge, MA: MIT Press.

Wiesel, T. N., & Hubel, D. H. (1963). Effects of visual deprivation on morphology and physiology of cells in the cat's lateral geniculate body. *Journal of Neurophysiology*, *26*, 978–993.

This page intentionally left blank

Index

Page numbers in **bold** indicate illustrations, figures, or tables. Page numbers followed by n indicate information in chapter endnotes, for example 198n.4 indicates note four on page 198.