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

Oxford is a registered trademark of Oxford University Press

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.