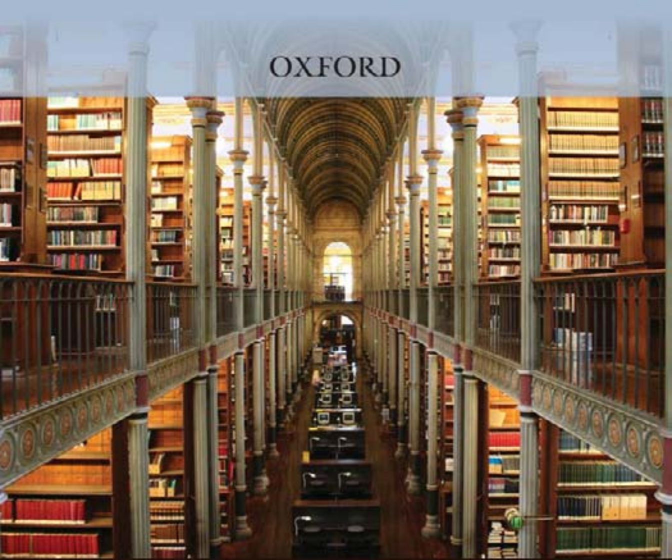
OXFORD



Social Neuroscience

Toward Understanding The
Underpinnings of The Social Mind

Alexander Todorov
Susan T. Fiske
Deborah A. Prentice

Social Neuroscience

Oxford Series in Social Cognition and Social Neuroscience

Series Editor: Ran R. Hassin

The New Unconscious

Edited by Ran R. Hassin, James S. Uleman, and John A. Bargh

Oxford Handbook of Human Action

Edited by Ezequiel Morsella, John A. Bargh, and Peter M. Gollwitzer

Self Control in Society, Mind, and Brain

Edited by Ran R. Hassin, Kevin Ochsner, and Yaacov Trope

Social Neuroscience

Edited by Alexander Todorov, Susan T. Fiske, and Deborah A. Prentice

Social Neuroscience

TOWARD UNDERSTANDING THE UNDERPINNINGS OF THE SOCIAL MIND

Edited by

ALEXANDER TODOROV

SUSAN T. FISKE

DEBORAH A. PRENTICE

OXFORD
UNIVERSITY PRESS
2011

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

Todorov, Alexander B.
Social neuroscience : toward understanding the underpinnings of the social mind / Alexander Todorov,
Susan T. Fiske & Deborah Prentice. — 1st ed.
p. cm. — (Oxford series in social cognition and social neuroscience)
Includes bibliographical references and index.
ISBN 978-0-19-531687-2
1. Cognitive neuroscience. 2. Social psychology. I. Fiske, Susan T.
II. Prentice, Deborah A. III. Title.
QP360.5.T63 2011
612.8′233—dc22
2010009165

9 8 7 6 5 4 3 2 1

Printed in China

CONTENTS

I. UNDERSTANDING AND REPRESENTING OTHER PEOPLE

II. UNDERSTANDING AND REPRESENTING SOCIAL GROUPS

III. REGULATION OF SOCIAL BEHAVIOR

IV. NAVIGATING SOCIAL LIFE

CONTRIBUTORS

Reginald B. Adams, Jr., PhD
Department of Psychology
The Pennsylvania State University
University Park, PA

Nalini Ambady, PhD
Psychology Department
Tufts University
Medford, MA

David M. Amodio, PhD
Assistant Professor
Department of Psychology and Center for Neural Science
New York University
New York, NY

Antoine Bechara, PhD
Professor
Department of Psychology
University of Southern California
Los Angeles, CA

Jennifer S. Beer
Assistant Professor
Department of Psychology
University of California
Davis, CA

Gary G. Berntson, PhD
Professor
Department of Psychology
Ohio State University
Columbus, OH

Jamil P. Bhanji
Department of Psychology
University of California
Davis, CA

John T. Cacioppo, PhD
Tiffany and Margaret Blake Distinguished Service Professor
Director, Center for Cognitive and Social Neuroscience, and
Director, Arete Initiative of the Office of the Vice President for Research and National Laboratories
University of Chicago
Chicago, IL

William A. Cunningham
Associate Professor
Department of Psychology
The Ohio State University
Columbus, OH

Hanna Damasio M.D.
Department of Psychology
Dana Dornsife Professor of Neuroscience
Director of the Dana and David Dornsife Cognitive Neuroscience Imaging
Center at the University of Southern California

Naomi I. Eisenberger, PhD
Assistant Professor
Department of Psychology
University of California
Los Angeles, CA

Susan T. Fiske
Professor
Department of Psychology
Princeton University
Princeton, NJ

Maria Ida Gobbini
Department of Psychology
University of Bologna
Bologna, Italy

Joshua D. Greene
Assistant Professor of Psychology
Director of the Moral Cognition Lab
Harvard University

Eddie Harmon-Jones
Texas A&M University
Department of Psychology
College Station, TX

Cindy Harmon-Jones
Texas A&M University
Department of Psychology
College Station, TX

Lasana T. Harris, PhD
Assistant Professor
Department of Psychology and Neuroscience
Duke University
Durham, NC

Louise C. Hawkley, PhD
Center for Cognitive and Social Neuroscience
University of Chicago
Chicago, IL

James V. Haxby
Evans Family Distinguished Professor
Director, Center for Cognitive Neuroscience
Department of Psychological and Brain Sciences
Dartmouth College
Hanover, NH

Todd F. Heatherton
Department of Psychological and Brain Sciences
Dartmouth College
Hanover, NH

Tiffany A. Ito, PhD
Associate Professor
Department of Psychology and Neuroscience
University of Colorado
Boulder, CO

Adrianna C. Jenkins
Department of Psychology
Harvard University
Cambridge, MA

Marcia K. Johnson, PhD
Professor
Department of Psychology
Yale University
New Haven, CT

Amanda Kesek, MA
PhD Candidate
Institute of Child Development
University of Minnesota
Minneapolis, MN

Matthew D. Lieberman, PhD
Professor
Psychology, Psychiatry and Biobehavioral Sciences
University of California,
Los Angeles, CA

Jason P. Mitchell, PhD
Assistant Professor
Department of Psychology
Harvard University
Cambridge, MA

Kevin Ochsner
Associate Professor
Department of Psychology
Columbia University
New York, NY

Dominic J. Packer, PhD
Assistant Professor
Department of Psychology
Lehigh University
Bethlehem, PA

Elizabeth A. Phelps, PhD
Silver Professor
Psychology and Neural Science
New York University
New York, NY

Deborah A. Prentice
Alexander Stewart 1886
Professor of Psychology
Department of Psychology
Princeton University
Princeton, NJ

James K. Rilling, PhD
Associate Professor
Department of Anthropology and Department of Psychiatry and Behavioral Sciences
Emory University
Atlanta, GA

Alexander Todorov
Associate Professor of Psychology and Public Affairs
Department of Psychology and Woodrow
Wilson School of Public and International Affairs
Princeton University
Princeton, NJ

Daniel Tranel, PhD
Professor
Department of Neurology
University of Iowa
Iowa City, IA

Jamil Zaki
PhD Candidate
Department of Psychology
Columbia University
New York, NY

This page intentionally left blank

Introduction

Alexander Todorov, Susan T. Fiske, & Deborah A. Prentice

The field of social cognitive neuroscience has captured the attention of many psychologists in the last 10 years (Ocksner & Lieberman, 2001). Although a few prominent social psychologists (Blascovich & Mendes, 2010; Cacioppo 1994; Cacioppo et al. 2001) pioneered the use of psychophysiological methods in the 1990s and earlier, the real upsurge of social neuroscience research started at the beginning of this decade (Lieberman, 2010). A quick search on PsycINFO, the major search database for psychologists, reveals more than 350 publications referring to social neuroscience. All but 7 of them were published after 2000. Since then, social neuroscience research has been represented at every major social psychology conference. The increased interest in social neuroscience is also evident from the publication of many special journal issues dedicated to Social Neuroscience (*Neuropsychologia*, *NeuroImage*, *Journal of Cognitive Neuroscience*, *Journal of Personality and Social Psychology, Brain Research* to name a few journals), and the launching of two new journals dedicated to social cognitive neuroscience research (*Social Cognitive and Affective Neuroscience* and *Social Neuroscience).*

Undoubtedly, much of the spur for this new field came from the development of functional neuroimaging methods, making possible unobtrusive measurement of brain activation over time. Within 30 years, research questions have moved from simple validation questions—would flashing stimuli activate visual cortex (Lassen, Ingvar, & Skinhoj, 1978)—to questions about functional specialization of brain regions—are there regions in the inferior temporal cortex dedicated to face processing (Kanwisher, McDermott, & Chun, 1997; McCarthy, Puce, Gore, & Allison, 1997)—to questions that would have been considered intractable at that level of analysis just a decade ago. These "intractable" questions are the focus of the chapters of this book. How do we understand and represent other people? How do we represent social groups? How do we regulate our emotions and often socially undesirable responses?

The objective of this book is to introduce social cognitive neuroscience research that addresses questions of fundamental importance to social psychology, combining multiple methodologies in innovative ways. These methodologies include behavioral experiments, computer modeling, functional Magnetic Resonance Imaging (fMRI) experiments, event-related potential (ERP) experiments, and brain lesion studies. The book is divided into four sections. The first section deals with understanding and representing other people. The second section deals with representing social groups. The third section deals with the interplay of cognition and emotion in social regulation. The final section considers a range of novel questions that emerged in the context of social neuroscience research: understanding social exclusion as pain, deconstructing our moral intuitions, understanding cooperative exchanges with other agents, and the effect of aging on brain function and its implications for well-being.

One of the fundamental problems of social cognition is how we represent and understand other people. The first section of the book contains four chapters exploring this problem. Jenkins and Mitchell describe recent research suggesting that there are neural systems dedicated to processing of social information. Zaki and Ochsner review the subtle similarities and differences between the neural representations of self and others. Faces are some of the most salient social stimuli, and there is wealth of research on face perception with a primary focus on regions in exstrastriate visual cortex (Haxby, Hoffman, & Gobbini, 2001). However, as Gobbini explains in her chapter, the neural systems underlying face perception extend beyond these "core" regions. In particular, faces of significant others activate the same regions that are often observed in task-requiring inferences of mental states (Jenkins & Mitchell, this volume; Zaki & Ochsner, this volume). Todorov outlines a framework for understanding how unfamiliar faces are rapidly and automatically evaluated on social dimensions. The section concludes with Haxby's commentary on the chapters, in which he considers the central questions for understanding the social brain.

Another fundamental problem of social cognition is how we represent social groups. The second section of the book contains three chapters exploring this problem. Using ERP methods, Ito describes research demonstrating that inferences of social category membership are made within a few hundred milliseconds exposure to a face. Using similar methods, Amodio outlines a set of precise hypotheses about mechanisms of cognitive control of prejudiced responses. Harris and Fiske describe research showing that different social categorizations (Fiske, Cuddy, & Glick, 2007) result in very different neural responses, with members of highly stigmatized groups eliciting less activation in regions involved in understanding the minds of other people (Jenkins & Mitchell, this volume). The section concludes with Ambady and Adams's commentary on the chapters, in which they consider outstanding questions in this research, with a particular emphasis on the integration of multiple social cues in perception.

One of the major contributions of fMRI and other neuroscience methodologies to social psychology has been to put affect and cognition on similar footing. The third section of the book contains four chapters exploring the interplay of affect and cognition in social regulation. Packer, Kesek, and Cunningham review the complexities of dynamic interaction between evaluative, automatic processes and reflective, controlled processes. Using data from patients with lesions in the orbitofrontal cortex, Beer and Bhanji review research showing the various ways in which these lesions impair social functioning. E. Harmon-Jones and C. Harmon-Jones show the importance of not confounding valence (positive vs. negative) and motivational direction (approach vs. avoidance) for understanding social cognition. Lieberman considers the control mechanisms through which labeling negative feelings could lead to improved emotional well-being. The section concludes with Phelps's commentary on the chapters, in which she reflects on how the study of emotion further contributes to the interdisciplinary nature of social neuroscience.

The final section of the book contains four chapters exploring novel research topics that emerged in the context of social cognitive neuroscience research. Rilling reviews fMRI research in which participants engage in social interactions. Eisenberger reviews research demonstrating a surprising overlap between neural regions implicated in the experience of physical pain and regions implicated in the experience of social pain resulting from social exclusion. Cacioppo and his colleagues propose an interesting hypothesis explaining the decreased levels of depressive symptoms and increased well-being in older healthy adults by considering how the amygdala's sensitivity to positive and negative stimuli changes with aging. Drawing from behavioral and fMRI studies, Greene argues that the "core competence of the soul"—moral judgments—can be understood in mechanistic terms. The section concludes with a commentary by Heatherton, in which he considers what should be the building components of the social brain.

The book concludes with a commentary by Johnson, one of the pioneers of cognitive neuroscience research, in which she reflects on the short history and long future of social cognitive neuroscience.

We very much hope that you will enjoy this book and develop an interest in an exciting, rapidly developing, and expanding field that promises a richer and deeper understanding of the social mind.

References

Blascovich, J., & Mendes, W. B. (2010). Social psychophysiology and embodiment. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of Social Psychology* (5th ed.). New York: Wiley.

Cacioppo, J. T. (1994). Social neuroscience: Autonomic, neuroendocrine, and immune responses to stress. *Psychophysiology, 31*, 113–128.

Cacioppo, J. T., Berntson, G. G., Adolphs, R., et al. (Eds.) (2001). *Foundations of Social Neuroscience*. Cambridge, MA: MIT Press.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences, 11*, 77–83.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*, 223–233.

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.

Lassen, N. A., Ingvar, D. H., & Skinhoj, E. (1978). Brain function and blood flow. *Scientific American, 239*(4): 62–71.

Lieberman, M. D. (2010). Social cognitive neuroscience. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of Social Psychology* (5th ed.). New York: Wiley.

McCarthy, G., Puce, A., Gore, J.C., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neurosciences, 9*, 605–610.

This page intentionally left blank

PART I

Understanding and Representing Other People

This page intentionally left blank

CHAPTER 1
How Has Cognitive Neuroscience Contributed to Social Psychological Theory?

Adrianna C. Jenkins & Jason P. Mitchell

Even the most casual follower of developments within social psychology is unlikely to have missed the recent surge of studies adapting the methods of cognitive neuroscience to questions about the nature of human social cognition. The sudden growth spurt enjoyed by this enterprise is perhaps best indexed by the striking number of special issues, edited volumes, and conferences devoted to neuroscientific approaches to the study of social psychology. Moreover, these new methods have been applied to a dazzling variety of theoretical questions, from studies that use brain imaging to revisit long-standing questions of social psychological interest—such as those relating to attribution (Harris, Todorov, & Fiske, 2005) or impression formation (Mitchell, Cloutier, Banaji, & Macrae, 2006; Mitchell, Macrae, & Banaji, 2004, 2005)—to those using such methods to develop new theoretical insights about the ways humans go about navigating their social environment.

With this surge of neuro-imaging studies have come novel theoretical contributions to social psychological theory. This chapter reviews three such contributions. The first has been the somewhat unexpected observation that social cognition consistently elicits a distinct pattern of brain activity that distinguishes it from non-social cognition, strongly suggesting that the mental operations giving rise to human social abilities do not simply "piggyback" on general-purpose cognitive processes but instead rely on a set of processes specialized for social thought (Blakemore, Winston, & Frith, 2004; Adolphs, 2003; Mitchell, Heatherton, & Macrae, 2002; Mitchell, Macrae, & Banaji, 2004).

Second, recent neuro-imaging work has revitalized a question that, although of central importance to social cognition, has been relatively understudied by social psychologists—namely, what are the mechanisms that allow one person to successfully infer the mental states (thoughts, feelings, motivations) of others? By translating this question into a common, brain-based parlance, this recent approach has succeeded in putting social psychologists in touch with relevant concepts discussed earlier by philosophers and cognitive scientists—for example, "simulation" approaches to the problem of understanding other minds.

Third, perhaps the most unique contribution made to social psychology by the use of neuro-imaging has been the observation that brain regions subserving social cognition appear to have a special status in the brain. Specifically, these regions tend to have unusually high levels of activity even when perceivers are ostensibly at rest, suggesting that the human brain may have a special propensity for social thought. In the remainder of this chapter, we review various neuro-imaging data that have made novel contributions of these kinds to the study of social cognition, generating new insights into the ways in which human beings navigate their social world.

Contribution 1: the fundamental difference between social and nonsocial cognition

Although understanding the behavior of other people is crucial to our daily functioning and survival, making sense of other minds poses one of the greatest challenges to human cognition. Compared to the causes of everyday physical phenomena—falling tree branches or a waning moon—the factors influencing the behavior of people are far more elusive and far less consistent. Still, although we never have direct access to another person's mind, we are able to make sense of others by accounting for the fact that each of them has unique intentions, desires, beliefs, and motivations—that is, mental states like our own—guiding their behavior (Dennett, 1987). In other words, despite fairly impoverished input about the contents of others' mental states, perceivers are able to reconstruct the goings-on of other minds with tremendous richness and complexity.

Given the unique demands of thinking about other people, psychologists have become increasingly interested in whether social cognition can be accomplished on the basis of the same mental processes we deploy for nonsocial purposes or whether we instead draw on a set of processes specifically responsible for meeting the unique cognitive demands of interpersonal interaction. As Adolphs (2003) recently asked, "Are there processes, are there neural structures, that are in some way designed, specialized, and best understood as subserving the perception of socially relevant stimuli and the guidance of social behavior?" (p. 124). Such questions have long been a source of inquiry not only to social psychologists but also among developmental psychologists considering the components of a theory of mind, comparativists probing the unique features of human cognition, and clinicians seeking to understand the roots of social disorders such as autism.

Indeed, research with neuropsychological patients has previously hinted at some sort of dissociation between social and nonsocial reasoning. The now-fabled tale of Phineas Gage, the construction foreman who suffered a pervasive disruption of social functioning following insult to ventral aspects of medial prefrontal cortex, provided perhaps the first scientific evidence of such a dissociation. Dr. John Harlow, who studied Gage after his accident, provided a detailed account of Gage's abrupt change in personality and social abilities in the relative absence of deficits outside the social realm (Harlow, 1868; *see also* Damasio, 1994). Present-day patients with damage to this same ventromedial region of prefrontal cortex have also been observed to exhibit selective social and emotional impairments on a variety of experimental tasks (Bar-On, Tranel, Denburg, & Bechara, 2003; Shamay-Tsoory, Aharon-Peretz, Tomer, & Berger, 2003; Bechara, Damasio, Tranel, & Damasio, 1997) as well as in everyday life (Anderson, Tranel, Barrash, & Bechara, 2006; Barrash, Tranel, & Anderson, 2000). Finally, additional neuropsychological evidence in favor of a dissociation between social and nonsocial cognition has come from studies of individuals with autism, who often have stark deficits in social functioning despite proficiency in nonsocial areas (Baron-Cohen, O'Riordan, Stone, & Plaisted 1999; Peterson & Siegal, 1998; Happé, 1995; Leslie & Thiass, 1992; Zaitchik, 1990; Mundy, Sigman, Ungerer, & Sherman 1986; Baron-Cohen, Leslie, & Frith, 1986).

Recent studies have used neuro-imaging techniques to extend these neuropsychological observations, providing converging evidence that social cognition recruits a set of brain regions over and above those recruited by nonsocial tasks. Specifically, social-cognitive tasks have been observed to preferentially recruit a consistent set of neural regions: medial prefrontal cortex (mPFC), the superior temporal sulcus (STS), medial parietal cortex (precuneus), and lateral parietal cortex, including the temporoparietal junction (TPJ). In the first studies using neuro-imaging to examine mental state inference, Goel et al. (1995) and Fletcher et al. (1995) observed increased activity in these regions when participants considered a person's mental state. Specifically, Goel and colleagues asked participants to assess whether or not an historical figure (Christopher Columbus) would know how to use various objects (e.g., a broom, a radio), comparing brain activity during this task to that elicited when participants instead considered the physical properties of the same objects. Fletcher and colleagues compared patterns of neural activation when participants read stories requiring either a mental state attribution or an attribution of physical causality to be understood. Beginning with these initial studies, the observation that these particular regions (especially mPFC) are differentially engaged by tasks requiring mental state attribution has been among the most consistent effects in cognitive neuroscience.

More recent neuro-imaging work has directly probed the separateness of social cognition in the human brain by following two complementary approaches. A first approach has held constant perceivers' task while varying the social relevance of the targets (i.e., making the same kind of judgment about social vs. nonsocial entities). For example, Mitchell, Heatherton, and Macrae (2002) asked participants to consider whether each in a series of adjectives could appropriately be used to describe either people (represented by proper names) or inanimate objects (either articles of clothing or kinds of fruit). Analyses investigated whether distinct neural regions subserve semantic knowledge about different classes of entities. Interestingly, whereas the same regions appeared to be involved in semantic knowledge of fruit and clothing, the comparisons between judgments of people and both kinds of objects revealed strikingly distinct patterns of neural activity. Specifically, judgments of objects differentially activated a set of regions previously implicated in semantic knowledge, including left ventrolateral prefrontal cortex and left inferotemporal cortex. In contrast, judgments of people engaged a qualitatively distinct set of regions that included dorsal and ventral mPFC, right lateral parietal cortex, and left superior temporal cortex. That is, although the task itself remained identical in all cases ("can this adjective describe this noun?"), the pattern of brain activity elicited by judgments of people was qualitatively distinct from that elicited by judgments of objects, overlapping considerably with the set of brain regions that has been implicated previously in other social-cognitive tasks.

Studies by Gallagher et al. (2002) and McCabe et al. (2001) also examined differences linked to social cognition by holding perceivers' task constant while manipulating the targets they considered. In these studies, participants were scanned while playing computerized games against one of two opponents: either a computer or what they believed to be another person (in actual fact, participants always played against a computer). Despite playing the same game throughout the study, two distinct patterns of neural activity emerged as a function of whether participants believed they were interacting with a computer or another human being. Again, it was only when participants believed they were playing against another person that the particular set of brain regions subserving social cognition—including mPFC and STS—became engaged.

Saxe and colleagues (Saxe & Kanwisher, 2003; Saxe & Wexler, 2005; Saxe & Powell, 2006) have observed similar results using yet a third way to modulate social-cognitive processing within a single task that manipulates the target of judgment. Participants in these studies read short vignettes that involve either "false belief" scenarios (in which perceivers must attribute a belief to a target that is inconsistent with the actual state of the world) or "false photograph" scenarios (in which perceivers must recognize that a camera or map contains a representation that is no longer consistent with reality). These two types of scenarios are structurally similar in that perceivers must access a representation that conflicts with what they know to be true but differ in whether or not this false representation exists in the mind of another person. Consistent with the claim that social cognition recruits a distinct set of cognitive processes, false belief tasks reliably elicit greater activity (relative to false photograph tasks) in areas implicated in social cognition: mPFC and TPJ.

In contrast to studies that keep the task constant while manipulating the target of participants' judgments, a second approach has held targets constant while varying the nature of the task. As such, these studies have examined neural differences across the various dimensions along which perceivers can judge other people, including their mental states and their physical attributes. For example, one such study (Mitchell, Macrae, & Banaji, 2004) examined the neural correlates of memory as a function of a social or nonsocial orienting task. While undergoing fMRI scanning, participants viewed sentences containing information about recent events in a person's life (e.g., "closed the elevator door before anyone else could get on"), during which time they were alternately instructed to complete one of two tasks: form an impression of the person (*impression formation* trials) or remember the order in which the events occurred (*sequencing* trials). Participants subsequently performed a memory task, allowing the neural activity associated with items that later went on to be remembered (hits) to be segregated from that associated with items that later when on to be forgotten (misses), as a function of whether the item was initially encountered as part of the social or nonsocial orienting task.

Analyses revealed two main findings. First, impression formation was associated with significantly greater activity in dorsal mPFC (compared to sequencing trials), extending previous findings by suggesting that what drives this region's response is not the mere presence of a person but specifically the consideration of a person's mental states. Second, two distinct brain regions were associated with subsequent memory. Whereas successful memory for information encoded in the *impression formation* task was associated with activity in mPFC, memory for information encoded in the *sequencing* task was associated with activity in the right hippocampus. That is, a single factor—the relative social demands of the orienting task—was enough to determine which brain regions distinguished successful from unsuccessful memory encoding.

A follow-up study (Mitchell, Macrae, & Banaji, 2005) strengthened these conclusions by combining the two approaches—that is, by systematically varying the social relevance of both the targets and the tasks. In this experiment, participants were similarly asked to either form impressions of the target or remember the order in which events occurred; however, targets in this case were either people or inanimate objects (cars and computers). Stimulus sentences described events that involved either a person (e.g., "carried the old woman's groceries across the street") or an object ("recently received a new paint job"). Supporting the conclusion that the observed neural dissociation in the previous study hinged upon the social versus nonsocial nature of the judgment (and not on any particular features of impression formation tasks more generally), mPFC was preferentially engaged by the specific combination of the social orienting task and the socially relevant target—that is, impression formation about other people (relative to forming impressions about cars and computers and to sequencing events about cars, computers, and people).

Taken together, these studies supply a growing body of evidence that a specific set of neural regions subserves the task of navigating through the complex world of human social interaction. Importantly, this evidence could not have come solely from studies of brain-damaged patients, neuropsychological studies of individuals with autism or other disorders, or studies involving behavioral tasks alone. Neuro-imaging has facilitated experiments that contrast social and nonsocial reasoning within a single, healthy individual, leading to repeated observations that even the seemingly fundamental cognitive abilities of categorization, problem-solving, and memory may recruit different cognitive processes when deployed for social as opposed to nonsocial purposes.

Contribution 2: the component processes of social cognition

Having homed in on the particular set of neural regions responsible for social cognition, researchers have recently begun using neuroimaging to identify the specific cognitive processes of which social cognition consists. How exactly does one person go about making inferences about the mental states of another? A surge of recent studies has begun to shed new light on questions about what cognitive mechanisms support our ability to reason about the minds of other people.

Over the past 20 years, two contrasting cognitive accounts have emerged to explain how humans might go about inferring the mental states of others. On one hand, a class of ideas known as *simulation theory* (Heal, 1986; Gordon, 1986) suggests that perceivers may use their own minds as a kind of "model" for the mind of another person, drawing on their own thoughts, experiences, and reactions when anticipating or inferring those of another. This view rests on the appreciation that in inferring the contents of the mind of another person, each of us has a powerful resource at our disposal: our own mind. In short, simulation accounts suggest that perceivers tap into self-knowledge when considering others' mental states.

In contrast, a view known as *theory-theory* posits that perceivers instead rely on more objective, less personal knowledge when inferring others' mental states (Gopnik & Wellman, 1992, 1994; Gopnik & Meltzoff, 1997). This view suggests that throughout development, humans accumulate something akin to a set of "laws" about the behavior of other humans, such as "people act based on their needs" or "if a person has not slept in many hours, she will become tired and need to sleep." Perceivers then deploy these "laws" to predict and make sense of human behavior, much as one might use the laws of Newtonian mechanics to predict and make sense of the behavior of objects.

Although simulation and theory-theory have historically been portrayed as mutually exclusive, emerging evidence suggests that perceivers likely rely on a combination of both approaches when making sense of other minds (Ames, 2004a, 2004b; Mitchell, Banaji, & Macrae, 2005; *see* Saxe, 2005, and Mitchell, 2005, for discussion). Most recently, we have used neuro-imaging to begin to disentangle the circumstances under which the mind may draw on simulation and theory-like approaches to infer the mental states of others. In doing so, we have capitalized on a central prediction of simulation theory—namely, that using self-reference to understand another's mind only serves as an appropriate strategy if perceivers assume that the person in question experiences the same mental states in roughly the same situations as they themselves do. In other words, simulation should be a useful cognitive strategy for understanding others only to the extent that a perceiver believes that a target is relevantly similar to self. As such, perceived similarity between self and other should moderate the extent to which an individual mentalizes in a self-referential manner.

Across three studies, we have found evidence suggesting that the brain distinguishes between mentalizing about similar versus dissimilar others. In an initial fMRI study (Mitchell et al., 2005), participants made two kinds of judgments about individuals in a series of photographs, either a *mentalizing* judgment that oriented participants to targets' mental states (how pleased was this person to have his/her photo taken?) or a *nonmentalizing* judgment that instead oriented participants to targets' physical characteristics (how symmetrical is the person's face?). After scanning, participants considered each target a second time and were asked to indicate how similar they perceived the person to be to themselves (using a 4-point scale).

Analyses identified brain areas in which activity correlated with subsequent ratings of similarity, revealing that a region of ventral mPFC responded preferentially to photographs of people perceived to be similar to the participant but only during the mentalizing task. Importantly, ventral mPFC has been implicated repeatedly in earlier studies of self-referential thought, in which participants were instructed to report on their own, first-person mental states, preferences, or personality traits (Schmitz, Kawahara-Baccus, & Johnson, 2004; Vogeley et al., 2004; Johnson et al., 2002; Kelley et al., 2002; Gusnard, Akbudak, Shulman, & Raichle, 2001). Interestingly, we observed preferential activation of this ventral mPFC region during mentalizing about similar others although participants were instructed neither to attend to their own mental states nor to consider the similarity of the targets during scanning. The observed overlap between brain regions engaged during self-referential thought and mentalizing about similar others suggests that when considering the mental state of a person who is perceived to be similar to themselves, perceivers simultaneously engage in self-referential thinking. Overall, this pattern of data is consistent with the possibility that perceivers make reference to some aspects of self-knowledge when mentalizing and that they do so specifically for similar others, two key predictions of simulation accounts of social cognition.

A second study (Mitchell, Macrae, & Banaji, 2006) extended these observations by asking participants to make a variety of judgments about three targets: a person with liberal political views, a person with conservative political views, and themselves. Participants were scanned while judging how likely each target would be to hold each in a series of opinions and preferences (e.g., "get frustrated sitting in traffic"). Participants made each judgment three times: once for each target (liberal, conservative) and one trial on which they indicated their own opinion about the statement. After scanning, the extent to which each subject identified with each target was assessed with a version of the Implicit Association Test (IAT) that measured the degree to which the participant more closely associated self with the liberal versus the conservative target. This IAT measure allowed us to retroactively designate a "similar" and "dissimilar" target for each participant, facilitating analyses of brain regions that responded preferentially during judgments of those perceived to be similar versus dissimilar. Analyses revealed a double dissociation between the regions of mPFC associated with judgments of similar versus dissimilar others. Whereas thinking about the dissimilar other was associated with activity in a *dorsal* aspect of mPFC, thinking about the similar other was associated with activity in a more *ventral* mPFC region, nearly identical to the one observed for similar others in the first study.

Finally, a third study used a repetition suppression paradigm to provide converging evidence that perceivers engage in self-referential thought during mentalizing about similar others. Repetition suppression is the observation that activity in the brain region(s) associated with a given process is typically reduced upon repeated engagement in that process (Grill-Spector, Henson, & Martin, 2006). In this study (Jenkins, Macrae, & Mitchell, 2008), we investigated whether judgments about self and similar (but not dissimilar) others could be thought to recruit "the same process." Following a similarity manipulation identical to that of Mitchell et al. (2006), participants made rapid, mentalistic judgments about a similar target, a dissimilar target, and self (e.g., how much does the person enjoy skiing?) in pairs. In a given pair, participants first made a judgment about one of the targets, immediately followed by a judgment about self. Following the speeded judgment task, participants completed an independent "self-localizer" task (Kelley et al., 2002), in which they rated the applicability of a series of personality traits to themselves and to a familiar but not personally known other (George W. Bush), allowing us to identify a region of ventral mPFC that responded preferentially to judgments of self compared to judgments of another person. We then interrogated this region with respect to the speeded judgment task on which participants had judged themselves and the similar and dissimilar targets. Consistent with the hypothesis that perceivers spontaneously engage in self-referential thought when making mentalistic judgments about similar (but not dissimilar) others, activity in ventral mPFC was reduced when perceivers made judgments of self immediately following judgments of similar others, and no such suppression was observed for dissimilar-self pairs. These data suggest that judgments of similar others, unlike judgments of dissimilar others, proceed in a self-referential manner, providing further support for an account on which different processes are recruited for mentalizing under different circumstances.

These studies make three main contributions to our understanding of the processes of mental state inference. First, perceivers seem to automatically and implicitly assess the similarity of another person to self when making judgments about other minds. Second, distinct cognitive processes are recruited when performing identical mentalizing tasks for targets who are perceived to be similar versus dissimilar. Finally, the cognitive processes recruited when considering the mental states of a similar other may overlap with those implicated in self-referential thought. Taken together, these findings suggest that simulation may be a powerful tool for understanding other minds but that, like any tool, it may be best suited to a specific purpose: in this case, understanding the mental propensities of the minds we take to be most like our own.

Contribution 3: the primacy of social cognition

In the first section of this chapter, we reviewed evidence suggesting that social cognition draws on a distinct set of cognitive processes that are dissociable from those involved in nonsocial thought. Although we suggest that this observation is noteworthy in its own right, perhaps the most intriguing contribution of cognitive neuroscience to our understanding of this "social network" comes from the observation that the regions implicated in social cognition behave differently than neural regions not implicated in social thought. In this section, we suggest that social cognition may not only be "separate" from nonsocial cognition but also in some ways "special" as compared to it.

This suggestion derives from the unexpected observation that when not engaged in an active task, metabolic activity in the brain reaches an equilibrium that is marked by the same ratio of oxygen to glucose consumption across the brain. However, although this oxygen to glucose ratio remains constant, brain regions differ from one another in how much *overall* glucose and oxygen they metabolize while at rest.[1] These differences in overall resting metabolic activity suggest that some brain regions are tonically more "active" than others.

Intriguingly, the particular set of brain regions that demonstrate this heightened baseline activity shows remarkable overlap with the set of regions repeatedly implicated in social cognition. Specifically, Gusnard and Raichle (2001) observed elevated baseline activity in a set of regions consisting of ventral and dorsal mPFC and lateral parietal cortex (including STS and TPJ), as well as the precuneus. This observation that regions implicated in social cognition are also those with the highest resting metabolic rates suggests that the baseline state of brain activity may subserve sustained social-cognitive processing or a readiness to engage in it. In particular, high tonic activity in these regions may support a propensity to attribute mental states to entities encountered in our environment (Mitchell, 2006).

The hypothesis that the brain's baseline state may support a tonic readiness for mental state attribution is supported by additional behavioral and neural observations. First, beginning with the classic experiments of Heider and Simmel (1944), psychologists have noted the pervasive human tendency to anthropomorphize—that is, to explain the behavior of nonhuman entities on the basis of human-like mental states. Such a tendency seems readily explainable if tonic activity in regions responsible for social cognition predisposes people to perceive other entities as animate beings.

Second, another of the most consistent findings in cognitive neuroscience has been the observation that these same high-metabolism regions that subserve social-cognitive processing also consistently *deactivate* when perceivers engage in tasks that are nonsocial in nature (for reviews, *see* Shulman et al., 1997; Kawashima et al., 1995). In other words, when perceivers are asked to engage in tasks that do not require consideration of another person's mind, activity in these brain regions appears to be actively suppressed. Such deactivations below baseline are consistently observed even when participants engage in tasks that are structurally identical to social tasks but simply lack a mental state component, including those mentioned in the first section of this chapter, such as remembering the order in which information about a person was presented (Mitchell et al. 2004, 2005, 2006),

[1] These two metabolites are especially important in the context of neuro-imaging, because it is a local change to the ratio of glucose to oxygen consumption that produces the fMRI BOLD signal. Although a detailed discussion of the physiological basis of fMRI is beyond the scope of the present chapter, the observation that the metabolism of glucose and oxygen naturally fall into equilibrium at rest means that this state is not associated with fMRI "activations" over baseline. Instead, these neurophysiological observations demonstrate that resting states represent a default state of human cognitive activity. For a fuller account of these ideas, see Gusnard & Raichle (2001).

forming an impression of an inanimate object (Mitchell, Macrae, & Banaji, 2005), judging the symmetry of a person's face (Mitchell, Banaji, & Macrae, 2005), or solving false photograph problems (Saxe & Kanwisher, 2003).

Activity in the regions responsible for social cognition is thus characterized by behavior that is qualitatively distinct from that of the rest of the brain. Whereas modulations in other (nonsocial) brain regions take the form of significant increases in activity when the region is "in use," followed by a return to baseline upon completion of the relevant task, social regions often show very little increase above their already elevated baseline when "in use" in a social task, along with marked decreases below baseline activity when the brain engages in tasks not involving mental state attribution. Taken together, these observations suggest that humans may navigate the world in a perpetual state of readiness to perceive other beings as social agents, and that to reason appropriately about nonsocial things, this propensity for automatic mental state attribution may need to be temporarily suppressed.

Finally, in addition to shedding light on the processes and propensities supporting the human ability to mentalize, these observations suggest a promising new lens through which to view disorders of social function. For example, a recent study by Kennedy, Redcay, and Courchesne (2006) found that autistic participants failed to demonstrate the typical deactivations in mPFC when completing tasks on which control subjects deactivate the social network, suggesting that they may suffer from low resting activity in brain regions responsible for social thought. The questions raised by these observations of the baseline state and its accompanying deactivations are a unique product of the mixing of social psychology with cognitive neuroscience—one that will surely guide some of the field's most interesting work as social neuroscience continues to come into its own.

Conclusions

Since the advent of fMRI, cognitive neuroscience and social psychology have colluded to generate novel insights into the nature of human social cognition. Like the best of such cross-disciplinary scientific partnerships, the relationship between cognitive neuroscience and social psychology is particularly symbiotic. Social psychology has helped make sense of previously mysterious findings in neuroscience, such as those associated with observations of the baseline state and deactivations. At the same time, neuroscientific observations have begun to illuminate the cognitive basis of various social psychological phenomena, such as the human tendency to overattribute mental states.

Already, social neuroscience has enriched our view of the theoretical landscape of social psychological inquiry in unique and important ways. The observation that social cognition draws on its own particular set of cognitive processes has begun to change the ways in which we think about such myriad topics as autism and psychopathy, the nature of semantic knowledge, and the course of human evolution. Evidence that perceivers draw on distinct cognitive processes as a function of the target's similarity to themselves has widespread implications for understanding the origins of stereotyping, in-group and out-group biases, and selective cognitive impairments in specific aspects of mentalizing. Finally, the observation that the brain's baseline state may reflect a special propensity for social thought has already begun to change our approaches to understanding social disorders and what may be the unique features of the human mind. Through these and other contributions, the emerging field of social cognitive neuroscience joins philosophical, evolutionary, and developmental efforts to discover new truths about the social world, en route to the fullest possible understanding of human behavior.

References

Ames, D.R. (2004a). Inside the mind reader's tool kit: projection and stereotyping in mental state inference. *Journal of Personality & Social Psychology, 87*(3), 340–353.

Ames, D.R. (2004b). Strategies for social inference: a similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of Personality & Social Psychology, 87*(5), 573–585.

Adolphs, R. (2003). Cognitive neuroscience of human social behavior. *Nature Reviews Neuroscience, 1*, 165–178.

Anderson, S.W., Tranel, D., Barrash, J., & Bechara, A. (2006). Impairments of emotion and real-world complex behavior following childhood- or adult-onset damage to ventromedial prefrontal cortex. *Journal of the International Neuropsychological Society, 12*(2), 224–235.

Baron-Cohen, S., Leslie, A.M., & Frith, U. (1986). Mechanical, behavioural and intentional understanding of picture stories in autistic children. *British Journal of Developmental Psychology, 4*, 113–125.

Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders, 29*(5), 407–418.

Bar-On, R., Tranel, D., Denburg, N.L., & Bechara, A. 2003). Exploring the neurological substrate of emotional intelligence. *Brain, 126*, 1790–2000.

Barrash, J., Tranel, D., & Anderson, S.W. (2000). Acquired personality disturbances associated with bilateral damage to the ventromedial prefrontal region. *Developmental Neuropsychology, 18*, 355–381.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A.R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275*(5704), 1293–1295.

Blakemore, S.J., Winston, J., & Frith, U. (2004). Social cognitive neuroscience: where are we heading? *Trends in Cognitive Science, 8*(5), 216–222.

Damasio, A.R. (1994). *Descartes' Error: Emotion Reason, and the Human Brain*. New York, NY: Grosset/Putnam.

Dennett, D. (1987) *The Intentional Stance*. Cambridge, MA: MIT Press.

Fletcher, P.C., Happé, F., Frith, U., Baker, S.C., Dolan, R.J., Frackowiak, R.S., & Frith, C.D. (1995). Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition, 57*(2), 109–128.

Galinsky, A.D., & Moskowitz, G.B. (2000). Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology, 78*(4), 708–724.

Gallagher, H.L., Jack, A.I., Roepstorff, A., & Frith, C.D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage, 16*(3 Pt. 1), 814–821.

Goel, V., Grafman, J., Sadato, N., & Hallett, M. (1995). Modeling other minds. *Neuroreport, 6*(13), 1741–1746.

Gopnik, A., & Meltzoff, A. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Gopnik, A., & Wellman, H.M. (1992). Why the child's theory of mind is really a theory. *Mind and Language, 7*(1), 145–171.

Gopnik, A., & Wellman, H.M. (1994). The theory-theory. In L.A. Hirschfeld & S.A. Gelman (Eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 257–293). New York, NY: Cambridge University Press.

Gordon, R. (1986). Folk psychology as simulation. *Mind and Language, 1*, 158–171.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences, 10*(1), 14–23.

Gusnard, D.A., Akbudak, E., Shulman, G.L., & Raichle, M.E. (2001). Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. *Proceedings of the National Academy of Sciences, USA, 98*, 4259–4264.

Gusnard, D.A., & Raichle, M.E. (2001). Searching for a baseline: functional imaging and the resting human brain. *Nature Reviews Neuroscience, 2*, 685–694.

Happé, F.G. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development, 66*(3), 843–855.

Harlow, J.M. (1868). Recovery from the passage of an iron bar through the head. *Publications of the Massachusetts Medical Society, 2*, 327–347.

Harris, L.T., Todorov, A., & Fiske, S.T. (2005). Attributions on the brain: neuro-imaging dispositional inferences, beyond theory of mind. *Neuroimage, 28*(4), 763–769.

Heal, J. (1986). Replication and functionalism. In Butterfield, J. (Ed.), *Language, Mind and Logic*. Cambridge: Cambridge University Press.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57*, 243–259.

Heistad, D.D., & Kontos, H.A. (1983). Cerebral circulation. In Shepherd, J.T., Abboud, F.M., & Geiger, S.R. (Eds.), *Handbook of Physiology*, section 2, *The Cardiovascular System*, vol. 3, *Peripheral Circulation and Organ Blood Flow*, part 1, (pp. 137–182). Bethesda, MD: American Physiological Society.

Jenkins, A.C., Macrae, C.N., & Mitchell, J.P. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proceedings of the National Academy of Sciences, USA, 105*(11), 4507–4512.

Johnson, S.C., Baxter, L.C., Wilder, L.S., Pipe, J.G., Heiserman, J.E., & Prigatano, G.P. (2002). Neural correlates of self-reflection. *Brain, 125*, 1808–1814.

Kawashima, R., O'Sullivan, B.T., & Roland, P.E. (1995). Positron emission tomography studies of cross-modality inhibition in selective attentional tasks: closing the 'mind's eye'. *Proceedings of the National Academy of Sciences, USA, 92*, 5969–5972.

Kelley, W.M., Macrae, C.N., Wyland, C.L., Calgar, S., Inati, S., & Heatherton, T.F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience, 14*(5), 785–794.

Kennedy, D.P., Redcay, E., & Courchesne, E. (2006). Failing to deactivate: resting functional abnormalities in autism. *Proceedings of the National Academy of Sciences, USA, 103*(21), 8275–8280.

Leslie, A.M., & Thaiss, L. (1992). Domain specificity in conceptual development: neuropsychological evidence from autism. *Cognition, 43*, 225–251.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences, USA, 98*(20), 11,832–11,835.

Mitchell, J.P. (2006). Mentalizing and Marr: an information processing approach to the study of social cognition. *Brain Research, 1079*, 66–75.

Mitchell, J.P. (2005). The false dichotomy between simulation and theory–theory: the argument's error. *Trends in Cognitive Sciences, 9*(8), 363–364.

Mitchell, J.P., Banaji, M.R., & Macrae, C.N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience, 17*(8), 1306–1315.

Mitchell, J.P., Cloutier, J., Banaji, M.R., & Macrae, C.N. (2006). Medial prefrontal dissociations during processing of trait diagnostic and nondiagnostic person information. *Social Cognitive and Affective Neuroscience, 1*(1), 49–55.

Mitchell, J.P., Heatherton, T.F., & Macrae, C.N. (2002). Distinct neural systems subserve person and object knowledge. *Proceedings of the National Academy of Sciences, USA, 99*, 15,238–15,243.

Mitchell, J.P., Macrae, C.N., & Banaji, M.R. (2004). Encoding specific effects of social cognition on the neural correlates of subsequent memory. *Journal of Neuroscience, 24*(21), 4912–4917.

Mitchell, J.P., Macrae, C.N., & Banaji, M.R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50*(4), 655–663.

Mitchell, J.P., Macrae, C.N., & Banaji, M.R. (2005). Forming impressions of people versus inanimate objects: social-cognitive processing in the medial prefrontal cortex. *NeuroImage, 26*, 251–257.

Mundy, P.C., Sigman, M., Ungerer, J., & Sherman, T. (1986). Defining the social deficits of autism: the contribution of nonverbal communication measures. *Journal of Child Psychology and Psychiatry, 27*, 657–669.

Peterson, C., & Siegal, M. (1998). Changing focus on the representational mind: concepts of false photographs, false drawings and false beliefs in deaf, autistic and normal children. *British Journal of Developmental Psychology, 16*, 301–320.

Saxe, R. (2005). Against simulation: the argument from error. *Trends in Cognitive Sciences, 9*(4), 174–179.

Saxe, R., & Kanwisher, N. (2003). People thinking about people: fMRI studies of theory of mind. *Neuroimage, 19*(4), 1835–1842.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: the role of the right temporoparietal junction. *Neuropsychologia, 43*(10), 1391–1399.

Saxe, R., & Powell, L.J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological Science, 17*(8), 692–699.

Schmitz, T.W., Kawahara-Baccus, T.N., & Johnson, S.C. (2004). Metacognitive evaluation, self-relevance, and the right prefrontal cortex. *Neuroimage, 22*, 941–947.

Shamay-Tsoory, S.G., Aharon-Peretz, J., Tomer, R., & Berger, B.D. (2003). Characterization of empathy deficits following prefrontal brain damage: the role of right ventromedial prefrontal cortex. *Journal of Cognitive Neuroscience, 15*(3), 324–337.

Shulman, G., Fiez, J., Corbetta, M., Buckner, R.L., Miezin, F.M., Raichle, M.E., & Petersen, S.E. (1997). Common blood flow changes across visual tasks II: decreases in cerebral cortex. *Journal of Cognitive Neuroscience, 9*, 648–663.

Vogeley, K., May, M., Ritzl, A., Falkai, P., Zilles, K., & Fink, G.R. (2004). Neural correlates of first-person perspective as one constituent of human self-consciousness. *Journal of Cognitive Neuroscience, 16*, 817–827.

Zaitchik, D. (1990). When representations conflict with reality: the preschooler's problem with false beliefs and false photographs. *Cognition, 35*, 41–68.

CHAPTER 2
You, Me, and My Brain: Self and Other Representations in Social Cognitive Neuroscience

Jamil Zaki & Kevin Ochsner

"Say this blanket represents all the matter and energy in the universe, okay? This is me, this is you. And over here, this is the Eiffel Tower, right, it's Paris!"

—Bernard Jaffe, *I Heart Huckabees*

In the film "I Heart Huckabees," the magic of digital effects allows characters to see bits of themselves appear in the pixilated faces of others. Soon after experiencing this literal mirroring of him- or herself in someone else, each character is moved to act compassionately, even toward previous enemies. Vaguely echoing Eastern philosophy, one of the main characters (as quoted above) claims that in moments of clarity, people come to understand that they are actually made from the same blanket, so to speak, and that self/ other distinctions are an illusion.

Does dissolving boundaries between ourselves and others actually help us to navigate the social world? Do we, in fact, understand the mental and emotional states of others using processes that are similar to those we use to think about ourselves? As with most psychological questions this broad, the answer is most likely both yes and no. On the one hand, behavioral research suggests that quite often we use ourselves as a template or anchor when trying to piece together the contents of someone else's mind. This overlap is attested to by our astonishing success at quickly inferring and learning from the goals of others, a type of learning that would only be possible if we understood others as operating much like we do, in pursuit of goals much like our own (Tomasello, 2000). These tendencies produce predictable errors as well. For example, before their 4th birthday, the majority of children despotically assume that others see the world the way they do, and it takes the development of inhibitory control to quell this tendency and enable children to understand that others have thoughts and desires independent from their own (Carlson & Moses, 2001). Similarly, adults will incorrectly guess that things they have just learned (e.g., the brand names of two sodas in a taste test) will be known by others who have not had the benefit of having the answers told to them, and it takes cognitive effort to override this assumed overlap and correctly judge other people's state of knowledge (Epley, Keysar, Van Boven, & Gilovich, 2004).

On the other hand, most adults are easily able to gauge the difference between their own mental states and those of others, and we do so countless times every day. Planning surprise birthday parties and imagining what Christopher Columbus would think about a Corvette are just two examples of situations in which perceivers are able to separate their minds from those of others and use rule-based processing to infer the contents of those others' mental states.

How do we reconcile our tendencies to think of others as being similar to us with the importance and ease of seeing ourselves as different from others? To address this issue, this chapter adopts a social cognitive neuroscience (SCN) approach, using information about the brain to constrain thinking about the psychological processes involved in perceiving people. We review neuroimaging research on self-perception, emotion, and social cognition with an eye toward understanding the person perception processes that lead to our dual tendencies to see others as both like and not like ourselves. Our framework differentiates between two modes of processing information about people—one that is a quick, direct, and bottom-up and another that is deliberative, reflective, and top-down. We then examine whether self and other overlap may depend critically on which mode of processing perceivers are engaging.

Toward this end, the remainder of the chapter is divided into three parts. First, we describe elements of the SCN approach that guide the formulation of our framework. Then, in the second and most detailed section of the paper, we review and synthesize recent imaging research on self and other perception in both direct and reflective modes of processing. This section unpacks each cell of the 2*2 matrix created by crossing the target being perceived (self or other) with the mode of processing used to perceive that target (direct or reflective). For each cell, we draw on a growing neuroimaging literature to help constrain our thinking about the information processing steps that characterize self-perception and other perception. By qualitatively examining commonalities and differences among activation foci from previous studies on self-perception, emotion, and social cognition, we can identify neural systems engaged by each processing mode and for each type of social perceptual target. Finally, in the third section, we use prosocial behavior as an example of how knowledge about neural representations of self and other can help inform our understanding of long-standing social psychological questions.

A SOCIAL COGNITIVE NEUROSCIENCE APPROACH

Social cognitive neuroscience emerged in the past decade as a combination of the theories and methods of its parent disciplines: social psychology and cognitive neuroscience (Ochsner, 2007; Ochsner & Lieberman, 2001). True to its heritage, SCN's goal is to understand the abilities necessary to effectively navigate the social world at multiple levels of analysis, bridging descriptions of social and emotional behaviors and experiences to models of their underlying psychological processes and neural bases.

SCN differs from its parent disciplines in a few important ways. Most obviously, SCN differs from its social psychological parent in its use of neuroscience data to constrain and inform psychological theory (Lieberman, 2007). But there is another important, and perhaps less obvious, way in which SCN is distinguished from the other of its parents. In contrast to much—but not all—of cognitive neuroscience research, SCN emphasizes the core social psychological idea that situations or contexts determine how we think and act (Ochsner, 2007). So central is this idea to social psychology that, as Matthew Lieberman put it, "…if a social psychologist was going to be marooned on a deserted island and could only take one principle of social psychology with him it would undoubtedly be the 'power of the situation'" (Lieberman, 2005). The same might be said of the social cognitive neuroscientist.

Previously, we have argued that the goal of SCN is to construct multilevel models of the way in which one's current context—which includes both the external situation and one's internal states and traits—constrains how we construe the meaning of social cues (Ochsner, 2007). Whereas the cognitive neuroscientist might want to understand the brain systems involved in perceiving faces or facial expressions of emotion, a social cognitive neuroscientist might want to take that understanding further by asking how one's interaction goals (e.g., to form an impression or to connect empathically), beliefs about the other person's intentions (e.g., whether they intend to help or to deceive), or similarity between themselves and a target (e.g., ethnically or politically) shape the cognitive processes and neural systems engaged by perceiving that same emotional expression.

In the sections that follow, the SCN approach will guide a systematic review of recent

Table 2–1 **Person Perception Phenomena Included in Meta-Analysis Grouped as a Function of Mode of Processing and Target of Processing**

		Target	
		Self	Other
Mode of processing	Reflected	Traits, emotion, preferences	Traits, emotion, beliefs, knowledge, familiarity, intentions
	Direct	Pain, arousal, emotion, agency	Traits (stereotypes), intentions, goal–oriented movement, emotion

functional imaging research exploring distinctions in the neural activation corresponding to distinctions between targets (self or other) and modes of processing (direct or reflected). Tables 2–1 and 2–2 indicate the phenomena and studies that were included in each cell of this 2*2 matrix.

Neuroimaging data can help constrain our theories about how these processes interact in two ways: first, by showing that two or more types of behavior that could be viewed as similar—such as explicit and implicit memory formation, actually depend on different information processing mechanisms,; (Schacter, Alpert, Savage, Rauch, & Albert, 1996) and, second, by showing that two types of behavior that were thought to be different, such as visual perception and visual imagery, actually depend on similar mechanisms (Kosslyn & Ochsner, 1994; Kosslyn, Thompson, & Alpert, 1997). Furthermore, aggregating results of several studies allows for examining the reliability of relevant findings, such as the activation of a certain brain region during a certain task type (cf. Phan, Wager, Taylor, & Liberzon, 2002).

With this in mind, the review below describes how different systems of brain regions come into play as a function of situational (i.e., context-specific) goals to understand thoughts, emotions, or traits, goals that in turn lead one to engage in direct or reflective modes when perceiving different kinds of social targets (i.e., one's self or other people).

FROM DATA TO THEORY: BUILDING A SOCIAL COGNITIVE NEUROSCIENCE FRAMEWORK FOR UNDERSTANDING SELF–OTHER REPRESENTATION

If Bernard Jaffe's notion that all people are cut from the same fabric is to be treated as more than a post-hippie platitude, it needs to be grounded in empirical research findings. The goal of this section is to use a review of brain imaging data to bring ideas about self–other similarity down to the brain. To accomplish this goal, we first briefly review past SCN work that has attempted to identify the neural correlates of direct and reflective modes of processing information about the self and about others. This work sets the stage for our review of neural systems implicated in direct and reflective modes of processing for self and other.

Dual-process models in Social Cognitive Neuroscience

Explanations of behavior that appeal to the interplay between direct or automatic processing on the one hand, and reflective or controlled processes on the other are about as old as experimental psychology itself. In social psychology, many such dual-process models have been offered to explain person perception phenomena ranging from stereotyping and dispositional inference to emotion regulation (Gilbert, 1999; K. N. Ochsner & Gross, 2004).

Although the details vary from theory to theory, most models agree upon the basic properties of a direct and automatic mode of processing as opposed to a controlled and reflective one (for several examples of such theories, *see* Chaiken & Trope, 1999). Automatic processes are thought to operate without the costly and cumbersome need to bring mental contents into our awareness for deliberation. Through the simple perception of stimuli that activate mental representations of emotions, stereotyped out-groups, our self-concept, and so on, automatic processes can guide the formation of impressions, can shape judgments and decisions, can generate emotions, and may even queue up goals that motivate and guide actions (e.g. Bargh, Gollwitzer, Lee-Chai, Barndollar, & Trotschel, 2001). By contrast,

Table 2–2 Studies Included in Meta-Analysis as a Function of Mode of Processing and Target of Processing

		Target			
		Self		Other	
		Author/year	Phenomenon	Author/Year	Phenomenon
Mode of processing	Reflected	Fossati 03	Trait Attribution	Brunet 00	Intentions
		Moran 06	Trait Attribution	Calarge 03	Mental States
		Hutcherson 05	Emotion	Castelli 00	Intentions
		Kircher 05	Trait Attribution	Gallagher 00	Beliefs
		Kjaer 02	Trait Attribution	Goel 95	Knowledge
		Lou 04	Trait Recall	Moran 06	Trait Attribution
		Ochsner 04	Emotion	Hynes 06	Mental States, Emotion
		Ochsner 05	Trait Attribution	Lou 04	Trait Attribution
		Ruby 01	Intentions	Mitchell 04	Impression Formation
		Schmitz 04	Trait Attribution	Mitchell 05a	Impression Formation
		Seger 04	Preference	Mitchell 05b	Mental States
		Phan 04	Emotion	Mitchell 05c	Mental States
		Kelley 02	Trait Attribution	Mitchell 06	Impression Formation
				Ochsner 04	Emotion
				Ruby 01	Intentions
				Saxe 03	Beliefs
				Schmitz 04	Trait Attribution
				Seger 04	Preferences
				Vollm 06	Mental States, Emotion
	Direct	Aalto 05	Emotion	Botvinick 05	Pain
		Botvinick 05	Pain	Carr 03	Emotion
		Cato 04	Emotion	Chaminade 02	Intention
		Farrer 02	Intention	Decety 02	Intention
		Hutcherson 05	Emotion	Decety 03	Emotion
		Morrison 04	Pain	Farrow 01	Emotion
		Paradiso 99	Emotion	Jackson 05	Pain
		Schaefer 05	Emotion	Jackson 06	Pain
		Singer 04	Pain	Morrison 04	Pain
		Sugiura 00	Self Recognition	Ramnani 04	Agency
		Taylor 03	Emotion	Saarela 06	Pain
				Singer 04	Pain
				Winston 03	Emotion
				Hooker 03	Intention/movement
				Pelphrey 04	Intention/movement

controlled processes are recruited when, for whatever reason, we need to reflect on or control the impressions, feelings, thoughts, or actions generated by processes operating automatically outside our awareness. Typically reflective control occurs either because we have the explicit goal to be deliberative in a given situation or because of some error or problem produced by the direct mode of processing. Depending on the theory, these two types of processes have been described as working either in competition or in collaboration, either simultaneously or exclusive of one other, and with or without sharing information (Gilbert, 1999).

Recently, dual-process models have begun to inform SCN analyses of person perception

(Keysers & Gazzola, 2007; Lieberman, 2007, in press; Zaki & Ochsner, 2009), emotion (Ochsner & Feldman Barrett, 2001), and emotion regulation (Ochsner & Gross, 2005). In general, these models posit that the direct and bottom-up route for perceiving people or generating emotion depends on brain systems different from, but partially overlapping with, those involved in the reflective mode of processing. Although the neural players implicated in the direct mode may vary from context to context, depending on the specific features of the stimulus at hand (e.g., whether it is painful, visual, auditory, verbal, or pictorial, and so on), for reflective control one player takes center stage for virtually all behaviors. The prefrontal cortex (PFC) is thought to be essential for most aspects of reflective processing, and current work is examining the role of discrete frontal regions in holding information in memory, selective attention, inhibiting prepotent impulses, and higher-order reasoning.

Self and Other Perception in Social Cognitive Neuroscience

As discussed above, questions about whether we see others as we see ourselves have been central to behavioral research for many years. SCN work begun to investigate this issue by asking a related question: whether judgments about one's own states and traits depend on brain systems similar to judging the states and traits of others. This question has been asked in parallel by two different literatures in the field. The first has to do with the neural overlap underlying *conceptual* representations of the self and others and has most often been associated with research on theory of mind. One region in particular—the medial prefrontal cortex (mPFC)—consistently plays a key role in judgments about both self and other, but the nature of mPFC's involvement it is not yet clear. Some studies have found greater activity in ventral portions of mPFC when thinking about one's own as compared to a non-close other's traits (Fossati et al., 2003; Kelley et al., 2002; Macrae, Moran, Heatherton, Banfield, & Kelley, 2004; Northoff et al., 2006). Studies of theory of mind and perspective taking have found activations in more dorsal areas of mPFC occurring while subjects make judgments about the mental states of others (Fletcher et al., 1995; Gallagher et al., 2000; Goel, Grafman, Sadato, & Hallett, 1995; Mitchell, Heatherton, & Macrae, 2002). Other work suggests that the mPFC regions involved in making judgments about one's self and someone else's mental state may overlap (K. N. Ochsner et al., 2005) and that furthermore, this overlap may be moderated by how similar perceivers feel to the people they make judgments about (Mitchell, Banaji, & Macrae, 2005; Mitchell, Macrae, & Banaji, 2006).

The second literature concerns the overlap between the brain areas underlying *motor* representations of self and other and has been centered in research on so-called "mirror neurons" in the premotor cortex of nonhuman primates. These neurons fire both when primates perform an action and when they see another animal performing the same action (Rizzolatti, Fogassi, & Gallese, 2001). This overlap in neural action representations has been reproduced in humans, and a growing number of studies have explored overlapping representations of sensory experiences as well. For example, one fMRI study exposed unlucky participants to aversive odors as well as faces expressing disgust and showed an overlap in activation of the insula for both of these conditions (Wicker et al., 2003). Similar studies have shown overlaps in the perception of pain (Botvinick et al., 2005; Jackson, Meltzoff, & Decety, 2005; Morrison, Lloyd, di Pellegrino, & Roberts, 2004; Singer et al., 2004), touch (Keysers et al., 2004), and basic emotions (Carr, Iacoboni, Dubeau, Mazziotta, & Lenzi, 2003; Leslie, Johnson-Frey, & Grafton, 2004).

"Motor theories" of social cognition and empathy, largely based on the mirror neuron literature, suggest that social cognitive abilities are mediated largely by the fast, automatic, and bottom-up activation of representations of internal states that perceivers see in others. These representations are overlapping, or "shared," to the extent that they are recruited both when one engages in an action and when one sees someone else engaging in the same action. An assumption made by these motor theories is that the bottom-up or stimulus-driven activation of "shared" affective representations creates the feeling in a perceiver that he or she would experience if an event being witnessed was experienced personally. For example, seeing someone else get kicked in the shins may cause a perceiver to wince automatically, actually feeling some measure of discomfort themselves. Motor theories take this and other similar phenomena as a starting point to propose that, in fact, many of our judgments about other people (predicting their actions, intentions, and beliefs) are built on similar overlapping representations (Gallese, Keysers, & Rizzolatti, 2004).

One problem with such accounts is that although they provide explanations of how we understand actions, they fare worse when used to explain our understanding of feelings and beliefs, especially when perceptual inputs are absent or ambiguous. There are many such cases in everyday life, such as when a depressed person has flat affect, when a healthy individual is not emotionally expressive (Zaki, Bolger, & Ochsner, 2008), when someone smiles simply to be polite (Ansfield, 2007), or when someone has a false belief that a perceiver does not share (Jacob & Jeannerod, 2005). Alternative theories propose that in these cases, perceivers use rule-based, top-down processing to dissociate representations of self and other and in this way may be able to infer states in others that differ from their own (Saxe, 2005). In this way, perceiving an ambiguous behavior may have much in common with perceiving any kind of ambiguous visual object: when an incoming percept is not correctly classified using bottom-up processes, the top-down use of an attention and stored knowledge can guide a perceiver to test hypotheses about what she is perceiving or guide her toward goal relevant stimuli (Posner, 1980).

Upshot

On one hand, current work provides some intriguing initial models of how we engage in direct/bottom-up and reflective/top-down modes of perception, but the models have yet to explain how and when the engagement of each mode depends on the target of judgment—self or other. On the other hand, current work has made progress toward clarifying when similar neural representations may underlie perception of and judgments about self and other, but controversies exist as to when and how such "shared representations" or common brain regions are recruited during these processing steps. In the next section, we show how simultaneously accounting for both the mode and the target of judgment may help in resolving these ambiguities.

Toward a Dual-Process Framework for Self–Other Perception

Before discussing the results of our division of previous work, it is worth commenting on the phenomena we chose to include in each analysis, as well as to recap our goals in this review. First, although we included various person perception phenomena from Table 2–1 in our analysis, we have chosen to emphasize the perception of emotions in self and other in our discussion. This is because emotion is the perceptual attribute most clearly present in all four cells of our processing mode * target matrix. For example, as can be seen in Table 2–1, although one can reflect on one's own or someone else's traits, neuroimaging studies of *direct* processing of trait information are rare.

Second, by using a factorial approach, we hoped to isolate patterns of activations from previous studies that would map onto either a main effect of self versus other perception or onto direct versus reflective modes of processing. We then used this framework to probe for interaction effects of perceptual target with processing mode. Specifically, as discussed above, prior work had suggested that neural representations of self and other would overlap, but we expected that the extent of overlap would in some way depend on the mode of processing being engaged. Such interactions could suggest that, in fact, when considering how much people tend to view themselves and social targets as overlapping, it is critical to understand the mode they are using to view those social targets.

Main Effects of Target and Processing Mode

Type of Target: Self versus Other:

We first collapsed activations across all studies of both direct and reflective processing modes and separated them only by the target of perception to test the hypothesis that the processes used to perceive self and other are represented in discrete neural structures. The resulting images clearly show that such a broad distinction cannot be made based on brain data (Fig. 2–1). Studies of both self-perception and other perception have reported activations in regions of the brain associated with processing information about emotions, traits, and intentions. Importantly, across the large majority of studies, both the dorsal and ventral mPFC were activated regardless of whether subjects focused on themselves or someone else.

Furthermore, a host of other areas involved in emotion perception and social cognition, including the superior temporal sulci (STS), anterior insula (AI), amygdala, and posterior cingulate (PCC) were also engaged by both self-perception and other perception. Each of these regions may play important roles in person perception generally. For example, the STS has been implicated in decoding the social meaning of nonverbal cues such as eyes that vary in the direction of gaze, moving lips and forms with biologically possible motion, and tasks involving the assessment of theory of mind or trait attribution (Pelphrey, Morris, Michelich, Allison, & McCarthy, 2005; Saxe, Xiao, Kovacs, Perrett, & Kanwisher, 2004). By contrast, the AI has been implicated in representing internal bodily states, as well as in pain processing. However, it has also been shown to become active while subjects focus

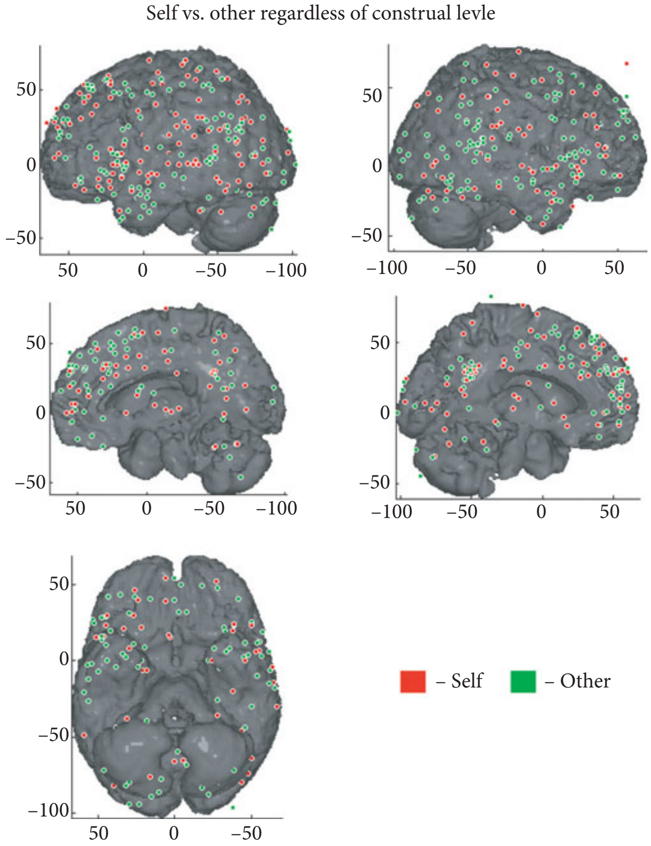


Fig. 2–1 Main effect of target (self vs. other) on neuro-imaging activation peaks.

on the pain and bodily states of other people, suggesting that it is not specific to self-perception (Botvinick et al., 2005; Keysers et al., 2004; Wicker et al., 2003). Similarly, the PCC has been associated with self-directed thought, as well as drawing attention to salient external cues (Vogt, Vogt, & Laureys, 2006). Furthermore, PCC shows high functional connectivity with the mPFC, suggesting that these regions work together during reflection about both one's self or someone else (Lou et al., 2004).

Briefly, two differences between self- and other-related activation peaks are worth noting. First, other-related activations in posterior mPFC tended to be located more dorsally than self-related activations. That is, whereas self-related activation peaks were observed along the cortex adjacent to the corpus callosum, other-related peaks were more often dorsal to the cingulate gyrus. It is known that mPFC evolved in a radial fashion, with the architectonically ancient three layered cingulate gyrus gradually developing into adjacent six-layered portions of mPFC proper. That fact rather intriguingly suggests a developmental relationship between regions involved in perceiving oneself and those involved in perceiving others. That being said, this separation is by no means complete and taken alone does not shed light on the nature of the computations performed by these regions (which are discussed below). Second, more activation peaks in the thalamus and hypothalamus occurred for self than for other. The hypothalamus is critical to regulating autonomic responses to emotionally salient stimulus and also shares connections with brain regions involved in other aspects of emotion processing, such as the subgenual anterior cingulate and orbitofrontal cortex (Morecraft, Geula, & Mesulam, 1992; Nagai, Critchley, Featherstone, Trimble, & Dolan, 2004). Activation of the hypothalamus preferentially during self-related processing may reflect increased effects of autonomic arousal and sensory processing when perceiving or making judgments about internal states than when observing or inferring the presence of such states in others.

Nevertheless, the most striking pattern between self and other was that of overlap. This is not to say that there is a total overlap between the processing steps perceivers use to understand themselves and others. If this was the case, then complex social situations and crowded subway platforms would be difficult to maneuver. Still, these differences do not appear as discrete, consistent separations between targets across all task types.

Mode of Processing: Direct versus Reflective:

When collapsing across targets and instead comparing activation peaks found in studies of direct versus reflective processing, much clearer patterns of separation emerge (Fig. 2–2). This contrast showed a dissociation of activation peaks in the mPFC and ACC, such that reflective processing of traits, emotions, and mental states tended to activate more anterior points within these regions, whereas direct experience of emotion or pain more commonly activated posterior mPFC and ACC, regardless of whether the target was self or other.

This anterior to posterior gradient is consistent with the idea that high-level, reflective, secondary appraisals about one's own or another person's emotions are neurally and cognitively separable from primary appraisals of the potential threat value of stimuli, supporting findings of individual studies. For example, Kalisch et al. (2006) induced anxiety through anticipation of painful shock while subjects performed concurrent working memory tasks involving either low or high cognitive load. Although autonomic arousal and self-reported anxiety were not affected by the amount of cognitive effort the secondary task required, a rostral mPFC region became more engaged for anxiety versus nonanxiety conditions only under low load—that is, when participants could attend to their anxiety. This finding, along with many others that directly manipulate the need for high-level reflective appraisals suggests that rostral MPFC underlies appraisals of internal and emotional states when subjects can attend to and reflect on those states, but not otherwise. This is also consistent with theories about the function of (especially ventral and orbital) PFC that suggest it is a

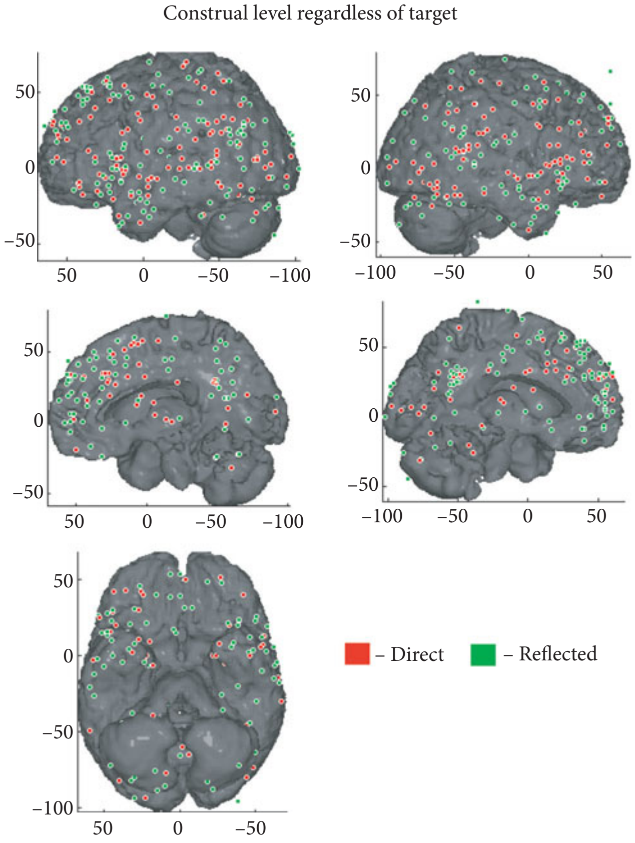


Fig. 2–2 Main effect of mode of processing (direct vs. reflective) on neuro-imaging activation peaks.

"zone of convergence," integrating information about internal bodily states via connections with the hypothalamus and AI with external cues processed in the superior temporal sulci and the amygdala (Floyd, Price, Ferry, Keay, & Bandler, 2001; Mesulam & Mufson, 1982; Rolls, 2004).

By contrast, the ACC may react more automatically and in a bottom-up fashion to the presence of goal-relevant, affectively salient stimuli. In keeping with this notion, a recent study used structural equation modeling to explore effective connectivity between the PFC, ACC, and amygdala while subjects viewed emotional faces and either rated the gender of the face (incidental or direct emotion processing) or the emotion (reflective processing). During direct processing, information from the amygdala traveled to the ACC and then to the PFC, whereas during reflective emotion, this pattern was reversed (de Marco, de Bonis, Vrignaud, Henry-Feugeas, & Peretti, 2006). Because amygdala activation can indicate an "early" cortical mechanism responding to emotional salience, the cortical region it projects to first may indicate the type of appraisal that is made about that stimuli. Thus, the connectivity pattern reported in that paper is consistent with the idea that under a reflective mode of processing, appraisals of emotional value are made in a "top-down" manner through the mPFC before reaching areas (such as the amygdala) more associated with automatic reaction to emotions of others (see also Keightley et al., 2003).

These data, along with the distribution of activation revealed by our plots, suggests that reflecting on emotional states depends on the engagement of medial prefrontal regions supporting high-level appraisal processes used to represent information about the nature of one's own, or someone else's, mental states. This kind of reflection may be important for other types of top-down processing, such as those involved in cognitive forms of emotion regulation that depend on the ability to know what someone is feeling. One such strategy is known as reappraisal: actively rethinking the meaning of an emotionally charged stimulus in ways that change the trajectory of your emotional response to it. Reappraisal may involve awareness of and reflection on the nature of one's own emotional response, as well as reflection on the intentions and beliefs of others. Thus, regions associated with reflective processing of mental states may serve dual duty, helping us perform social cognitive tasks as well as regulate our emotions. In either case, mPFC may communicate with cortical and subcortical regions involved in the direct/bottom-up processing of affective cues, either amplifying or modulating their activity according to the nature of the reflective demands (i.e. amygdala; see Beauregard, Levesque, & Bourgouin, 2001; Etkin, Egner, Peraza, Kandel, & Hirsch, 2006; K. N. Ochsner, Bunge, Gross, & Gabrieli, 2002; K. N. Ochsner et al., 2004). These reflective processes could be employed in cognitive therapy, in which clients are encouraged to reflect on their emotional states and their causes to be able to effectively modulate and dampen their reactions to affective cues (Goldapple et al., 2004; Mayberg, 1997).

Interaction Effects: Degree of Self-Other Overlap Depends on Processing Mode and Content

Our main effect contrasts for perceiving self versus other suggested that separating brain activations by the target of processing alone might resemble trying to slice a cake into the flour and sugar that went into it: although one can contemplate the separation conceptually, in actual practice, the two are hopelessly intertwined. Does this mean that the brain areas used to understand self and other are totally overlapping? Above, we hypothesized that the distinctions between processing different targets might emerge as meaningful depending on the mode of processing a perceiver engages. To test this idea, we plotted activation points for self and other for only one processing mode (i.e., either direct or reflective) at a time, thereby identifying activations associated with either direct or reflective modes of perceiving self or other. In addition, we separated activations associated with different types of judgment and/or stimulus content. In particular, we considered whether activations might segregate for studies involving pain, emotion, or more purely cognitive judgments about nonaffective beliefs. The goal was to determine whether distinct processing systems would subserve the perception of self and other but only when engaged in direct, as opposed to reflective, processing for specific types of stimulus or judgment content.

Direct Processing of Self and Other

Many theories suggesting an overlap between processes involved in self and other perception focus primarily on what we would term direct processing. As described above, these theories have relied mostly on data from studies of mirror neurons and their engagement during the observation of motor actions (Brass & Heyes, 2005; Jarvelainen, Schurmann, & Hari, 2004) as well as mirror-like responses during perception of pain, disgust, and touch in other people. Such studies of self–other neural overlap have influenced suggestions that perceivers understand social targets by automatically activating their own sensory, motor, and affect systems. In the following two subsections, we review studies exploring overlap in the neural systems used to perceive pain and emotion in the self and others.

Pain One of the most compelling cases for overlap in the brain systems involved in self–other perception comes from the results of studies of pain. It is important to our survival that nociceptive (i.e., noxious and painful) signals allow us to pull away from a hot stove; equally important is our ability to learn not to touch a stove someone else has pulled away from in pain. For more than two decades, vicarious conditioning studies have provided a laboratory model of this phenomenon by showing similar skin conductance and heart rate responses when perceivers observe others learning to “fear” conditioned stimuli and when the perceivers themselves are being conditioned (Olsson & Phelps, 2004; Vaughan & Lanzetta, 1980). Imaging studies have focused on a parallel phenomenon, known as "empathic pain," and have observed activity in overlapping regions of ACC and AI both when one experiences pain directly and when one sees someone else experiencing pain (Botvinick et al., 2005; Jackson, Brunet, Meltzoff, & Decety, 2006; Morrison et al., 2004; Singer et al., 2004). The fact that these two regions are associated primarily with affective responses to painful stimuli have been taken to suggest that instead of understanding someone else's pain in a cold and cognitive manner, we feel it as we would our own.

Although the finding of overlapping activity for self and other pain has been highly influential to theories of empathy, important differences for self-pain and other pain have been observed. The process of understanding someone else's pain requires not just an affective response to that pain but a number of additional processing steps as well. For example, one might need to attend to nonverbal, visual cues such as facial expression or body language that can be indicative of another person's response to a painful stimulus. What's more, some understanding of the motivational relevance of a painful situation for someone else may be used to constrain one's understanding of a target's pain experience. Theoretically, these additional types of processing steps should recruit neural systems beyond those commonly supporting the representation of pain affect in self and other, including medial prefrontal regions described earlier that are important for reflecting on the nature of one's mental states and posterior cortical regions (such as the STS) important for interpreting nonverbal cues. By contrast, the direct perception of one's own pain may differentially depend on regions important for the perception of one's own body and the generation of physiological responses important for coping with a noxious stimulus. Regions such as

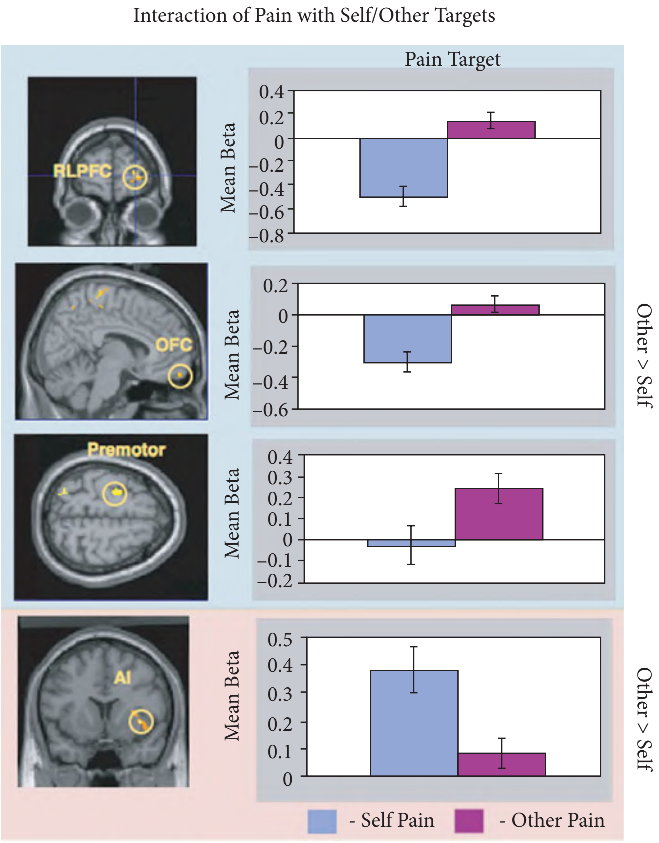


Fig. 2–3 Interaction effects from our recent study of empathy for pain (Ochsner et al., 2008). Orbitofrontal (OFC) and rostrolateral prefrontal cortex (RLPFC), as well as premotor regions, became more active during "other pain" as opposed to self pain. The anterior insula (AI) showed the opposite pattern.

the anterior insula, hypothalamus, and thalamus (described earlier as being important for perception of bodily states and sensations) might be expected to play an role in these processes.

To explore this possibility, Ochsner and colleagues (Ochsner et al., 2008) had participants complete two tasks: in a self-pain task, participants were exposed to both nonpainful and painful thermal stimuli; in an other pain task, participants viewed of others in painful and nonpainful situations. As has been shown in previous work, we identified overlapping regions of AI and ACC more active for painful than for nonpainful stimuli in both tasks. In addition, we found that perception of pain and others preferentially engaged a host of additional regions associated with reflective processing of mental states, including orbitofrontal cortex and rostrolateral PFC. By contrast, posterior sections of the AI were preferentially engaged by self-pain (Fig. 2–3). These findings suggested that as a common affective pain matrix is engaged by both self-pain and other pain, additional functional systems are necessary to fully decode the meaning of painful experiences experienced personally or perceived in others.

We further hypothesized that although self-pain and other pain both involve activation of the AI and ACC, this activation may be part of different cognitive and neural network activity in each case. To test this, we employed functional connectivity analyses. Whereas main effect contrasts that average activity across time and individuals may be insensitive to regions whose activity across two conditions co-vary, functional connectivity analyses are sensitive to such dynamic fluctuations (Friston et al., 1997). In the context of empathy for pain, these analyses showed that during other pain as opposed to self-pain, overlap areas in the ACC and AI become more connected to mPFC regions associated with theory of mind, whereas during self-pain, ACC and AI become more connected to the hypothalamus and periaqueductal gray regions associated with processing autonomic responses (Zaki, Ochsner, Hanelin, Wager, & Mackey, 2007). Based on these findings, we created a schematic representation of brain networks involved in perceiving self-pain and other pain (Fig. 2–4). Such a model can be used as an example of dissociating a seemingly similar process in self and other by probing interaction effects in the brain.

To provide further support for the dissociation of self-processing and other processing in the context of pain, we plotted activations from previous studies of pain perception in self and other (Fig. 2–5). Although the authors of these studies emphasized overlap for self-perception and other perception in the affective pain matrix, Figure 2–5 shows that there are important differences as a function of the target of pain. Whereas self-pain more commonly activates the thalamus and areas along the central sulcus, other pain activated mPFC, bilateral ventrolateral PFC, and OFC, as well as visual association areas. Furthermore, all activation peaks anterior to the genu of the corpus callosum, representing associative regions of PFC, occurred during other pain perception only.

Although these differential activations seldom are discussed in theoretical accounts of

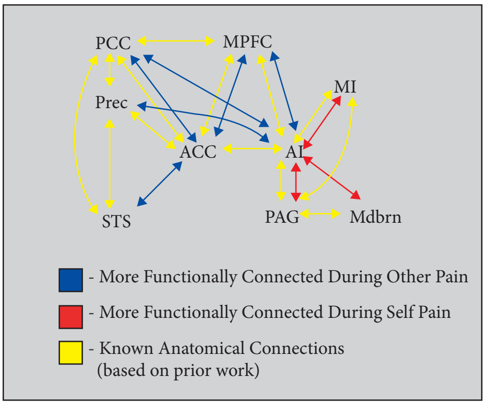

PCC
MPFC
Prec
MI
ACC
AI
STS
PAG
Mdbrn
🟦 - More Functionally Connected During Other Pain
🟥 - More Functionally Connected During Self Pain
🟨 - Known Anatomical Connections (based on prior work)

Fig. 2–4 A circuit model diagramming the interaction of brain areas during self and other pain only, as well as interactions occurring during both types of pain. Connections in the model are based both on connectivity analyses from Zaki et al. (2007), and on existing information about intrinsic physical connections between these regions. mPFC, medial prefrontal cortex; STS, superior temporal sulcus; MI, mid insula; ACC, anterior cingulate cortex; AI, anterior insula; PAG, periaquedictal gray; Prec, precuneus; PCC, posterior cingulate cortex; Mdbrn, Midbrain.

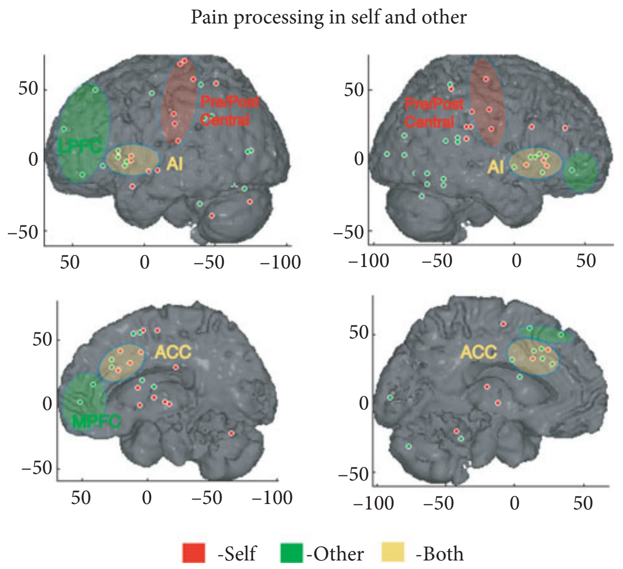


Fig. 2–5 Neuro-imaging activation plots demonstrating the effect of target (self vs. other) on pain perception.

empathic pain, they are important in at least two ways. First, they suggest that although neural overlap between self-pain and other pain processing may exist in the ACC and AI, the functional role of activity in these regions may differ in each context, depending on the additional regions with which the ACC and AI are interconnected. Second, they provide means for explaining paradoxical effects of viewing pain in certain contexts. For example, during competition, one's own goal and those of someone else directly conflict. In these cases, it may be adaptive for perceivers to "turn off" otherwise automatic reactions to the pain of others (e.g., during athletic competitions or, more extremely, during war). In keeping with this notion, both autonomic and neural activity evoked by watching others in pain is reduced or reversed when the people in pain are in an adversarial or competitive relationship with a perceiver (Lanzetta & Englis, 1989; Singer et al., 2006). Under the hypothesis that processing of pain in self and other largely overlap, these effects would be difficult to explain. However, the recruitment of prefrontal regions important for perceiving the intentions of others could modulate the amount of AI and ACC activity perceivers engage while observing another person in pain, depending on how a perceiver feels about or relates to that target.

Emotion Emotional stimuli do not necessarily require reflective awareness of them to affect the way we feel, act, or engage in cognitive processing. This fact was taken advantage of by the producers of *The Exorcist*, who included grotesque subliminal images in their film, causing moviegoers to become terrified and nauseated while watching the film although they couldn't quite pinpoint why. Before being discovered, these producers managed to show, in thousands of unwary subjects, the extent to which emotional cues we do not experience consciously can affect our mood. Importantly, emotion without reflection can affect other aspects of our cognitive and even perceptual functioning, such as how much money we will spend while shopping or the part of a photograph to which we attend (Gasper & Clore, 2002).

Perception of emotional cues without reflection also has discrete neural correlates. Masked emotional stimuli can cause amygdala activation outside of awareness (Whalen et al., 2004; Whalen et al., 1998), although this finding has been contested (Pessoa, Japee, & Ungerleider, 2005; Pessoa, McKenna, Gutierrez,

& Ungerleider, 2002). Interestingly, the amygdala is preferentially engaged by faces displaying fear, even over other potentially threat-related emotions such as anger (Whalen et al., 2001). Given that the amygdala is connected to sensory systems via only a few synapses, this suggests that some of the fastest processing we use to assess potential threat may rely on cues about the emotional experiences of others who may be responding to something we should be avoiding. This possibility raises what by now should be an obvious question: Does the neural activity accompanying perception of someone else's fear resemble the neural activity we exhibit in response to our own fear? Or, to extend William James' already overextended phrase, does a perceiver become frightened by *someone else* running from a bear? If so, does that perceiver's fear originate in an understanding of the frightened sprinter, or does the perceiver simply become primed for fear and vigilance outside of his awareness?

A few studies have argued that the latter may be true. This work extends the logic of studies examining so-called "shared representations" to the domain of perceiving facial expressions of emotion. By and large, findings have supported the theory that when we see someone else's emotional face, we "feel" the same thing they do, by virtue of activating brain regions similar to those activated when we experience the emotion we see them expressing. For example, both seeing and imitating emotional facial expressions activates the amygdala and AI (along with classic mirror neuron regions in the inferior frontal and premotor cortices), suggesting overlap between perception and sensation of emotions (Carr et al., 2003; but see also Leslie et al., 2004).

Although these data suggest that direct processing of self-emotion and other emotion cues may recruit at least partially overlapping neural circuitry, this is certainly not the entire story. Although the amygdala is associated with generating physiological components of emotional responses, an early meta-analysis of emotions found that more frontal regions, including the mPFC and ACC, are actually the most commonly recruited by emotional stimuli (Phan et al., 2002), and more recent meta-analyses suggest that these regions are associated with emotional experience, whereas the amygdala is not (Barrett, Mesquita, Ochsner, & Gross, 2007). Furthermore, an observational learning paradigm found that while watching someone else receive shock activated the amygdala, only subjects' own fear of being shocked engaged ACC (Olsson, Nearing, & Phelps, 2007). This suggests that the perception of emotions experienced by another person may commonly trigger a "warning bell" to the self that danger is present but does not engage prefrontal systems associated with higher-level, reflective processing of mental states and intentions.

To parse the regions associated with processing of self-emotion and other emotion cues under direct and reflective modes of processing, we selected activation peaks from a group of emotion-related neuroimaging studies. In doing so, we defined a "direct" mode of emotion processing as any emotional response that a subject experiences or sees someone else experience but does not attend to or judge explicitly. Contrasts were included in the "direct self" category if they asked participants to passively look at aversive or amusing scenes or videos or required participants to make a nonemotional judgment about those stimuli (e.g., "was this photograph taken indoors or outdoors?"). Contrasts were included in "direct other" category if they asked participants to passively attend to or make nonemotional judgments about emotionally expressive faces or body movements.

Resulting plots are shown in Figure 2–6. The greatest degree of overlap between direct processing of self-emotion and other emotion cues occurred in anterior and posterior sections of mPFC, dorsal to the genu of the corpus callosum. These regions have been shown to respond to emotional stimuli in general (Phan et al., 2002) but, as reviewed above, also respond during tasks requiring reflective processing of mental states including, theory of mind tasks and action monitoring (Amodio & Frith, 2006). Self and other stimuli also produced heavily overlapping patterns of activity

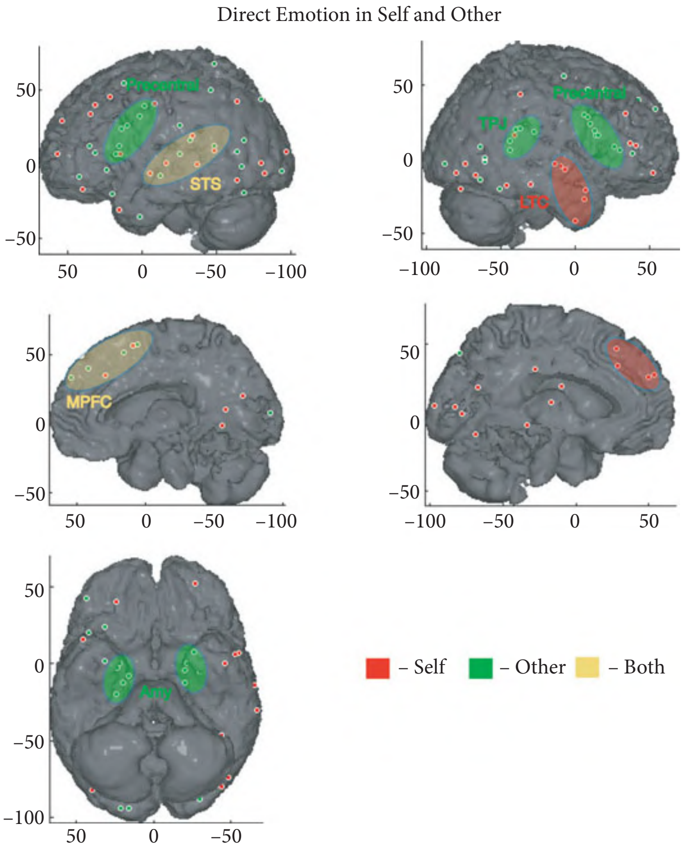


Fig. 2–6 **Neuro-imaging activation plots demonstrating the effect of target (self vs. other) on the direct processing of emotion.**

in left STS regions associated with the perception of nonverbal social cues (Pelphrey, Morris, & McCarthy, 2004).

The fact that these regions are engaged both by reflective processing of social targets in general, and by the direct processing of affective cues regardless of target, highlights the important role that understanding the intentions of others plays in appraising the affective significance of stimuli. Indeed, many appraisal theories of emotion postulate that specific computations about the intentions of others determine whether or not we feel angry or sad, happy, or surprised in response to the actions of other people (Scherer, Schorr, & Johnstone, 2001).

Perhaps as important as these regions of overlap, self-processing and other processing of emotion also showed disparate patterns of activations in several brain regions. Although “direct other” emotional stimuli more commonly activated bilateral premotor cortex, amygdala and right temporoparietal junction (TPJ), “direct self” emotions showed unique activation peaks along the right temporal pole, medial occipital lobe, and thalamus. The premotor and TPJ activations in the “other” condition are consistent with previous accounts of “motor empathy” in which covert imitation plays some role in processing emotional cues from others (Iacoboni, 2005; Iacoboni et al., 1999). The TPJ is often associated with making inferences about the mental states of others (Saxe & Kanwisher, 2003; Saxe & Wexler, 2005) as well as the disengagement of spatial attention more generally (Corbetta, Kincade, Ollinger, McAvoy, & Shulman, 2000); as such, its presence preferentially in “other direct” emotion may suggest attempts to orient to alternative interpretations of other people’s affective responses. The activation of amygdala during other emotion, and of the thalamus in self emotion, respectively, suggests that qualitatively different processes underlie each type of emotion. In keeping with our discussion of the perception of pain, perceiving emotions in others may depend on systems sensitive to detecting potentially goal relevant features of the environment, whereas experiencing our own emotions may involve greater monitoring of internal bodily states.

Summary In reviewing studies of pain and emotion, we found that under a direct mode of processing, the brain regions engaged by perceiving self and other partially overlap, corresponding with the emphasis of many studies on so-called "shared representations" in empathy and social cognition. These overlaps occur mainly in cortical regions (i.e., AI and ACC) used for integrating emotional cues or sensations into coherent second order (i.e., nonsensory) representations of affective states. However, we also found striking dissociations between self- and other-related activation peaks. Specifically, watching others feeling pain or expressing emotion engaged motor cortex, which may help us understand intentions underlying others' actions, as well as the amygdala, which may trigger vigilance in response to the perception of others feeling threatened. On the other hand, the experience of self-pain and emotion consistently involved postcentral gyrus, thalamus, and hypothalamus—areas associated with processing information about bodily states and sensations. Furthermore, connectivity analyses of perceiving pain in the self and in others revealed that only other pain causes ACC and AI to become functionally connected with the mPFC, an area associated with mental state inference (Zaki et al., 2007). Together, these findings indicate that although perceivers may experience responses to their own pain and emotion that are similar to those experienced when they perceive pain and emotion in others, the functional networks through which these sensations are created may be importantly different.

Reflective Processing of Self and Other

The patterns of dissociation between self and other we observed when participants are in a direct mode of processing are the product of differential recruitment of systems important for processing be sensory information available for direct personal experience as compared to the indirect observation of others. To the extent that reflective processing integrates lower-level sensory and perceptual cues into higher-order representations, we would expect similar systems to support the reflective processing of multiple types of cues, including those associated with the perception of emotion in oneself and other people.

Emotion To explore this hypothesis, we plotted activation peaks from several studies of reflective emotion processing in Figure 2–7. To date, there are few studies of the reflective processing of pain. As described above, we constrained our plots to show the results of main effect contrasts requiring explicit judgment of affective states. The "reflective self" category included any contrast in which participants were asked to rate their own experience while viewing emotional stimuli, whereas the "reflective other" category included contrasts where participants rated the emotional state of someone else in a picture, vignette, or cartoon. We included both contrasts comparing judgment to no judgment and contrasts comparing affective judgments to judgments about external stimulus features (e.g., emotional state vs. gender of someone in a picture). Because we were also interested in the relationship between qualitatively different types of reflections about others, we plotted "reflective other" studies in which subjects made nonemotional mental state judgments about others in vignettes, pictures, and cartoons separately from those where participants made judgments about the enduring personality traits of targets (which, in all cases, involved both emotional and nonemotional judgments, Fig. 2–8).

Several distinctions emerged in these plots. First, reflective emotion processing showed several regions of overlap for both self-targets and other targets. These overlaps included activations in precuneus and posterior cingulate cortex (PCC), the mPFC, bilateral temporal poles, and medial OFC. These findings are important because virtually all of these regions have been

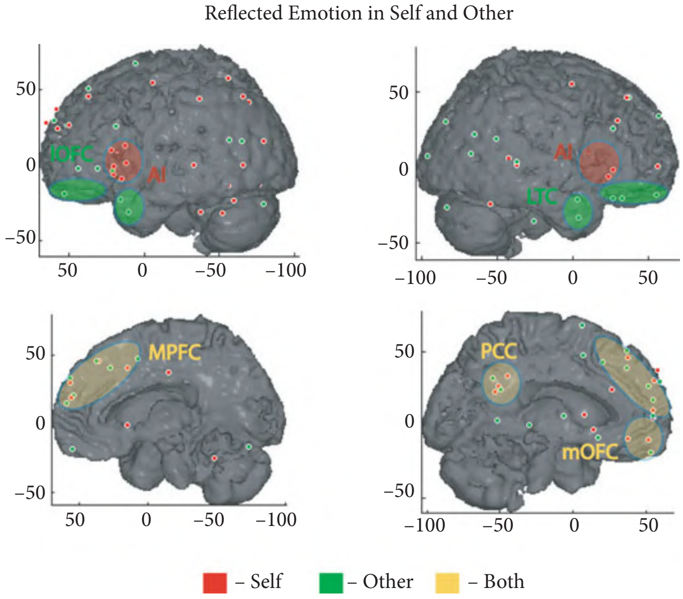


Fig. 2–7 **Neuro-imaging activation plots demonstrating the effect of target (self vs. other) on reflective processing of emotion.**

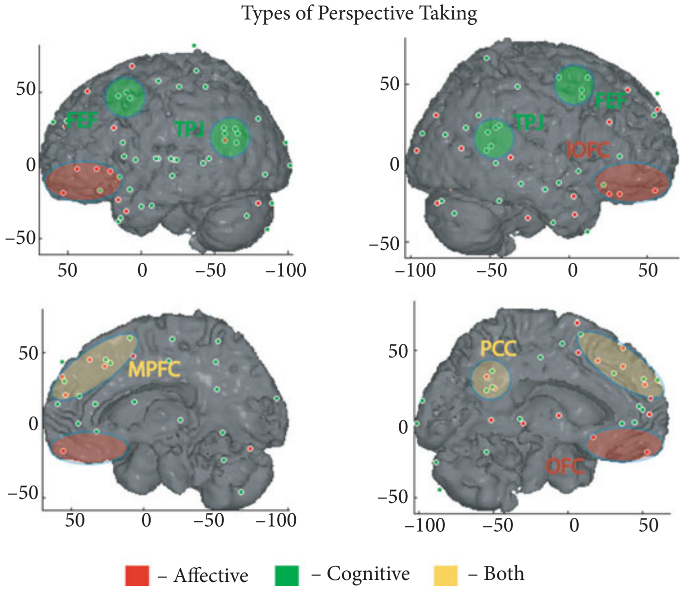


Fig. 2–8 **Neuro-imaging activation plots demonstrating the effect of perspective taking type (cognitive vs. affective) on neural activity.**

previously described as important for mental state attribution in general (Frith & Frith, 2003). The present analysis highlights once again the importance for emotion of regions previously associated with social cognition and mental state attribution in general. Activity in numerous subregions of mPFC, including anterior and ventral portions of this region, was not surprising, given that mPFC is central to both inferences about internal states (Amodio & Frith, 2006; Mitchell, Neil Macrae, & Banaji, 2005) and emotional experience, as described earlier.

Activity in two additional overlap regions—the precuneus and PCC—is worthy of additional discussion, as they have not been discussed previously. Activity in the precuneus is often related to both visuospatial imagery and self-focused attention (Cavanna & Trimble, 2006; Gusnard, Akbudak, Shulman, & Raichle, 2001; Kelley et al., 2002) visual perspective taking in a first person (Vogeley & Fink, 2003) or third person (Ruby & Decety, 2001) point of view. Importantly, the precuneus does not have connections with any primary sensory cortices but does have efferent connections to the STS and ACC and may be involved in directing attentional resources to salient social or emotional stimuli (Lou et al., 2004). Similarly, the PCC is often recruited in self-referential mental and emotional tasks, and Vogt et al. (2006) have suggested that ventral PCC may play a part in a ventral attentional stream, sending information about potentially salient stimuli to the vACC through direct reciprocal connections. Together, common activation in these regions suggests that perceivers use similar mechanisms for self-perception and other perception to direct attentional resources to emotional cues.

Dissociations between activity associated with reflective judgments of self and other were subtler than the analogous differences described in the context of direct emotion processing. These differences may be less reliable and are deserving of attention and future research designed to unpack their functional significance. For present purposes, we merely note reflective judgments of other people's emotions more commonly recruited extrastriate and medial occipital cortices, which is consistent with the fact that these tasks involved explicit attention to people, mostly in visual scenes. In addition, whereas self-related judgments more commonly recruited inferior frontal regions, other-related judgments more often recruited lateral orbitofrontal regions. Given that both of these regions are associated with response selection and response in addition, and that their precise computational roles remain a hot topic of debate, it is not yet clear what this result might mean.

Overall, however, the most striking feature of these plots is the commonality of activity regardless of the target of perception. Importantly, this differs from the pattern observed for direct processing of emotion, which showed recruitment of both common and distinct regions for self and other. Together, these patterns suggest that when the self or someone else is viewed as an object of reflection, a network of regions comes into play that is involved in directing attention, interpreting social cues, and inferring internal states. By contrast, in the absence of reflective processing, the direct and bottom-up perception of emotion from low-level cues recruits different systems depending on the type of perceptual input associated with each target (visceral for self vs. visual for other).

*Distinct neural substrates for different types of reflective judgment*The reflective mode of processing offers myriad possible ways of attending to, and elaborating on, our judgments about ourselves and other people. We might, for example, think about how someone feels as compared to what they are thinking, and such differences in focus might involve different underlying neural circuitry. To determine whether the *way* in which we reflect on our own or others' mental states depends on different underlying neural systems, we examined separately activations related to emotional as compared to nonemotional mental state judgments (i.e., false belief tasks). This analysis revealed a dissociation in brain regions recruited by cognitive as opposed to affective inferences about other people (Fig. 2–8). Whereas cognitive judgments more commonly recruited bilateral TPJ and frontal eye fields (FEFs), affective judgments more commonly recruited orbital frontal and anterior vmPFC regions.

TPJ is associated with mental state judgments (Saxe & Kanwisher, 2003; Saxe & Wexler, 2005) and also with shifting attention towards behaviorally relevant stimuli in—for example, external cueing tasks (Kincade, Abrams, Astafiev, Shulman, & Corbetta, 2005). FEF is engaged during tasks requiring increased attention to and working memory for visuospatial stimuli, including when one attempts to inhibit reflexive tendencies to shift one's eyes toward a visual stimuli (Curtis & D'Esposito, 2003). Activations in these regions when drawing inferences about cognitive, but not affective, states could suggest that cognitive inferences depend to a greater extent on the mental manipulation of information about stimuli in the external world. This could especially be the case given that often (as in a false belief task), cognitive inferences require participants to keep two disparate mental states (their own and their target's) in mind, as well as overriding the prepotent desire to impose their own mental states and knowledge on a target. Theory of mind critically relies on executive function—and especially on inhibitory control—and the two develop in parallel (Carlson & Moses, 2001). When our own perspectives and someone else's differ (i.e., we have knowledge that a target does not), making accurate judgments about their state requires us to adjust from our own state, a process that is attentionally demanding. Activation of FEF and TPJ during mental state inference may reflect the unique attentional demands of keeping multiple mental states in mind simultaneously.

Engagement of OFC and related ventral mPFC regions when drawing affective inferences could be related to the role these regions play in representing the motivational value of stimuli. Single-unit recording, lesion, and functional imaging studies of conditioning and reinforcement learning have long implicated OFC and ventromedial PFC in representing the current motivational or affective value of stimuli as it changes over time as a function of one's current goals (Barrett et al., 2007; Rolls, 2004). OFC also shares strong connections with the hypothalamus, which projects to brain-stem nuclei controlling autonomic outflow, and its activity has been shown to co-vary with skin conductance responses (cf. Nagai, Critchley, Featherstone, Fenwick et al., 2004). By contrast, the amygdala has been thought to encode relatively enduring, context-free and stimulus-driven associations between perceptual cues and physiological responses (Schoenbaum, Chiba, & Gallagher, 1999). The OFC could therefore play an important role in representing either one's own or another person's current affective state.

This hypothesis could explain the role of OFC in the perception of emotion in self and other. Consider, for example, the results of a recent study in which participants saw emotional or neutral pictures and then rated their affect for the subsequent 20 seconds after the pictures disappeared. After viewing negative pictures, subjects commonly reported feeling sustained emotion after the picture itself was gone. Although time-courses of amygdala activity tracked with the presence of negative pictures, lateral OFC activity tracked participants' sustained self-reported emotional response (Garrett & Maddock, 2006). In this study, OFC reflects the personal experience and generation of an emotional response to a stimulus. Interestingly, antisocial and psychopathic patients, as well as patients with orbitofrontal and vmPFC damage, show blunted autonomic reactions to expected stressors (Bechara, Tranel, Damasio, & Damasio, 1996; Raine, Lencz, Bihrle, LaCasse, & Colletti, 2000), as well as in anticipation of unpredictable stressors (Roberts et al., 2004). This suggests that they may be unable to generate context-appropriate affective responses.

Now consider the results of other studies suggesting that affective representations in OFC may help us understand the emotions generated in other people. OFC patients don't understand social faux pas (Stone, Cosmides, Tooby, Kroll, & Knight, 2002) and also fail to experience normal levels of self-conscious emotion in social interactions that would engender either pride or embarrassment in healthy individuals (Beer, Heerey, Keltner, Scabini, & Knight, 2003). Self-conscious emotions like these are important in social interactions because they tell us when our own behavior has had intended (pride) or unintended (embarrassment) consequences for others. To the extent that damage to OFC renders us unable to experience these emotions normally, we may make become inappropriately boastful, forward, or rude.

Summary Comparisons of patterns of neural activity associated with a reflective mode of processing for self and other showed much more overlap and fewer differences than did the same comparison for the direct mode of processing. This suggests that when making explicit judgments about people, perceivers tap into a common set of cognitive and affective processes regardless of whether they are reflecting about themselves or someone else. Perceivers direct their attention to salient cues, infer internal states, and also create corresponding autonomic and emotional states in themselves when trying to infer emotions in others and when inferring false beliefs may use inhibitory control to separate their point of view from their target's.

Conclusions and future directions

Now that we have taken this whirlwind tour of the data on direct and reflective modes of processing for self and other targets, we can take a moment to recap where we've been and then revisit some questions we began with to see if we're any closer to answering them than when we started.

The premise of this chapter was that we could gain insight into the processes mediating perception of one's own feelings and thoughts, or those of other people, by using data from functional neuro-imaging studies. We felt that that common and distinct patterns of activity associated with the mode of processing—reflective or direct—and the target of perception—self or other—could be used to address this question. Our method was to perform a qualitative meta-analysis of studies examining the perception of one's own or other people's affective states. Our results suggested two conclusions. First, when perceivers reflect on the emotions of others, they do so using mechanisms similar to those they use to process their own emotions. Second, in the absence of reflective attention, overlapping but distinct processes are used to represent your own or other people's affective states.

Do these data help us understand whether representational overlap between of our own emotions and those of others allow smooth navigation of the social world, and whether it could stimulate prosocial behavior, as suggested in *I Heart Huckabees*? This question is important not just because it relates to the fanciful premise of a moderately successful existential film but because the ability of neuroscience data to address it may provide a litmus test for our current SCN models of social behavior.

Not coincidentally, this question is also the subject of a longstanding debate in social psychology. Daniel Batson and colleagues have argued that we help others because of a selfless *empathic concern* we feel for them. For example, in a series of studies, Batson asked participants to decide whether they would like to perform a fun task with the potential of earning money or a boring task for which they would not get paid. Whichever task they did not choose would be given to another person whom the participant would not meet. An experimenter gave each participant a coin to flip in case they wanted to make a "fair" choice. Before deciding, subjects were either *(1)* not given instructions, *(2)* told to imagine themselves in the other person's situation, or *(3)* told to think of the other person's feelings while they made their decision. Thinking of oneself in someone else's situation caused participants to flip the coin more but not to assign the other person to the more desirable task, whereas thinking of the other person's emotions at the time caused most participants to take on the more boring task for themselves (Batson et al., 2003). These results and others support Batson's view that perspective taking and emotional empathy are at the root of prosocial behavior towards others (Batson et al., 1991; Batson et al., 1988), including social out-groups toward whom we might otherwise find threatening (Batson et al., 1997; see also Eisenberg & Miller, 1987).

Other researchers have disagreed, however, with Batson's idea that prosocial behavior is impersonal or selfless in nature. Several studies have claimed that the effect of empathy on prosocial behavior is moderated (or replaced) by a sense of similarity—or overlap—between self and other. That is, we help people only because we feel connected to them in some way, and their suffering causes us suffering as well (Cialdini, Brown, Lewis, Luce, & Neuberg, 1997; Cialdini et al., 1987). From this viewpoint, empathy may create a feeling of similarity between a participant and the person whose perspective they are taking (Davis, Conklin, Smith, & Luce, 1996). In the end, Cialdini and colleagues argue that it is only because of a desire to reduce our own suffering that we choose to help others. For example, one study related volunteer AIDS workers' motivation to their resulting helping behavior and found that empathic concern moderated helping only if the patient that volunteers worked with was a member of their ingroup. This effect was replicated in a laboratory paradigm testing spontaneous helping behavior for a confederate who participants believed had hepatitis (Sturmer, Snyder, & Omoto, 2005).

If helping behavior is driven by an observer's own distress while seeing someone else's suffering, then it makes sense that we should preferentially help those closest to us. Although it is painful to read news stories about natural disasters happening in foreign countries, this pain may be fundamentally different from what we feel when a friend or family member is injured. In line with this suggestion, individuals are more likely to ascribe secondary emotions (i.e., shame, pride) to in-group members, suggesting that we attend more to their emotional states, allowing us to feel a greater sense of overlap with them and to feel more distress at their distress (Leyens et al., 2000). Group membership in this context can be defined by a situation, rather than by traits such as race or gender. This could explain the lack of empathy subjects had for competitors in Lanzetta (1989) and Singer's (2006) work.

Does our review of the neuro-imaging literature on self-perception and other perception suggest that helping behavior is mediated by emotional perspective taking (as claimed by Batson) or that it instead depends on an overlap between self and other (as claimed by Cialdini and others)? Our review indicates that although self-perceptions and other perceptions differ importantly when one is processing information in a direct, unreflective manner, paying attention to someone else's emotional state increases the similarity of regions used to perceive one's own emotions and those of another person. In other words, to the extent that an observer attends to and reflects on the emotional states of a target, a richer, more reflectively elaborated representation of that target's state begins to emerge for the observer. Behavioral data converges with imaging data by suggesting that this reflective representation more closely approximates how the observer views herself: perspective taking causes observers to rate targets as more similar to themselves (Davis et al., 1996) and to engage more overlapping neural activity when judging themselves and targets (Ames, Jenkins, Banaji, & Mitchell, 2008).

Applying our models of the brain bases of self and other perception to real-world dilemmas such as the motivations for prosocial behavior remains a speculative pursuit but one which we feel can nonetheless be fruitfully expanded through further use of brain imaging data. Hopefully, this chapter has served to illustrate how such data can be used begin building theories of person perception that link psychological processes to their neural bases. It remains for future work to take the next step and link this work directly to behavior in prosocial contexts to determine whether the presence of "shared representations" truly mediates one's desire to help, or at least makes one feel like part of an existential blanket.

References

Ames, D. L., Jenkins, A. C., Banaji, M. R., & Mitchell, J. P. (2008). Taking another person's perspective increases self-referential neural processing. *Psychol Sci, 19*(7), 642–644.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nat Rev Neurosci, 7*(4), 268–277.

Ansfield, M. (2007). Smiling when distressed: when a smile is a frown turned upside down. *Pers Soc Psychol Bull, 33*(6), 763–775.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trotschel, R. (2001). The automated will: nonconscious activation and pursuit of behavioral goals. *J Pers Soc Psychol, 81*(6), 1014–1027.

Barrett, L. F., Mesquita, B., Ochsner, K. N., & Gross, J. J. (2007). The experience of emotion. *Annu Rev Psychol, 58*, 373–403.

Batson, C. D., Batson, J. G., Slingsby, J. K., Harrell, K. L., Peekna, H. M., & Todd, R. M. (1991). Empathic joy and the empathy-altruism hypothesis. *J Pers Soc Psychol, 61*(3), 413–426.

Batson, C. D., Dyck, J. L., Brandt, J. R., Batson, J. G., Powell, A. L., McMaster, M. R., et al. (1988). Five studies testing two new egoistic alternatives to the empathy-altruism hypothesis. *J Pers Soc Psychol, 55*(1), 52–77.

Batson, C. D., Lishner, D. A., Carpenter, A., Dulin, L., Harjusola-Webb, S., Stocks, E. L., et al. (2003). "…As you would have them do unto you": does imagining yourself in the other's place stimulate moral action? *Pers Soc Psychol Bull, 29*(9), 1190–1201.

Batson, C. D., Polycarpou, M. P., Harmon-Jones, E., Imhoff, H. J., Mitchener, E. C., Bednar, L. L., et al. (1997). Empathy and attitudes: can feeling for a member of a stigmatized group improve feelings toward the group? *J Pers Soc Psychol, 72*(1), 105–118.

Beauregard, M., Levesque, J., & Bourgouin, P. (2001). Neural correlates of conscious self-regulation of emotion. *J Neurosci, 21*(18), RC165.

Bechara, A., Tranel, D., Damasio, H., & Damasio, A. R. (1996). Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex. *Cereb Cortex, 6*(2), 215–225.

Beer, J. S., Heerey, E. A., Keltner, D., Scabini, D., & Knight, R. T. (2003). The regulatory function of self-conscious emotion: insights from patients with orbitofrontal damage. *J Pers Soc Psychol, 85*(4), 594–604.

Botvinick, M., Jha, A. P., Bylsma, L. M., Fabian, S. A., Solomon, P. E., & Prkachin, K. M. (2005). Viewing facial expressions of pain engages cortical areas involved in the direct experience of pain. *Neuroimage, 25*(1), 312–319.

Brass, M., & Heyes, C. (2005). Imitation: is cognitive neuroscience solving the correspondence problem? *Trends Cogn Sci, 9*(10), 489–495.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Dev, 72*(4), 1032–1053.

Carr, L., Iacoboni, M., Dubeau, M. C., Mazziotta, J. C., & Lenzi, G. L. (2003). Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proc Natl Acad Sci U S A, 100*(9), 5497–5502.

Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain, 129*(Pt 3), 564–583.

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual Process Theories in Social Psychology*. New York: Guilford Press.

Cialdini, R. B., Brown, S. L., Lewis, B. P., Luce, C., & Neuberg, S. L. (1997). Reinterpreting the empathy-altruism relationship: when one into one equals oneness. *J Pers Soc Psychol, 73*(3), 481–494.

Cialdini, R. B., Schaller, M., Houlihan, D., Arps, K., Fultz, J., & Beaman, A. L. (1987). Empathy-based helping: is it selflessly or selfishly motivated? *J Pers Soc Psychol, 52*(4), 749–758.

Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P., & Shulman, G. L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nat Neurosci, 3*(3), 292–297.

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends Cogn Sci, 7*(9), 415–423.

Davis, M. H., Conklin, L., Smith, A., & Luce, C. (1996). Effect of perspective taking on the cognitive representation of persons: a merging of self and other. *J Pers Soc Psychol, 70*(4), 713–726.

de Marco, G., de Bonis, M., Vrignaud, P., Henry-Feugeas, M. C., & Peretti, I. (2006). Changes in effective connectivity during incidental and intentional perception of fearful faces. *Neuroimage, 30*(3), 1030–1037.

Eisenberg, N., & Miller, P. A. (1987). The relation of empathy to prosocial and related behaviors. *Psychol Bull, 101*(1), 91–119.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *J Pers Soc Psychol, 87*(3), 327–339.

Etkin, A., Egner, T., Peraza, D. M., Kandel, E. R., & Hirsch, J. (2006). Resolving emotional conflict: a role for the rostral anterior cingulate cortex in modulating activity in the amygdala. *Neuron, 51*(6), 871–882.

Fletcher, P. C., Happe, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., et al. (1995). Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition, 57*(2), 109–128.

Floyd, N. S., Price, J. L., Ferry, A. T., Keay, K. A., & Bandler, R. (2001). Orbitomedial prefrontal cortical projections to hypothalamus in the rat. *J Comp Neurol, 432*(3), 307–328.

Fossati, P., Hevenor, S. J., Graham, S. J., Grady, C., Keightley, M. L., Craik, F., et al. (2003). In search of the emotional self: an FMRI study using positive and negative emotional words. *Am J Psychiatry, 160*(11), 1938–1945.

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., & Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage, 6*(3), 218–229.

Gallagher, H. L., Happe, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia, 38*(1), 11–21.

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends Cogn Sci, 8*(9), 396–403.

Garrett, A. S., & Maddock, R. J. (2006). Separating subjective emotion from the perception of emotion-inducing stimuli: an fMRI study. *Neuroimage, 33*(1), 263–274.

Gasper, K., & Clore, G. L. (2002). Attending to the big picture: mood and global versus local processing of visual information. *Psychol Sci, 13*(1), 34–40.

Gilbert, D. (1999). What the mind's not. In S. Chaiken & Y. Trope (Eds.), *Dual Process Theories in Social Psychology* (pp. 1–12). New York: Guilford Press.

Goel, V., Grafman, J., Sadato, N., & Hallett, M. (1995). Modeling other minds. *Neuroreport, 6*(13), 1741–1746.

Goldapple, K., Segal, Z., Garson, C., Lau, M., Bieling, P., Kennedy, S., et al. (2004). Modulation of cortical-limbic pathways in major depression: treatment-specific effects of cognitive behavior therapy. *Arch Gen Psychiatry, 61*(1), 34–41.

Gusnard, D. A., Akbudak, E., Shulman, G. L., & Raichle, M. E. (2001). Medial prefrontal cortex and self-referential mental activity: relation to a default mode of brain function. *Proc Natl Acad Sci USA, 98*(7), 4259–4264.

Iacoboni, M. (2005). Neural mechanisms of imitation. *Curr Opin Neurobiol, 15*(6), 632–637.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science, 286*(5449), 2526–2528.

Jackson, P. L., Brunet, E., Meltzoff, A. N., & Decety, J. (2006). Empathy examined through the neural mechanisms involved in imagining how I feel versus how you feel pain. *Neuropsychologia, 44*(5), 752–761.

Jackson, P. L., Meltzoff, A. N., & Decety, J. (2005). How do we perceive the pain of others? A window into the neural processes involved in empathy. *Neuroimage, 24*(3), 771–779.

Jacob, P., & Jeannerod, M. (2005). The motor theory of social cognition: a critique. *Trends Cogn Sci, 9*(1), 21–25.

Jarvelainen, J., Schurmann, M., & Hari, R. (2004). Activation of the human primary motor cortex during observation of tool use. *Neuroimage, 23*(1), 187–192.

Kalisch, R., Wiech, K., Critchley, H. D., & Dolan, R. J. (2006). Levels of appraisal: a medial prefrontal role in high-level appraisal of emotional material. *Neuroimage, 30*(4), 1458–1466.

Keightley, M. L., Winocur, G., Graham, S. J., Mayberg, H. S., Hevenor, S. J., & Grady, C. L. (2003). An fMRI study investigating cognitive modulation of brain regions associated with emotional processing of visual stimuli. *Neuropsychologia, 41*(5), 585–596.

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *J Cogn Neurosci, 14*(5), 785–794.

Keysers, C., & Gazzola, V. (2007). Integrating simulation and theory of mind: from self to social cognition. *Trends Cogn Sci, 11*(5), 194–196.

Keysers, C., Wicker, B., Gazzola, V., Anton, J. L., Fogassi, L., & Gallese, V. (2004). A touching sight: SII/PV activation during the observation and experience of touch. *Neuron, 42*(2), 335–346.

Kincade, J. M., Abrams, R. A., Astafiev, S. V., Shulman, G. L., & Corbetta, M. (2005). An event-related functional magnetic resonance imaging study of voluntary and stimulus-driven orienting of attention. *J Neurosci, 25*(18), 4593–4604.

Kosslyn, S. M., & Ochsner, K. N. (1994). In search of occipital activation during visual mental imagery. *Trends Neurosci, 17*(7), 290–292.

Kosslyn, S. M., Thompson, W. L., & Alpert, N. M. (1997). Neural systems shared by visual imagery and visual perception: a positron emission tomography study. *Neuroimage, 6*(4), 320–334.

Lanzetta, J. T., & Englis, B. G. (1989). Expectations of cooperation and competition and their effects on observers vicarious emotional responses. *J Pers Soc Psychol, 56*(4), 543–554.

Leslie, K. R., Johnson-Frey, S. H., & Grafton, S. T. (2004). Functional imaging of face and hand imitation: towards a motor theory of empathy. *Neuroimage, 21*(2), 601–607.

Leyens, J., Rodriguez-Torres, R., Vaes, J., Demoulin, S., Rodriguez-Perez, A., & Gaunt, R. (2000). The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups. *Pers Soc Psychol Rev, 4*(2), 186–197.

Lieberman, M. D. (2005). Principles, processes, and puzzles of social cognition: an introduction for the special issue on social cognitive neuroscience. *Neuroimage, 28*(4), 745–756.

Lieberman, M. D. (2007). Social cognitive neuroscience: a review of core processes. *Annu Rev Psychol, 58*, 259–289.

Lieberman, M. D. (In Press). The X and C systems. In E. Harmon-Jones & P. Winkielman (Eds.), *Fundamentals of Social Neuroscience*. New York: Guilford Press.

Lou, H. C., Luber, B., Crupain, M., Keenan, J. P., Nowak, M., Kjaer, T. W., et al. (2004). Parietal cortex and representation of the mental Self. *Proc Natl Acad Sci USA, 101*(17), 6827–6832.

Macrae, C. N., Moran, J. M., Heatherton, T. F., Banfield, J. F., & Kelley, W. M. (2004). Medial prefrontal activity predicts memory for self. *Cereb Cortex, 14*(6), 647–654.

Mayberg, H. S. (1997). Limbic-cortical dysregulation: a proposed model of depression. *J Neuropsychiatry Clin Neurosci, 9*(3), 471–481.

Mesulam, M. M., & Mufson, E. J. (1982). Insula of the old world monkey. III: efferent cortical output and comments on function. *J Comp Neurol, 212*(1), 38–52.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *J Cogn Neurosci, 17*(8), 1306–1315.

Mitchell, J. P., Heatherton, T. F., & Macrae, C. N. (2002). Distinct neural systems subserve person and object knowledge. *Proc Natl Acad Sci USA, 99*(23), 15238–15243.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50*, 1–9.

Mitchell, J. P., Neil Macrae, C., & Banaji, M. R. (2005). Forming impressions of people versus inanimate objects: social-cognitive processing in the medial prefrontal cortex. *Neuroimage, 26*(1), 251–257.

Morecraft, R. J., Geula, C., & Mesulam, M. M. (1992). Cytoarchitecture and neural afferents of orbitofrontal cortex in the brain of the monkey. *J Comp Neurol, 323*(3), 341–358.

Morrison, I., Lloyd, D., di Pellegrino, G., & Roberts, N. (2004). Vicarious responses to pain in anterior cingulate cortex: is empathy a multisensory issue? *Cogn Affect Behav Neurosci, 4*(2), 270–278.

Nagai, Y., Critchley, H. D., Featherstone, E., Fenwick, P. B., Trimble, M. R., & Dolan, R. J. (2004). Brain activity relating to the contingent negative variation: an fMRI investigation. *Neuroimage, 21*(4), 1232–1241.

Nagai, Y., Critchley, H. D., Featherstone, E., Trimble, M. R., & Dolan, R. J. (2004). Activity in ventromedial prefrontal cortex covaries with sympathetic skin conductance level: a physiological account of a "default mode" of brain function. *Neuroimage, 22*(1), 243–251.

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain-A meta-analysis of imaging studies on the self. *Neuroimage, 31*(1), 440–457.

Ochsner, K. (2007). Social cognitive neuroscience: historical development, core principles, and future promise. In A. Kruglanksi & E. T. Higgins (Eds.), *Social Psychology: A Handbook of Basic Principles* (pp. 39–66). New York, NY: Guilford Press.

Ochsner, K. N., Beer, J. S., Robertson, E. R., Cooper, J. C., Gabrieli, J. D., Kihsltrom, J. F., et al. (2005). The neural correlates of direct and reflected self-knowledge. *Neuroimage, 28*(4), 797–814.

Ochsner, K. N., Bunge, S. A., Gross, J. J., & Gabrieli, J. D. (2002). Rethinking feelings: an FMRI study of the cognitive regulation of emotion. *J Cogn Neurosci, 14*(8), 1215–1229.

Ochsner, K. N., & Gross, J. (2004). Thinking makes it so: a social cognitive neuroscience approach to emotion regulation. In R. F. Baumeister & K. Vohs (Eds.), *The Handbook of Self-Regulation* (pp. 221–255). New York: Guilford Press.

Ochsner, K. N., & Gross, J. J. (2005). The cognitive control of emotion. *Trends Cogn Sci, 9*(5), 242–249.

Ochsner, K. N., & Lieberman, M. D. (2001). The emergence of social cognitive neuroscience. *Am Psychol, 56*(9), 717–734.

Ochsner, K. N., Ray, R. D., Cooper, J. C., Robertson, E. R., Chopra, S., Gabrieli, J. D., et al. (2004). For better or for worse: neural systems supporting the cognitive down- and up-regulation of negative emotion. *Neuroimage, 23*(2), 483–499.

Ochsner, K. N., Zaki, J., Hanelin, J., Ludlow, D. H., Knierim, K., Ramachandran, T., et al. (2008). Your pain or mine? Common and distinct neural systems supporting the perception of pain in self and others. *Soc Cogn Affect Neurosci, 3*(2), 144–160.

Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: the neural systems of social fear transmission. *Soc Cogn Affect Neurosci, 2*(1), 3–11.

Olsson, A., & Phelps, E. A. (2004). Learned fear of "unseen" faces after Pavlovian, observational, and instructed fear. *Psychol Sci, 15*(12), 822–828.

Pelphrey, K. A., Morris, J. P., & McCarthy, G. (2004). Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. *J Cogn Neurosci, 16*(10), 1706–1716.

Pelphrey, K. A., Morris, J. P., Michelich, C. R., Allison, T., & McCarthy, G. (2005). Functional anatomy of biological motion perception in posterior temporal cortex: an FMRI study of eye, mouth and hand movements. *Cereb Cortex, 15*(12), 1866–1876.

Pessoa, L., Japee, S., & Ungerleider, L. G. (2005). Visual awareness and the detection of fearful faces. *Emotion, 5*(2), 243–247.

Pessoa, L., McKenna, M., Gutierrez, E., & Ungerleider, L. G. (2002). Neural processing of emotional faces requires attention. *Proc Natl Acad Sci USA, 99*(17), 11458–11463.

Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage, 16*(2), 331–348.

Posner, M. I. (1980). Orienting of attention. *Q J Exp Psychol, 32*(1), 3–25.

Raine, A., Lencz, T., Bihrle, S., LaCasse, L., & Colletti, P. (2000). Reduced prefrontal gray matter volume and reduced autonomic activity in antisocial personality disorder. *Arch Gen Psychiatry, 57*(2), 119–127; discussion 128–119.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nat Rev Neurosci, 2*(9), 661–670.

Roberts, N. A., Beer, J. S., Werner, K. H., Scabini, D., Levens, S. M., Knight, R. T., et al. (2004). The impact of orbital prefrontal cortex damage on emotional activation to unanticipated and anticipated acoustic startle stimuli. *Cogn Affect Behav Neurosci, 4*(3), 307–316.

Rolls, E. T. (2004). The functions of the orbitofrontal cortex. *Brain Cogn, 55*(1), 11–29.

Ruby, P., & Decety, J. (2001). Effect of subjective perspective taking during simulation of action: a PET investigation of agency. *Nat Neurosci, 4*(5), 546–550.

Saxe, R. (2005). Against simulation: the argument from error. *Trends Cogn Sci, 9*(4), 174–179.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage, 19*(4), 1835–1842.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia, 43*(10), 1391–1399.

Saxe, R., Xiao, D. K., Kovacs, G., Perrett, D. I., & Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia, 42*(11), 1435–1446.

Schacter, D. L., Alpert, N. M., Savage, C. R., Rauch, S. L., & Albert, M. S. (1996). Conscious recollection and the human hippocampal formation: evidence from positron emission tomography. *Proc Natl Acad Sci USA, 93*(1), 321–325.

Scherer, K. R., Schorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal Processes in Emotion*: Oxford University Press.

Schoenbaum, G., Chiba, A. A., & Gallagher, M. (1999). Neural encoding in orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning. *J Neurosci, 19*(5), 1876–1884.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science, 303*(5661), 1157–1162.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature, 439*(7075), 466–469.

Stone, V. E., Cosmides, L., Tooby, J., Kroll, N., & Knight, R. T. (2002). Selective impairment of reasoning about social exchange in a patient with bilateral limbic system damage. *Proc Natl Acad Sci USA, 99*(17), 11531–11536.

Sturmer, S., Snyder, M., & Omoto, A. M. (2005). Prosocial emotions and helping: the moderating role of group membership. *J Pers Soc Psychol, 88*(3), 532–546.

Tomasello, M. (2000). *The Cultural Origin of Human Cognition*. Cambridge, MA: Harvard University Press.

Vaughan, K. B., & Lanzetta, J. T. (1980). Vicarious instigation and conditioning of facial expressive and autonomic responses to a model's expressive display of pain. *J Pers Soc Psychol, 38*(6), 909–923.

Vogeley, K., & Fink, G. R. (2003). Neural correlates of the first-person-perspective. *Trends Cogn Sci, 7*(1), 38–42.

Vogt, B. A., Vogt, L., & Laureys, S. (2006). Cytology and functionally correlated circuits of human posterior cingulate areas. *Neuroimage, 29*(2), 452–466.

Whalen, P. J., Kagan, J., Cook, R. G., Davis, F. C., Kim, H., Polis, S., et al. (2004). Human amygdala responsivity to masked fearful eye whites. *Science, 306*(5704), 2061.

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. *J Neurosci, 18*(1), 411–418.

Whalen, P. J., Shin, L. M., McInerney, S. C., Fischer, H., Wright, C. I., & Rauch, S. L. (2001). A functional MRI study of human amygdala responses to facial expressions of fear versus anger. *Emotion, 1*(1), 70–83.

Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust. *Neuron, 40*(3), 655–664.

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: the interpersonal nature of empathic accuracy. *Psychol Sci, 19*(4), 399–404.

Zaki, J., & Ochsner, K. (2009). The need for a cognitive neuroscience of naturalistic social cognition. *Ann N Y Acad Sci, 1167*, 16–30.

Zaki, J., Ochsner, K. N., Hanelin, J., Wager, T., & Mackey, S. C. (2007). Different circuits for different pain: patterns of functional connectivity reveal distinct networks for processing pain in self and others. *Soc Neurosci, 2*(3 & 4), 276–291.

CHAPTER 3
Distributed Process for Retrieval of Person Knowledge

Maria Ida Gobbini

"The visual appearance of a face … is immediately and obligatorily transformed into the representation of a person (with dispositions and intentions) before having access to consciousness."—Leslie Brothers (1989), *The social brain: A project for integrating primate behavior and neuropsychology in a new domain,* Concepts in Neuroscience, 1: 27–51, p. 35.

"…Even the simple act which we describe as "seeing someone we know" is to some extent an intellectual process. We pack the physical outline of the person we see with all the notions we have already formed about him, and in the total picture of him which we compose in our minds those notions have certainly the principle place. In the end they come to fill out so completely the curve of his cheeks, they blend so harmoniously in the sound of his voice as if it were no more than a transparent envelope, that each time we see the face or hear the voice it is these notions which we recognize and to which we listen…."
—Marcel Proust, *In search of lost time*

Face perception plays a fundamental and multifaceted role in social communication. We have proposed that face perception is mediated by a distributed neural system that includes numerous brain regions, including face-selective regions in extrastriate visual cortex (the "core system") and areas for other functions such as emotion, action understanding, and person knowledge (Haxby, Hoffman, & Gobbini, 2000). In this chapter, I focus on the aspects of this distributed system that are involved in recognition of familiar faces and the concomitant activation of associated person knowledge and emotion. Finally, I present an amplified version of our model that explicitly incorporates the systems for familiar face recognition and for understanding facial gestures such as expression and eye gaze.

In humans and nonhuman primates, neuro-imaging of the hemodynamic response to faces as compared to other categories of objects has consistently identified areas that show a stronger response to faces (Sergent, Otha, & MacDonald, 1992; Clark, Keil, Maisog, Courtney, Ungerleider, & Haxby, 1996; Kanwisher, McDermott, & Chun, 1997; McCarthy, Puce, Gore, & Allison, 1997; Tsao et al., 2006). In humans the most prominent face-selective area is in the lateral fusiform gyrus. Kanwisher et al. (1997) proposed that this area is a module specialized for face perception and named it the "Fusiform Face Area" (FFA). The existence of a module uniquely specialized for face perception is still a matter of controversy (Hanson, Matsuka, & Haxby, 2004; Hanson & Halchenko, 2007). In an ongoing debate, the principal alternative accounts are the "distributed object form topography" hypothesis (faces and different categories of objects are represented by distributed and overlapping patterns of responses in the ventral temporal cortex; Haxby, Gobbini, Furey, Ishai, Schouten, & Pietrini, 2001), and the expert visual recognition hypothesis (the stronger response in the FFA is driven by expertise and not by faces *per se*; Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999).

Beside the FFA, neuro-imaging experiments have shown that other areas also respond significantly more to faces as compared to other objects, such as the inferior occipital gyri (OFA) and the posterior superior temporal sulcus (pSTS). Moreover, areas that are involved in cognitive functions other than face perception, such as emotion and social cognition, also respond significantly to faces.

Our model for a distributed neural system that mediates face perception in humans (Haxby, Hoffman, & Gobbini, 2000) (Fig. 3–1) was inspired by the cognitive model proposed by Bruce and Young (1986). In this model we suggested that face perception is mediated by spatially distributed processes across multiple regions. We divided the face-responsive regions into two systems: the "core system" and the "extended system." The "core system" consists of the FFA, OFA, and pSTS and is involved in encoding the visual appearance of a face. The FFA and the pSTS mediate the encoding of two broad classes of visual information about the appearance of faces. Although the FFA is involved in the representation of the invariant features of faces, the pSTS is involved in encoding dynamic features of faces (but *see also* Calder & Young, 2005). The areas of the "extended system" are recruited in concert with the areas of the core system to process information conveyed by a face, such as biographical information, direction of attention, and emotion.

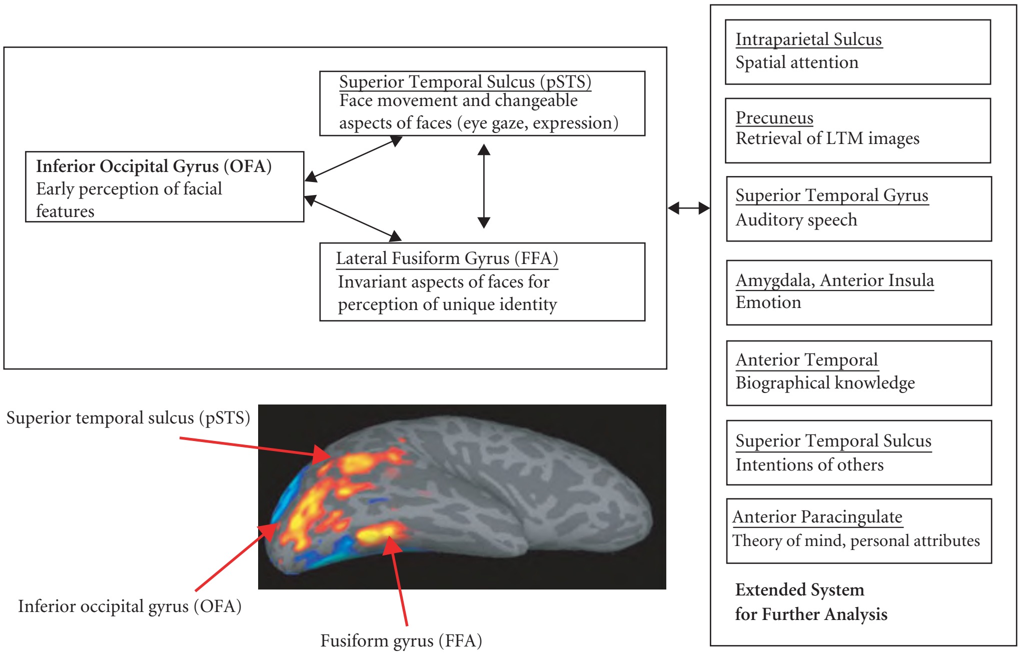

Superior Temporal Sulcus (pSTS)
Face movement and changeable aspects of faces (eye gaze, expression)
Inferior Occipital Gyrus (OFA)
Early perception of facial features
Lateral Fusiform Gyrus (FFA)
Invariant aspects of faces for perception of unique identity
Intraparietal Sulcus
Spatial attention
Precuneus
Retrieval of LTM images
Superior Temporal Gyrus
Auditory speech
Amygdala, Anterior Insula
Emotion
Anterior Temporal
Biographical knowledge
Superior Temporal Sulcus
Intentions of others
Anterior Paracingulate
Theory of mind, personal attributes
Extended System
for Further Analysis
Superior temporal sulcus (pSTS)
Inferior occipital gyrus (OFA)
Fusiform gyrus (FFA)

Fig. 3–1 Model of the distributed processing for face perception in humans. The core system (Inferior Occipital Gyrus, Fusiform Gyrus and Superior Temporal Sulcus) participates in the processing of perceptual characteristics of faces. The extended system is shared with other cognitive functions other than face perception and participates in extracting the meaning that a face can convey (Haxby, Hoffman, & Gobbini, 2000).

Most of the research that will be reviewed in this chapter involves studies of face perception with functional neuro-imaging, especially fMRI, a noninvasive technique that gives an indirect measure of neural activity as reflected in hemodynamic changes evoked by that activity.

Representation of familiar faces

Ready access to information about familiar individuals when we encounter them is necessary for effective social exchanges. Recognition of familiar faces, therefore, begins with recognizing the individual based on the appearance but also must include retrieval of personal information and emotional responses.

Previous neuro-imaging research on familiar face recognition has focused mostly on the fusiform gyrus and anterior temporal regions. The modulation of responses to faces in the fusiform gyrus based on familiarity has been inconsistent across studies (Gorno-Tempini et al. 1998; Dubois et al. 1999; Henson, Shallice, & Dolan, 2000; Leveroni, Seidenberg, Mayer, Mead, Binder, & Rao, 2000; Nakamura et al., 2000; Rossion, Schiltz, Robaye, Pirenne, & Crommelinck, 2001; Gobbini, Leibenluft, Santiago, & Haxby, 2004; Leibenluft, Gobbini, Harrison, & Haxby, 2004; Rotshtein, Henson, Treves, Driver, & Dolan, 2005). The anterior temporal cortex, by contrast, has shown stronger activation for a variety of familiar as compared to unfamiliar stimuli, such as names (Gorno Tempini et al., 1998), familiar landscapes (Nakamura et al., 1998), and faces (Sergent, Ohta, & MacDonald, 1992; Gorno Tempini et al., 1998; Leveroni et al., 2000; Rotshtein et al., 2005), suggesting that this region plays a role in the retrieval of biographical or autobiographical information.

The lack of consistent results in the response to faces in the fusiform gyrus could result from several reasons. First, different experiments have used different types of familiar faces. In some cases, the familiar faces were famous familiar faces, in other cases faces of acquaintances, and in other cases faces that became familiar through experimental training. Hence, the different types of familiar faces were characterized by different types of social and emotional attachment. Second, different experiments have used different tasks that placed different demands on attention. Recognition of a familiar individual extends beyond the visual representation of that person's face, and, therefore, the fusiform gyrus may be influenced by top-down modulation from areas of the extended system (Gobbini & Haxby, 2007). Research in social psychology has provided evidence for the spontaneous activation of traits and attitudes associated with perceived individuals (Bargh, Chen, & Burrows 1996; Greenwald & Banaji 1995; Todorov & Uleman 2002; Todorov, Gobbini, Evans, & Haxby, 2007). Furthermore, the representation of significant others is richer in thoughts, feelings, and emotions as compared to the representation of nonsignificant others (Andersen & Cole, 1990; Andersen Glassman & Gold, 1998). As someone becomes more familiar, the inferences made about that individual are related more to "psychological mediating variables" (such as goals and beliefs) and less to broad uncontextualized traits (for example, "aggressive" or "friendly") (Idson & Mischel, 2001).

To further investigate the neural representation of familiar faces, we designed three fMRI experiments. In all of these experiments, different groups of participants performed the same type of task: "one back repetition detection" based on identity ("Is this person the same as the one shown in the previous picture?"). Consecutive pictures of the same person were always different images. The task was based on the perceptual characteristics of the faces and did not demand person evaluation or knowledge retrieval. The purpose of the task was to induce equal attention to all the stimuli without explicit retrieval of information about the person. Hence, any effect of familiarity on the neural responses to faces most likely reflected the spontaneous retrieval of person knowledge.

In this chapter, I summarize the findings of these three fMRI experiments and discuss them in relation to relevant literature. I propose that recognition of familiar individuals is the result of a spatially distributed process that involves not only perceptual areas but also areas that are involved in cognitive and social functions other than visual perception. I highlight how visual familiarity, the spontaneous retrieval of person knowledge, and the emotional response are all integral components for recognition of familiar individuals. I also focus attention on different aspects of person knowledge that are retrieved during recognition of familiar individuals. Person knowledge refers to many kinds of information about familiar individuals, including personal traits, beliefs, goals, mental states, attitudes, and intentions, as well as more objective types of information such as biographical and episodic memories. I also highlight the role of emotional responses during recognition of familiar individuals, concentrating in particular on the modulation of activity in the amygdala and the insula. The differential hemodynamic response to familiar faces in those regions supports their role in social interactions and person recognition. Finally, I amplify our hypothesis about the role of the regions identified in these experiments and propose a modified version of our model of the distributed human neural system for face perception (Haxby, Hoffman, & Gobbini, 2000), concentrating on those brain regions that play a key role in the recognition of familiar faces (Gobbini & Haxby, 2007).

FMRI STUDIES ON FAMILIAR FACE RECOGNITION

The three fMRI experiments explored different aspects of face familiarity. In the first two fMRI experiments, we investigated familiarity that accrues naturally with years of exposure and social interactions. The third fMRI experiment was designed as a control for the first two studies to isolate the role of visual familiarity from the role of person knowledge during face recognition. In this third experiment we investigated visual, experimentally induced familiarity with no associated biographical or semantic information.

In the first two fMRI experiments, we contrasted the hemodynamic response to different groups of familiar faces characterized by different social and emotional attachment. In the first fMRI experiment we compared the neural response to personally familiar faces (faces of relatives and friends) versus the neural response to faces that are familiar because of the media (politicians, actors, singers, athletes) versus the neural response to strangers (*see* Gobbini et al., 2004, for more details). In the second fMRI experiment, we recruited mothers and measured neural responses while they viewed pictures of their own child, familiar but unrelated children, and unfamiliar children (*see* Leibenluft et al., 2004, for more details). In the third fMRI experiment, we induced visual familiarity experimentally (through a behavioral training prior to the imaging session) for a set of faces chosen randomly from a pool of novel faces and then recorded the neural responses to the visually familiar faces versus novel faces. During the behavioral training session, visual familiarity was induced by asking the participants to perform a delayed matching task that consisted of viewing an isolated feature and then assigning this feature to one of three target faces (*see* Gobbini & Haxby, 2006, for more details).

The results of these experiments demonstrated modulation of activity by familiarity in a distributed set of areas, including regions that have been associated with "theory of mind" (ToM) tasks, with retrieval of episodic memory, and with emotional response, as well as in extrastriate visual cortex in the fusiform gyrus.

RETRIEVAL OF PERSON KNOWLEDGE

We found modulation of the hemodynamic response to familiar faces based on the type of familiarity in the anterior paracingulate cortex (APC), the temporoparietal junction (TPJ), and in the posterior cingulate/precuneus (PCC/PC) (Fig. 3–2). Familiarity modulated activity in the APC, TPJ, and the PCC/PC, with a stronger response to personally familiar faces as compared to faces of famous familiar individuals and to faces of strangers, a stronger response to the face of one's own child as compared to the face of a familiar unrelated child, and a stronger response to familiar children as compared to unfamiliar children.

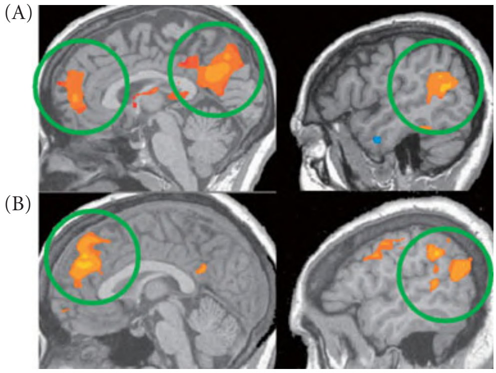


Fig. 3–2 Example of activation in the APC, the TPJ and in the PCC/PC for the contrast "personally familiar faces versus famous familiar faces" (Gobbini et al., 2004) (A) and for the contrast "one's own child versus familiar unrelated child" (Leibenluft et al., 2004) (B). The more familiar faces evoked a stronger response in these areas that are associated with retrieval of person knowledge.

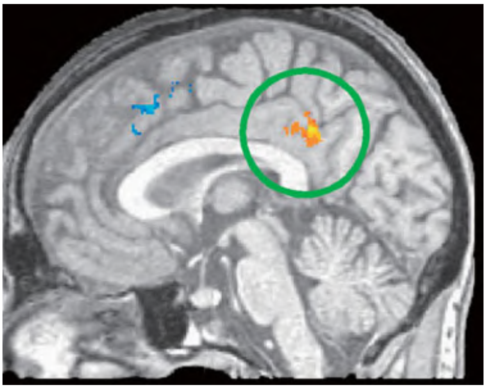


Fig. 3–3 The contrast "faces associated with experimentally induced visually familiarity versus novel faces" evoked a stronger response in the precuneus/posterior cingulate cortex but did not modulate the hemodynamic response in the theory of mind areas (Gobbini & Haxby, 2006).

The APC and the TPJ have been identified as consistently activated in neuro-imaging experiments exploring theory of mind, independently of the modality of input (Frith & Frith, 1999; McCabe, Houser, Ryan, Smith, & Trouard, 2001; Castelli, Happe, Frith, & Frith, 2000; Gallagher, Happe, Brunswick, Fletcher, Frith, & Frith, 2000; Saxe & Kanwisher, 2003; Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004; Amodio & Frith 2006; Frith & Frith 2006; Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007). "Theory of mind" refers to the capacity that allows one to explain and to predict someone else's behavior based on one's construal of that person's mental states (Leslie, 1994; Frith & Frith, 1999).

The APC and the TPJ are activated during tasks that require mentalizing (ToM). We propose that activity in the ToM areas and the PCC/PC is associated with the neural representation of information about familiar individuals such as personal traits, intentions, attitudes, transient mental states, and biographical information that is spontaneously retrieved in the act of recognition. We have hypothesized that the PCC/PC also plays an important role during recognition of familiar individuals and in the acquisition of familiarity with faces. More generally, the PCC/PC plays a role in the retrieval of episodic information and imagery from long-term memory (Ishai, Ungerleider, & Haxby, 2000; Burgess, Maguire, Spiers, & O'Keefe, 2001).

In our fMRI experiment, we used an implicit task (described above) that did not require explicit identification of the pictured individuals. The pattern of modulation of response by familiarity during performance of an implicit task suggests that person knowledge is retrieved spontaneously when we see someone we know.

In the third experiment, faces that were visually familiar but were not associated with person knowledge did not evoke stronger responses in the APC and TPJ. Notably, however, simple visual familiarity did evoke a stronger response in the PCC/PC as compared to novel faces (Fig. 3–3). The contrast between the results in this experiment as compared to those of the first two experiments supports the hypothesis that the APC and TPJ encode aspects related to personal traits, intentions, and transient mental states (Allison, Puce, & McCarthy, 2000; Mitchell, Heatherton, & Macrae, 2002; Amodio & Frith 2006; Heatherton, Wyland, Macrae, Demos, Denny, & Kelley, 2006), whereas the PCC/PC and the anterior temporal regions are involved in the retrieval of episodic memory (Burgess et al., 2001; Fletcher, Frith, Baker, Shallice, Frackowiak, & Dolan,

1995) and biographical information (Sergent et al., 1992; Damasio, Grabowski, Tranel, Hichwa, & Damasio, 1996; Gorno Tempini et al., 1998; Leveroni et al., 2000; Nakamura et al., 2000; Rotshtein et al., 2005).

Role of Theory of Mind Areas in Retrieval of Person Knowledge

Neuro-imaging research indicates that the APC and TPJ play a key role in mediating the representation of the personal attributes and mental states of others. A wide variety of tasks that require interpreting and predicting someone else's behavior activate both of these areas (Frith & Frith, 1999; Castelli et al., 2000; Gallagher & Frith, 2003; Saxe & Kanwisher, 2003; Gobbini et al., 2007). A stronger hemodynamic response has been recorded in these areas when subjects read stories or look at cartoons that require understanding that a character's beliefs are false (Gallagher et al., 2000; Saxe & Kanwisher, 2003; Gobbini et al., 2007), when subjects view animations with geometrical figures interacting in a way that implies specific mental states as compared to figures moving in random ways (Heider & Simmel animations) (Castelli et al., 2000; Martin & Weisberg, 2003; Gobbini et al., 2007), when subjects play competitive games against human partners as compared to a computer (Gallagher et al., 2000; McCabe et al., 2001; Rilling et al., 2004), and when subjects make moral decisions that involve awareness of the direct consequences to a victim who is clearly represented as an individual (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001).

The APC and TPJ play different roles in representing person knowledge. The APC cortex is more involved in encoding personal traits (Mitchell et al., 2002; Heatherton et al., 2006) and mental states of others (Calder et al., 2002; Amodio & Frith, 2006; Frith & Frith, 2006), whereas the TPJ plays a more general role in social cognition that is related more to the representation of other people's intentions (Allison et al., 2000; Hoffman & Haxby, 2000; Perrett et al., 1985; Puce & Perrett, 2003; Winston, Strange, O'Doherty, & Dolan, 2002; Gobbini et al., 2007). In earlier reports, including our own studies of face familiarity, the TPJ region has often been referred to as the posterior superior temporal sulcus (pSTS). In a recent study with a meta-analysis of earlier reports (Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007) we found that the region that is associated with ToM and person knowledge is in the more posterior and superior TPJ, as compared to the pSTS region that is associated more with the perception of biological motion and facial movement. This dissociation is consistent with the hypothesis of Saxe (2006).

The spontaneous retrieval of information about personal traits, intentions, mental states, and attitudes of someone we know in the act of recognition prepares one to interact appropriately and effectively with that person.

Episodic Memory

The PCC/PC is activated by a variety of familiar stimuli independently of the modality of input. For example, familiar faces, voices, and names, as compared to unfamiliar faces, voices, and names, evoke a stronger hemodynamic response in this region (Nakamura et al. 2001; Gorno Tempini et al., 1998; Suguira, Shah, Zilles, & Fink, 2005). Tasks requiring long-term memory or imagery also activate this region (Ishai et al., 2000; Burgess et al., 2001).

Our fMRI experiments demonstrate that the PCC/PC responds more strongly to familiar faces even in the absence of associated person knowledge. Unlike the effect of familiarity in the APC and TPJ, however, simple visual familiarity also increased the response to faces in the PCC/PC (Fig. 3–3; Gobbini & Haxby, 2006), suggesting that this region plays a role in the acquisition of visual familiarity (Kosaka et al., 2003). Because of the high number of repetitions of individual faces in the experiment on simple visual familiarity, we could track the change in response to faces that were unfamiliar at the beginning of the fMRI session but became visually familiar by virtue of simple repetition. We observed an increase of the response in the PCC/PC to the unfamiliar faces over the first 20 repetitions (Fig. 3–4). These findings support the hypothesis of a key role of the PCC/PC in the acquisition of familiarity with faces.

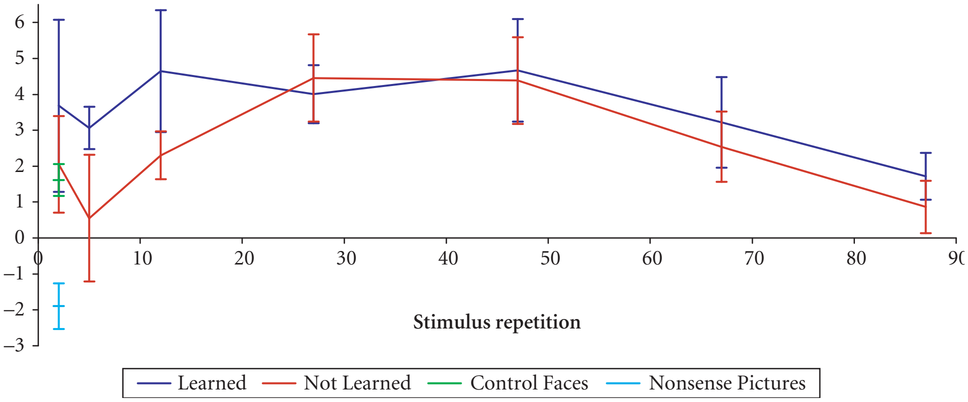


Fig. 3–4 Adaptation of the response to faces in the precuneus over multiple repetitions. Repetition of novel faces ("Not Learned") induced an increase in the hemodynamic response during the first 20 presentations as compared to faces that were visually familiar ("Learned") because of behavioral training. Subsequent repetitions induced a progressive decrease in the hemodynamic response with no significant difference between the two categories. Other control stimuli used in the fMRI experiment included seldom-repeated novel faces ("Control Faces": these were faces that were repeated only five to six times during the entire experiment) and scrambled version of the faces ("Nonsense Pictures") (for details on the experimental design and results, *see* Gobbini & Haxby, 2006).

Lesion studies of the anterior temporal areas have demonstrated impairment in accessing semantic information about people (Ellis, Young, & Critchley, 1989; Damasio et al., 1996). Several imaging experiments have also shown a consistently stronger neural response to familiar stimuli (faces, names, landscapes) in the anterior temporal regions, suggesting that these areas may be involved in representation of biographical or autobiographical information (Sergent et al., 1992; Gorno Tempini et al., 1998; Leveroni et al., 2000; Nakamura et al., 2000; Rotshtein et al., 2005).

The stronger response for the more familiar faces recorded in PCC/PC and in the anterior temporal regions might indicate the involvement of these areas in retrieval of episodic memories and biographical information associated with familiar individuals.

Emotional response

Another key component for the recognition of familiar individuals is the emotional response we experience when seeing someone we know.

Simple visual familiarity with faces, even when the visual familiarity is induced in an experimental setting, is sufficient to induce a weaker response in the amygdala as compared to novel faces (Dubois et al., 1999; Schwartz et al., 2003; Gobbini & Haxby, 2006). Personally familiar faces as compared to famous familiar faces and to faces of strangers also evoke a weaker response in the amygdala (Gobbini et al., 2004; Leibenluft et al., 2004) (Fig. 3–5).

Functional imaging studies have shown that the amygdala is sensitive to emotionally relevant stimuli with both positive and negative valence (Breiter et al., 1996; Canli, Sivers, Whitfield, Gotlib, & Gabrieli, 2002; Morris et al., 1996; Zalla et al., 2000). Studies of nonhuman primates and case reports of patients with selective lesions of the amygdala suggest that this anatomical structure plays a role in social interactions. Mature macaque monkeys with bilateral amygdala lesions exhibit socially uninhibited behavior and a lack of fear for stimuli that represent a potential threat (Klüver & Bucy, 1938). These findings suggest that the amygdala functions as a "social brake" and plays a role in producing a cautious attitude when approaching a new environment (Amaral, 2002). Patients with amygdala resection—especially if it involves the left

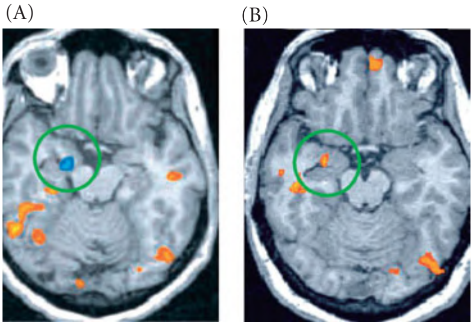


Fig. 3–5 Example of activations in the amygdala. (A) Personally familiar faces evoked a weaker response as compared to the famous familiar faces (Gobbini et al., 2004) but (B) the face of one's own child evoked a stronger response as compared to the face of a familiar unrelated child (Leibenluft et al., 2004). The right side of the brain is on the left side of each image (radiological convention).

amygdala—do not show enhanced perception for aversive stimuli (Anderson & Phelps, 2001). Furthermore, patients with bilateral amygdala lesions rate as trustworthy faces that normal subjects rate as unapproachable and untrustworthy (Adolphs, Tranel, & Damasio, 1998). In normal volunteers, perception of untrustworthy faces elicits activity in the amygdala during both explicit and implicit processing of faces (Winston et al., 2002; Engell, Haxby, & Todorov, 2007). These findings support the hypothesis that the amygdala may be sensitive to unexpected or unfamiliar events with potential biological importance (Davis & Whalen, 2001) such as unfamiliar faces (Gobbini & Haxby, 2007).

Reduced amygdala activity was found in response to the most familiar faces (relatives and friends) as compared to famous familiar faces and to faces of strangers (Gobbini et al., 2004) (Fig. 3–5) and in the response to the face of a lover (Bartels & Zeki, 2000). The reduced activity of the amygdala in response to personally familiar faces might reflect a lower level of vigilance when encountering someone we know. The stronger response of the amygdala to faces of strangers could reflect the role of this anatomical structure in mediating a cautious and wary attitude when encountering someone new.

Although viewing familiar unrelated children as compared to unfamiliar children induced a weaker response in the amygdala, consistent with the general effect of familiarity, viewing the face of one's own child evoked a stronger response in the amygdala (Leibenluft et al., 2004) (Fig. 3–5). The stronger amygdala response when seeing one's own child may reflect both the strong positive emotional attachment and the vigilant protectiveness of maternal feelings. Indeed, viewing the face of one's own child also evoked a stronger response in the insula. The insula appears to be associated with stimuli that evoke strong visceral sensations. The insula responds more strongly when viewing the face of one's beloved (Bartels & Zeki, 2000), or when viewing a loved one experiencing pain (Singer, Seymour, O'Doherty, Kaube, Dolan, & Frith, 2004), suggesting that the insula might play a role in mediating empathic reactions. Imitations of facial expressions also evoke a strong response in the insula (Carr, Iacoboni, Dubeau, Mazziotta, & Lenzi, 2003). Negatively valenced stimuli such as expressions of disgust (Calder, Lawrence, & Young, 2001; Phillips, Drevets, Rauch, & Lane, 2003) or being treated unfairly during negotiation games (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003) also elicit activity in the insula.

The intense attachment and protectiveness that characterizes the maternal relationship is reflected in increased activity in the amygdala and insula elicited by viewing the face of one's child.

Visual familiarity

As described above, reports on the effect of familiarity in perceptual areas have not been consistent across different neuro-imaging experiments.

Evoked potential studies have shown that modulation of the response by familiarity appears at a later latency than the first face-specific potentials. Whereas early face-specific evoked potentials are recorded in posterior temporal locations, the later potentials that are modulated by familiarity are recorded in frontal and parietal locations (Puce, Allison, Bentin, Gore, & McCarthy, 1998; Bentin, Deouell, & Soroker, 1999; Eimer 2000). Therefore, the early response to faces may represent a rapid feed-forward process that does not carry information about familiarity. Recognition of familiar individuals may be achieved through the interactions from other face-responsive areas at a later latency (Puce et al. 1999).

In our fMRI research on familiar face recognition, we found a complex, nonmonotonic modulation of activity in face-selective regions of ventral temporal cortex. Famous familiar faces evoked a weaker response in the fusiform gyrus as compared to the faces of strangers, but personally familiar faces evoked an equivalent response as compared to faces of strangers and a stronger response as compared to the famous familiar faces (Gobbini et al., 2004). Faces of familiar unrelated children evoked a weaker response than the faces of unfamiliar children, but the face of one's own child evoked a stronger response than the face of a familiar unrelated child (Leibenluft et al., 2004). Faces that were visually familiar with no associated person knowledge evoked a weaker response than novel faces (Gobbini & Haxby, 2006). We propose that the areas we have identified that encode person knowledge and participate in the emotional response to faces play a major role in modulating the response to familiar faces through feedback to the extrastriate cortex. Thus, the nonmonotonic modulation of response by familiarity in the fusiform gyrus may reflect both a weaker early response to visually familiar faces caused by more rapid or efficient processing and stronger later responses caused by top-down modulation associated with person knowledge and emotion.

Dissociation of the emotional response and visual recognition

Recognition of familiar individuals entails recognition of the visual appearance, spontaneous retrieval of person knowledge, and an appropriate emotional response. These components are all essential for successful recognition. Evidence that recognition of familiar individuals is the result of a distributed process involving multiple areas comes also from neuropsychological studies of patients. These studies demonstrate that the multiple components that participate in face recognition are dissociable and that the impairment of any one of these components can disrupt normal recognition. Classical examples are patients affected by prosopagnosia and Capgras' syndrome. Prosopagnosia is a neurological disorder characterized by the inability to explicitly recognize the identity of a familiar person based on visual appearance. Several lines of evidence, however, demonstrate that these patients can implicitly recognize familiar faces (Bauer 1984, 1986; De Haan, Bauer, & Greve, 1992; Tranel & Damasio, 1985), as evidenced by normal augmentation of skin conductance response to familiar as compared to unfamiliar faces.

By contrast, patients with Capgras' syndrome are able to recognize the identity of a familiar face but they deny the "authenticity" of such a face. These patients believe that familiar people, most frequently family members and friends, have been replaced by impostors, aliens, or robots (Capgras & Reboul-Lachaux, 1923; Ellis & Lewis, 2001; Hirstein & Ramachandran, 1997). These findings suggest that when visual recognition is accompanied by an altered emotional response, recognition of a familiar individual is not normal.

New model for face recognition

Various models have been proposed for the cognitive and neural systems that mediate face recognition (Bauer, 1984; Bruce & Young, 1986; Ellis & Lewis, 2001; Haxby, Hoffman, & Gobbini, 2000). Based on the data from our fMRI experiments on recognition of familiar faces, we recently proposed a functional model (Gobbini & Haxby, 2007), which is a modified version of our previous model of the distributed human neural system for face perception (Haxby, Hoffman, & Gobbini 2000). The original model suggested that face processing is spatially distributed across several cortical regions. The modified version of

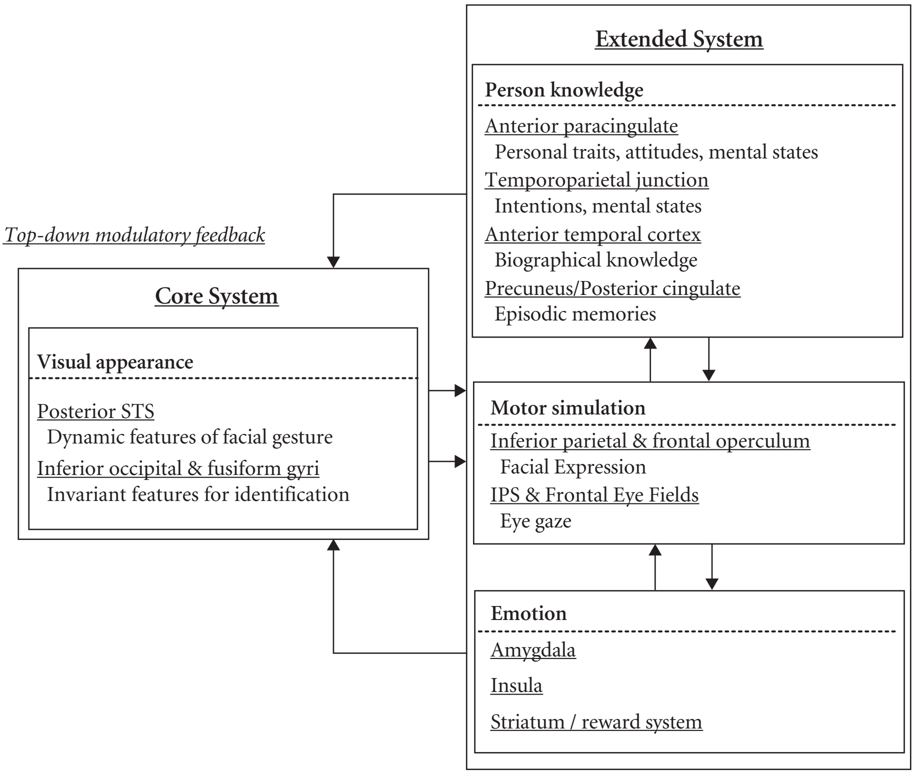


Fig. 3–6 Model of the distributed processing for recognition of familiar faces. The core system deals with the encoding of the visual appearance of a face whereas the extended system extracts further information helpful in recognizing a known individual. In this model, particular emphasis has been put on the areas that participate in retrieval of person knowledge and in the emotional response to familiar faces (Gobbini & Haxby, 2007).

the original model emphasizes the components that are fundamental for the representation of familiar individuals—namely, the systems that participate in the visual analysis of a face, the representation of person knowledge, and the emotional response to someone we know. The latest version of this model (Fig. 3–6) also includes neural systems that participate in processing perception of facial movements (expression and gaze) by simulating the motor programs that produce such movements (Carr et al., 2003; Grosbras & Paus, 1996; Montgomery & Haxby, 2008). As proposed in the original model, the areas that are face-responsive are grouped in the core system and in the extended system. As part of the core system, the fusiform gyrus participates in recognition of the visual appearance of a familiar face, based on aspects of facial structure that are invariant over facial movements. The pSTS, on the other hand, processes dynamic changes such as expression and eye gaze (Puce et al., 1998; Hoffman & Haxby, 2000; Winston et al., 2004; Engell & Haxby, 2007). As part of the extended system, we listed areas that we propose participate in familiar face recognition. We further divided these areas into those that play a role in the representation of person knowledge and those that play a role in the emotional response to familiar faces. The APC and the TPJ participate in retrieval of personal traits, intentions, goals, and mental states of familiar individuals. The anterior temporal regions and the PCC/PC are involved with the representation of semantic information and episodic memory.

Among the areas that participate in the emotional response, the amygdala plays a key role in recognition of familiar individuals. This anatomical structure shows a complex modulation based on the type of familiarity related to its possible role in mediating wary and vigilant reactions when encountering someone new and increased vigilance for one's own children. The insula shows an increased response for certain faces with whom one has a particularly intense emotional relationship.

Conclusions

Recognition of a familiar individual activates a distributed network of brain regions related not only to that person's visual appearance but also to knowledge about his or her personality traits, mental states, goals, and intentions, to episodic memories, and to one's emotional response. Recognition of visual appearance, the spontaneous retrieval of social and personal information, and the emotional response to someone we know are all necessary for successful and effective social interactions.

In this chapter, we have reviewed our data and relevant literature on recognition of familiar individuals and summarized the major findings in a modified version of a model of the distributed system for face perception in humans that highlights the areas that are spontaneously activated every time we meet someone we know.

Acknowledgment

I would like to thank James V Haxby for helpful discussions.

References

Adolphs R, Tranel D, & Damasio AR. (1998). The human amygdala in social judgment. *Nature*, 393, 470–474.

Allison T, Puce A, & McCarthy G. (2000). Social perception from visual cues: role of the STS region. *Trends Cogn Sci.* 4, 267–278.

Amaral DG. (2002). The primate amygdala and the neurobiology of social behavior: implications for understanding social anxiety. *Biol Psychiatry.* 51, 11–17.

Amodio DM, & Frith CD. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nat Rev Neurosci.* 7, 268–277.

Andersen SM, & Cole SW. (1990). "Do I know you?" The role of significant others in general social perception. *J Pers Soc Psychol.* 59, 384–399.

Andersen SM, Glassman NS, & Gold DA. (1998). Mental representations of the self, significant others, and nonsignificant others: structure and processing of private and public aspects. *J Pers Soc Psychol.* 75, 845–861.

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

Bartels A, & Zeki S. (2000). The neural basis of romantic love. *Neuroreport.* 11, 3829–3834.

Bargh JA, Chen M, & Burrows L. (1996). Automaticity of social behavior: direct effects of trait construct and stereotype-activation on action. *J Pers Soc Psychol.* 71, 230–244.

Bauer RM. (1984). Autonomic recognition of names and faces in prosopagnosia: a neuropsychological application of the Guilty Knowledge Test. *Neuropsychologia* 22, 457–469.

Bauer RM. (1986). The cognitive psychophysiology of prosopagnosia. In H. Ellis, M. Jeeves, F. Newcombe, & A.Young (Eds), *Aspects of face processing* (pp. 253–267). Dordrecht, The Netherlands: Martinus Nijhoff.

Bentin S, Deouell LY, & Soroker N. (1999). Selective visual streaming in face recognition: evidence from developmental prosopagnosia. *Neuroreport* 10, 823–827.

Breiter HC, Etcoff NL, Whalen PJ, et al. (1996). Response and habituation of the human amygdala during visual processing of facial expression. *Neuron* 17, 875–887.

Bruce V, & Young A. (1986). Understanding face recognition. *Br J Psychol.* 77, 305–327.

Burgess N, Maguire EA, Spiers HJ, & O'Keefe J. (2001). A temporoparietal and prefrontal network for retrieving the spatial context of lifelike events. *Neuroimage* 14, 439–453.

Calder AJ, Lawrence AD, & Young AW. (2001). Neuropsychology of fear and loathing. *Nat Rev Neurosci.* 2, 352–363.

Calder AJ, Lawrence AD, Keane J, et al. (2002). Reading the mind from eye gaze. *Neuropsychologia* 40, 1129–1138.

Calder AJ, & Young AW. (2005). Understanding the recognition of facial identity and facial expression. *Nat Rev Neurosci.* 6, 641–651.

Canli T, Sivers H, Whitfield SL, Gotlib IH, & Gabrieli JD. (2002). Amygdala response to happy faces as a function of extraversion. *Science* 296, 2191.

Capgras JMJ, & Reboul-Lachaux J. (1923). L'illusion des "sosies" dans un délire systématisé chronique. *Bull Soc Clin Med Ment.* 11, 6–16.

Carr L, Iacoboni M, Dubeau MC, Mazziotta JC, & Lenzi GL. (2003). Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proc Natl Acad Sci USA* 100, 5497–5502.

Castelli F, Happe F, Frith U, & Frith C. (2000). Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage* 12, 314–325.

Clark VP, Keil K, Maisog JM, Courtney S, Ungerleider LG, & Haxby JV. (1996). Functional magnetic resonance imaging of human visual cortex during face matching: a comparison with positron emission tomography. *Neuroimage* 4, 1–15.

Damasio H, Grabowski TJ, Tranel D, Hichwa RD, & Damasio AR. (1996). A neural basis for lexical retrieval. *Nature* 380, 499–505.

Davis M, & Whalen PJ. (2001). The amygdala: vigilance and emotion. *Mol Psychiatry.* 1, 13–34.

De Haan EH, Bauer RM, & Greve KW. (1992). Behavioural and physiological evidence for covert face recognition in a prosopagnosic patient. *Cortex* 28(1), 77–95.

Dubois S, Rossion B, Schiltz C, et al. (1999). Effect of familiarity on the processing of human faces. *Neuroimage* 9, 278–289.

Eimer M. (2000). The face-specific N170 component reflects late stages in the structural encoding of faces. *Neuroreport* 11, 2319–2324.

Ellis AW, Young AW, & Critchley EM. (1989). Loss of memory for people following temporal lobe damage. *Brain* 112, 1469–1483.

Ellis HD, & Lewis MB (2001). Capgras delusion: a window on face recognition. *Trends Cogn Sci.* 5, 149–156.

Engell AD, & Haxby JV. (2007). Facial expression and gaze-direction in human superior temporal sulcus. *Neuropsychologia* 45, 3234–3241.

Engell AD, Haxby JV, & Todorov A. (2007). Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *J Cogn Neurosci.* 9, 1508–1519.

Fink GR, Markowitsch HJ, Reinkemeier M, Bruckbauer T, Kessler J, & Heiss WD. (1996). Cerebral representation of one's own past: neural networks involved in autobiographical memory. *J Neurosci.* 16, 4275–4282.

Fletcher PC, Frith CD, Baker SC, Shallice T, Frackowiak RS, & Dolan RJ. (1995). The mind's eye—precuneus activation in memory-related imagery. *Neuroimage* 2, 195–200.

Frith CD, & Frith U. (1999). Interacting minds—a biological basis. *Science* 286, 1692–1695.

Frith CD, & Frith U. (2006). The neural basis of mentalizing. *Neuron* 50, 531–534.

Gallagher HL, Happe F, Brunswick N, Fletcher PC, Frith U, & Frith CD. (2000). Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia* 38, 11–21.

Gallagher HL, Frith CD. (2003). Functional imaging of "theory of mind." *Trends Cogn Sci.* 7, 77–83.

Gauthier I, Tarr MJ, Anderson AW, Skudlarski P, & Gore JC. (1999). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nat Neurosci.* 2, 568–573.

Gobbini MI, Leibenluft E, Santiago N, & Haxby JV. (2004). Social and emotional attachment in the neural representation of faces. *Neuroimage* 22, 1628–1635.

Gobbini MI, & Haxby JV. (2006). Neural response to the visual familiarity of faces. *Brain Res Bull.* 71, 76–82.

Gobbini MI, & Haxby JV. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia* 45, 32–41.

Gobbini MI, Koralek AC, Bryan RE, Montgomery KJ, & Haxby JV. (2007). Two takes on the social brain: a comparison of theory of mind tasks. *J Cogn Neurosci* 19, 1803–1814.

Gorno-Tempini ML, Price CJ, Josephs O, et al. (1998) The neural systems sustaining face and proper-name processing. *Brain* 121, 2103–2118.

Greene JD, Sommerville RB, Nystrom LE, Darley JM, & Cohen JD. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science* 293, 2105–2108.

Greenwald AG, & Banaji MR. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychol Rev.* 102, 4–27.

Grosbras MH, & Paus T. (2006). Brain networks involved in viewing angry hands or faces. *Cereb Cortex*. 16, 1087–1096.

Hanson SJ, Matsuka T, & Haxby JV. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? *Neuroimage* 1, 156–166.

Hanson SJ, & Halchenko YO. (2007). Brain reading using full brain support vector machines for object recognition: there is no "face" identification area. *Neural Comput*. 20, 1–18.

Haxby JV, Hoffman EA, & Gobbini MI. (2000). The distributed human neural system for face perception. *Trends Cogn Sci*. 46, 223–233.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, & Pietrini P, (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430.

Heatherton TF, Wyland CL, Macrae N, Demos KE, Denny BT, & Kelley WM. (2006). Medial prefrontal activity differentiates self from close others. *Soc Cogn Affect Neurosci*. 1, 18–25.

Henson R, Shallice T, & Dolan R. (2000). Neuroimaging evidence for dissociable forms of repetition priming. *Science* 287, 1269–1272.

Hirstein W, & Ramachandran VS. (1997). Capgras' syndrome: a novel probe for understanding the neural representation of the identity and familiarity of persons. *Proc R Soc Lond B Biol Sci*. 264, 437–44.

Hoffman EA, & Haxby JV. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nat Neurosci*. 3, 80–84.

Idson LC, & Mischel W, (2001). The personality of familiar and significant people: the lay perceiver as a social-cognitive theorist. *J Pers Soc Psychol*. 80, 585–596.

Ishai A, Ungerleider LG, & Haxby JV. (2000). Distributed neural systems for the generation of visual images. *Neuron* 28, 979–990.

Klüver H, & Bucy PC. (1938). An analysis of certain effects of bilateral temporal lobectomy in the rhesus monkey with special reference to "psychic blindness." *J Psychol*. 5, 33–54.

Kosaka H, Omori M, Iidaka T, et al. (2003). Neural substrates participating in acquisition of facial familiarity: an fMRI study. *Neuroimage* 20, 1734–1742.

Leibenluft E, Gobbini MI, Harrison T, & Haxby JV. (2004). Mothers' neural activation in response to pictures of their, and other, children. *Biol Psychiatry*. 56, 225–232.

Leslie AM. (1994). Pretending and believing: issues in the theory of ToMM. *Cognition* 50, 211–238.

Leveroni CL, Seidenberg M, Mayer AR, Mead LA, Binder JR, & Rao SM. (2000). Neural systems underlying the recognition of familiar and newly learned faces. *J Neurosci*. 20, 878–886.

McCabe K, Houser D, Ryan L, Smith V, & Trouard T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proc Natl Acad Sci USA* 98, 11,832–11,835.

McCarthy G, Puce A, Gore JC, & Allison T. (1997). Face-specific processing in the human fusiform gyrus. *J Cogn Neurosci*. 9, 605–610.

Mitchell JP, Heatherton TF, & Macrae CN. (2002). Distinct neural systems subserve person and object knowledge. *Proc Natl Acad Sci USA*, 99, 15,238–15,243.

Montgomery KJ, & Haxby JV. (2008). Mirror neuron system differentially activated by facial expressions and social hand gestures: a functional magnetic resonance imaging study. *J Cogn Neurosci*. 20, 1866–1877.

Morris JS, Frith CD, Perrett DI, et al. (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature* 383, 812–815.

Nakamura K, Kawashima R, Sato N, et al. (2000). Functional delineation of the human occipito-temporal areas related to face and scene processing. A PET study. *Brain* 123, 1903–1912.

Nakamura K, Kawashima R, Sugiura M, et al. (2001). Neural substrates for recognition of familiar voices: a PET study. *Neuropsychologia* 39, 1047–1054.

Phillips ML, Drevets WC, Rauch SL, & Lane R (2003). Neurobiology of emotion perception I: the neural basis of normal emotion perception. *Biol Psychiatry*. 54, 504–514.

Puce A, Allison T, Bentin S, Gore JC, & McCarthy G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *J Neurosci* 18, 2188–2199.

Puce A, Allison T, & McCarthy G. (1999). Electrophysiological studies of human face perception. III: Effects of top–down processing on face-specific potentials. *Cereb Cortex*. 9, 445–958.

Puce A, & Perrett D. (2003). Electrophysiology and brain imaging of biological motion. *Philos Trans R Soc Lond B Biol Sci*. 358, 435–445.

Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, & Cohen JD. (2004). The neural correlates of theory of mind within interpersonal interactions. *Neuroimage* 22, 1694–1703.

Rossion B, Schiltz C, Robaye L, Pirenne D, & Crommelinck M. (2001). How does the brain discriminate familiar and unfamiliar faces?: a PET study of face categorical perception. *J Cogn Neurosci.* 13, 1019–1034.

Rotshtein P, Henson RN, Treves A, Driver J, & Dolan RJ, (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nat Neurosci.* 1, 107–113.

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, & Cohen JD. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science* 300, 1755–1758.

Saxe R, & Kanwisher N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage* 19, 1835–1842.

Saxe R. (2006). Uniquely human social cognition. *Curr Opin Neurobiol.* 16, 235–239.

Schwartz CE, Wright CI, Shin LM, et al. (2003). Differential amygdalar response to novel versus newly familiar neutral faces: a functional MRI probe developed for studying inhibited temperament. *Biol Psychiatry.* 53, 854–862.

Sergent J, Ohta S, & MacDonald B. (1992). Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain* 115, 15–36.

Singer T, Seymour B, O'Doherty J, Kaube H, Dolan RJ, & Frith CD. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science* 303, 1157–1162.

Sugiura M, Shah NJ, Zilles K, & Fink GR. (2005). Cortical representations of personally familiar objects and places: functional organization of the human posterior cingulate cortex. *J Cogn Neurosci.* 17, 183–198.

Tranel D, & Damasio AR. (1985) Knowledge without awareness: an autonomic index of facial recognition by prosopagnosics. *Science* 228, 1453–1454.

Todorov A, & Uleman JS. (2002). Spontaneous trait inferences are bound to actors' faces: evidence from a false recognition paradigm. *J Pers Soc Psychol.* 83, 1051–1065.

Todorov A, Gobbini MI, Evans KK, & Haxby JV, (2007). Spontaneous retrieval of affective person knowledge in face perception. *Neuropsychologia* 45, 163–173.

Winston JS, Strange BA, O'Doherty J, & Dolan RJ. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nat Neurosci.* 5, 277–283.

Winston JW, Henson RNA, Fine-Goulden MR, & Dolan RJ. (2004). fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. *J Neurophysiol.* 92, 1830–1839.

Zalla T, Koechlin E, Pietrini P, et al. (2000). Differential amygdala responses to winning and losing: a functional magnetic resonance imaging study in humans. *Eur J Neurosci.* 12, 1764–1770.

CHAPTER 4
Evaluating Faces on Social Dimensions

Alexander Todorov

“We look at a person and immediately a certain impression of his character forms itself in us. A glance, a few spoken words are sufficient to tell us a story about a highly complex matter. We know that such impressions form with remarkable rapidity and with great ease. Subsequent observations may enrich or upset our view, but we can no more prevent its rapid growth than we can avoid perceiving a given visual object or hearing a melody” (Asch, 1948, p. 258).

Solomon Asch wrote these words 60 years ago. Since then, social psychologists have amassed evidence supporting his insights. People, indeed, are remarkably good at forming impressions of other people. First, as Asch noted, these impressions are formed from minimal information. They can originate in facial appearance (e.g., Bar, Neta, & Linz, 2006; Olson & Marshuetz, 2005; Willis & Todorov, 2006; Zebrowitz, 1999), “thin slices” of nonverbal behaviors (e.g., Albright, Kenny, & Malloy, 1988; Ambady, Hallahan, & Rosenthal, 1995; Ambady & Rosenthal, 1992), or behavioral information (e.g., Carlston & Skowronski, 1994; Todorov & Uleman, 2002, 2003, 2004; Uleman, Newman, & Moskowitz, 1996). Second, these impressions are formed rapidly and efficiently (Bar et al., 2006; Olson & Marshuetz, 2005; Willis & Todorov, 2006; Todorov, Pakrashi, & Oosterhof, 2009; Todorov & Uleman, 2003). For example, 33-millisecond exposure to a face is sufficient for people to make a trustworthiness judgment (Todorov et al., 2009). Third, these impressions are formed spontaneously and when cognitive resources are severely limited (Uleman, Blader, & Todorov, 2005). For example, even when people are engaged in a meaningless task of counting nouns while reading behavioral information, they form person impressions (Todorov & Uleman, 2003).

We have not moved beyond the insights of Asch in any fundamental way, but we have moved closer to understanding the cognitive processes underlying impression formation and their neural basis. As the present book testifies, person perception questions are addressed by a variety of novel methods such as fMRI (Chapters 1 and 3), event-related potentials (ERPs; Chapter 6), and the study of patients with brain lesions (Chapter 11).

In this chapter, I focus on how people form person impressions from facial appearance. Faces are a particularly rich source of social information. People use dynamic changes in the face, such as expression of emotions, to understand the immediate meaning of the situation and invariant facial features to identify other people (Haxby, Hoffman, & Gobbini, 2000). Faces are also a rich source of person inferences, although the accuracy of these inferences is dubious (Hassin & Trope, 2000; Olivola & Todorov, 2010; Todorov, 2008). Nevertheless, such person inferences predict important social outcomes (Hamermesh & Biddle, 1994; Hassin & Trope, 2000; Langlois et al., 2000; Montepare & Zebrowitz, 1998;

Zebrowitz, 1999), ranging from electoral success (Hall, Goren, Chaiken, & Todorov, 2009; Little, Burriss, Jones, & Roberts, 2007) to sentencing decisions (Blair, Judd, & Chapleau, 2004; Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006; Zebrowitz & McDonald, 1991). For example, inferences of competence from faces predict electoral success (Ballew & Todorov, 2007; Todorov, Mandisodza, Goren, & Hall, 2005) and inferences of dominance predict military rank attainment (Mazur, Mazur, & Keating, 1984; Mueller & Mazur, 1996).

Research on face evaluation or how people make personality inferences from facial appearance is situated within the existing cognitive neuroscience models of face perception in the section Cognitive Neuroscience Research on Face Perception. In the sections A Dimensional Model of Evaluation of Emotionally Neutral Faces and Computer Modeling of Face Trustworthiness and Face Dominance, I outline a model of face evaluation on social dimensions. According to this model (Oosterhof & Todorov, 2008), faces are evaluated on two primary, independent dimensions: valence and dominance. Evaluation on specific trait dimensions can be derived from the combination of these two dimensions. Consistent with theories that posit that evaluation of emotionally neutral faces is constructed from facial cues that have evolutionary significance (Zebrowitz, 2004; Zebrowitz & Montepare, 2006, 2008; Zebrowitz et al., 2003), I argue that face evaluation is an overgeneralization of adaptive mechanisms for guiding appropriate social behavior. Specifically, valence evaluation of faces is based on facial cues resembling emotional expressions signaling whether the person should be avoided or approached. Dominance evaluation is based on facial cues signaling the physical strength of the person. Functionally, these types of evaluation correspond to inferences about harmful intentions and the ability to cause harm (cf., Fiske et al., 2007). In section Emotion Overgeneralization Mechanisms, I present additional evidence for the hypothesis that evaluation of faces is an overgeneralization of perception of emotional expressions, and in section The Role of the Amygdala in Evaluation of Emotionally Neutral Faces, I review evidence for the involvement of the amygdala in valence evaluation of faces. Although people make rapid judgments from faces, they also change their minds in light of new person information. In section Beyond Facial Appearance: Impressions from Behaviors, I review evidence for the role of person knowledge in face perception, and in section Conclusions and Outstanding Questions, I conclude with a sample of outstanding questions for research on face evaluation.

Cognitive neuroscience research on face perception

Although there is a large body of cognitive neuroscience research on face perception, almost all of the studies in this tradition focus either on recognition of faces (e.g., Haxby et al., 1999, 2001; Kanwisher et al. 1997; McCarthy et al., 1997) or recognition of expressions of emotions (e.g., Adolphs, 2002, 2003; Calder et al., 2000, 2001; Phan et al., 2002; Morris et al., 1996; Philips et al., 1997). This research has led to great advances in the understanding of face perception. For example, functional imaging studies have shown that whereas areas in the fusiform gyrus are more responsive to facial identity information (Haxby et al., 1999, 2001; Kanwisher et al. 1997; McCarthy et al., 1997), areas in the superior temporal sulcus (STS) are more responsive to expression information (Allison et al., 2000; Hoffman & Haxby, 2000; Puce et al., 1998). Building on existing cognitive models of face processing (e.g., Bruce & Young, 1986), single-cell recording, and functional imaging data, a model of the neural system underlying face perception has been developed to capture the differences between processing of facial identity and emotional expressions (Haxby et al., 2000, 2002; *see also* Chapter 3).

According to this model, the major distinction in face processing is between invariant facial features and dynamic changes such as eye gaze and expressions. Whereas invariant facial features are critical for person recognition, dynamic facial changes communicate the mental states of others. This model can account for observed dissociations between processing of facial identity and emotional expressions (but *see* Calder & Young, 2005, for an alternative view). For example, there are prosopagnosics who, despite their inability to recognize faces, show normal perception of emotional expressions (Bentin et al., 2007; Damasio, Tranel, & Damasio, 1990; Duchaine, Parker, & Nakayama, 2003; Humphreys, Avidan, & Behrmann, 2007; Tranel, Damasio, & Damasio, 1988).

However, it is not clear how person inferences such as trustworthiness and competence fit in this distinction. Although such inferences are based on invariant facial features, expressions of emotions affect trait judgments (Knutson, 1996; Krumhuber et al., 2007), and it is possible to observe dissociations between processing of facial identity and impression formation. For example, with Brad Duchaine, we studied four developmental prosopagnosics with severe impairments in both memory for and perception of facial identity (Todorov & Duchaine, 2008). Despite this impairment, their judgments of face trustworthiness across three different face sets were within the normal range of control judgments, and the performance of two of the prosopagnosics was typical. This dissociation suggests that different mechanisms may underlie processing of facial identity and impression formation.

Because of the focus on recognition of faces and emotional expressions, there has been little cognitive neuroscience research on how faces are evaluated on social dimensions. Social cognition research has also largely ignored this topic despite the evidence for the importance of such evaluations. As Macrae and his colleagues have noted: "Although the human face conveys a wealth of potential information, social-cognitive research has focused almost exclusively on identifying the conditions under which categorical knowledge (i.e., stereotypes) is activated in response to available stimulus cues." (Macrae et al., 2005, p. 686).

The one social dimension of face evaluation that has been studied is face trustworthiness. Adolphs, Tranel, and Damasio (1998) showed that patients with bilateral amygdala damage show impaired discrimination between trustworthy- and untrustworthy-looking faces. Subsequent fMRI studies with normal individuals confirmed the involvement of amygdala in evaluation of faces on trustworthiness (Engell, Haxby, & Todorov, 2007; Said, Baron, & Todorov, 2009; Todorov, Baron, & Oosterhof, 2008; Winston, Strange, O'Doherty, & Dolan, 2002). But what do judgments of trustworthiness measure? As argued by Engell et al. (2007) and as I show below, the reason that the amygdala responds to face trustworthiness is that trustworthiness judgments are a good approximation of the general valence evaluation of faces (Todorov & Engell, 2008).

A Dimensional Model of Evaluation of Emotionally Neutral Faces

Although people engage in a variety of trait judgments from faces (e.g., Willis & Todorov, 2006), these judgments are highly correlated with each other. For example, for a set of standardized faces (Lundqvist et al., 1998) used in our research, judgments of trustworthiness correlated 0.83 with judgments of emotional stability, 0.75 with judgments of attractiveness, –0.76 with judgments of aggressiveness, and 0.63 with judgments of intelligence. Given the high correlations among judgments of different traits, it is almost impossible to identify *(1)* neural correlates specific to a trait dimension and *(2)* facial configurations that vary only along this dimension. For example, if the goal is to model the neural responses to faces as a function of multiple trait judgments, the high correlations among judgments introduce serious collinearity problems.

Instead of working with specific trait dimensions, we have undertaken an approach of reducing judgments on multiple trait dimensions to a few orthogonal dimensions that can account for these judgments (Oosterhof & Todorov, 2008; Todorov, 2008). In a data-driven approach, we first identified trait dimensions on which faces were spontaneously evaluated. Second, we collected judgments on these trait dimensions. Finally, we submitted these judgments to a principal components analysis (PCA) to identify the underlying dimensions of face evaluation.

The first principal component (PC) accounted for 63.3% of the variance of the mean trait judgments. All positive judgments (e.g., attractive, responsible) had positive loadings and all negative judgments (e.g., aggressive) had negative loadings on this component, suggesting that it can be interpreted as valence evaluation (Kim & Rosenberg, 1980; Rosenberg et al., 1968; cf. Osgood et al., 1957). The second PC accounted for 18.3% of the variance. Judgments of dominance, aggressiveness, and confidence had the highest loading on this component, suggesting that it can be interpreted as dominance evaluation. This two-dimensional structure of face evaluation is consistent with well-established dimensional models of social perception (Fiske et al., 2007; Wiggins, 1979; Wiggins et al., 1989). For example, starting with a large set of traits describing interpersonal relationships, Wiggins and colleagues have shown that these traits can be represented within a two-dimensional space defined by affiliation and dominance, dimensions that are similar to the dimensions identified in our research.

The PCA also showed that the valence and dominance dimensions can be approximated by single trait judgments (Fig. 4–1). Specifically, judgments of trustworthiness had the highest loading (0.94) on the first PC and were practically uncorrelated with the second PC (–0.06). Judgments of dominance had the highest loading (0.93) on the second PC and the lowest

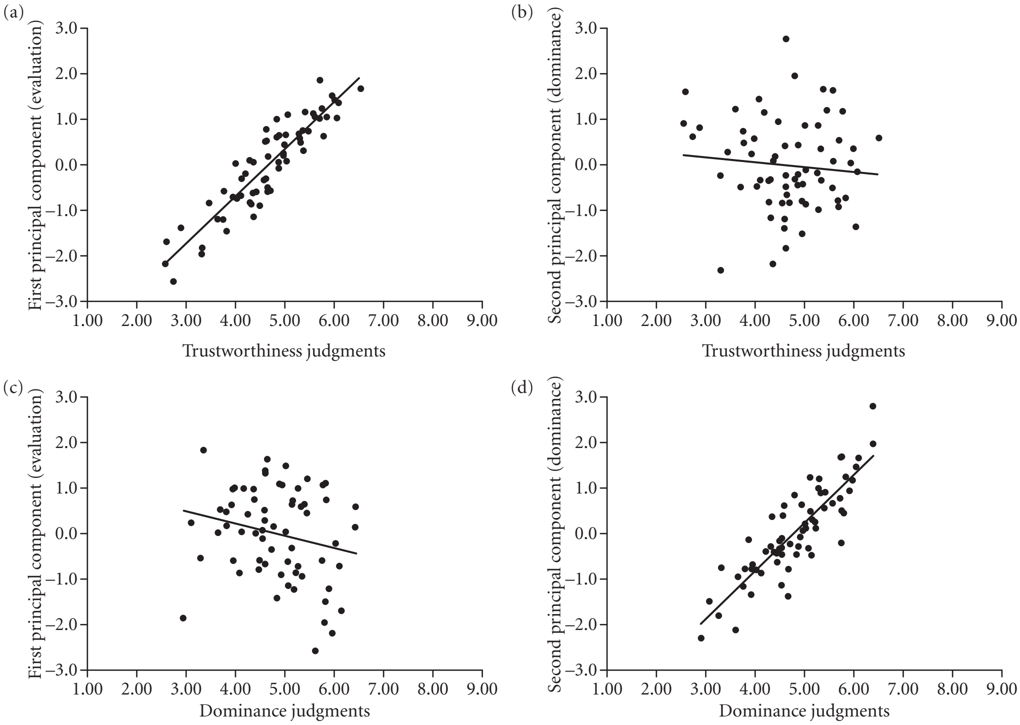


Fig. 4–1 Scatter plots of trustworthiness and dominance judgments from emotionally neutral faces and the first two principal components derived from a principal components analysis of judgments on 11 traits (other than trustworthiness and dominance) used to spontaneously characterize faces. Trustworthiness judgments and (a) the first valence component, and (b) the second dominance component. Dominance judgments and (c) the first valence component, and (d) the second dominance component. Each point is a face. The judgments were measured on a 9-point scale, ranging from 1 (not at all [trustworthy or dominant]) to 9 (extremely [trustworthy or dominant]). The lines represent the best linear fit.

loading on the first PC (–0.24). This was the case even when the principal components were obtained from an analysis excluding these two judgments to avoid biasing the PCA solution. As shown in Figures 4–1a and 4–1b, trustworthiness judgments were highly correlated with the first PC but not with the second PC. In contrast, as shown in Figures 4–1c and 4–1d, dominance judgments were highly correlated with the second PC but not with the first PC. Additional analyses showed that the two-dimensional solution is robust with respect to the set of traits used to estimate the PCs and the face stimuli (Oosterhof & Todorov, 2008).

Computer modeling of face trustworthiness and face dominance

Given the findings that judgments of trustworthiness and dominance can be used as approximations of the underlying dimensions—valence and dominance—of face evaluation, we built computer models for representing how faces vary on trustworthiness and dominance. To build trustworthiness and dominance dimensions, we used a data-driven statistical model of face representation, in which faces were represented as points in a multidimensional space (Blanz & Vetter, 1999, 2003; Singular Inversions, 2006). The input to this model was a database of faces that were laser-scanned in 3D. The shape of a 3D face was represented by the vertex positions (points in 3D Euclidian space) of a polygonal model of fixed mesh topology. Finally, using PCA, the representation of each face was reduced to a limited number of independent components. We worked with 50 dimensions (50 independent principal components) representing 3D face shape.

Using the face model, we randomly generated emotionally neutral faces. We used only White faces to avoid the influence of stereotypes on trait judgments. We asked participants to judge these faces on trustworthiness and dominance and used the mean trustworthiness and dominance judgments to find vectors (representing a weighted combination of the 50 principal components) in the 50-dimensional face space whose direction was optimal in changing trustworthiness and dominance, respectively. Specifically, these vectors were based on the best linear fit of the mean judgments as a function of the 50 shape components. Finally, to obtain an orthogonal solution (Fig. 4–2), we rotated the dominance vector to make it orthogonal to the trustworthiness vector (Oosterhof & Todorov, 2008).

To validate the computer models, first, we randomly generated faces. Second, for each face we created several versions that varied along the respective dimensions and, then, asked participants to rate the faces on these dimensions. These studies showed that the models of trustworthiness and dominance successfully manipulated trustworthiness and dominance of faces. Trustworthiness and dominance judgments of faces generated by the models tracked the

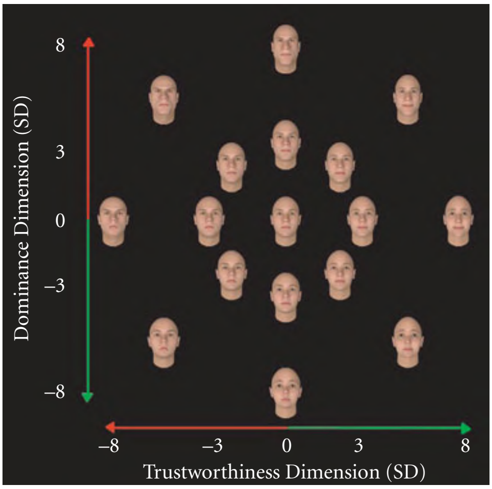


Fig. 4–2 A two-dimensional model of face evaluation. Examples of a face with exaggerated features on the two orthogonal dimensions—trustworthiness and dominance—of face evaluation. The changes in features were implemented in a computer model based on trustworthiness and dominance judgments of 300 emotionally neutral faces. The threat dimension shown on the diagonal from the 4th to the 2nd quadrant was obtained by rotating the trustworthiness dimension 45° clockwise and the dominance dimension 45° counterclockwise in the plane defined by the two dimensions. This threat dimension was practically identical to a dimension based on threat judgments of faces. The extent of face exaggeration is presented in SD units.

trustworthiness and dominance predicted by the model, respectively. Interestingly, whereas dominance judgments of faces generated by the dominance dimension were related in a linear fashion to the face dominance, trustworthiness judgments of faces generated by the trustworthiness dimension were related in a quadratic fashion to face trustworthiness. Specifically, people were more sensitive to changes at the negative end than at the positive end of the trustworthiness dimension.

The validation studies exemplify some of the advantages of using formal models of how faces vary on social dimensions (Todorov, 2008). First, these models can generate an unlimited number of faces that vary on the dimension of interest. Second, the variation of faces can be manipulated precisely (e.g., a face that is 3 SD above the center of the dimension vs. a face that is 3 SD below this center) and the range of differences maximized to detect subtle effects. For example, previous studies have failed to find that trait judgments are made after subliminal exposures to faces (Bar et al., 2006; Todorov et al., Exp. 2, 2009). However, the stimuli may not have been sufficiently different on the trait dimension of interest. We used faces generated by the trustworthiness dimension to test for subliminal effects. Untrustworthy (-3 SD) and trustworthy versions (3 SD) of faces were presented for 20 milliseconds and immediately masked by the neutral version of the faces (0 SD). Trustworthiness judgments of the neutral faces were more negative when these faces were preceded by untrustworthy than by trustworthy faces (Todorov et al., Exp. 3, 2009), although the recognition of the primes was at chance in a forced choice recognition task. These findings suggest that people can extract information for social judgments even when the faces are presented below their level of subjective awareness.

The third and probably most important advantage of computer models is that these models can be used as a discovery tool to identify the variations in facial cues that produce specific judgments. Although these models are holistic in the sense that they are not constrained by any set of facial features, they can be used to discover the important features *a posteriori*. By exaggerating the features specific to an evaluative dimension, we can identify the type of facial information used for this evaluation. For example, as shown in Figure 4–2, whereas faces at the negative extreme of the trustworthiness dimension (–8 SD) were no longer neutral and looked angry, faces at the positive extreme (8 SD) looked happy. Whereas faces at the negative extreme of the dominance dimension (–8 SD) looked extremely feminine, faces at the positive extreme (8 SD) looked extremely masculine.

Subsequent experiments confirmed that the two dimensions are sensitive to different types of facial information. As in the model validation studies, we randomly generated faces and created extreme versions of the faces on the trustworthiness and dominance dimensions. First, in a study in which participants were asked to categorize these faces as neutral or as expressing one of the six basic emotions, participants classified extremely exaggerated faces in the negative direction on the trustworthiness dimension (–8 SD, *see* Fig. 4–2) as angry and extremely exaggerated faces in the positive direction (4 and 8 SD) as happy. Although there were fewer emotion categorizations of faces that varied on the dominance dimension, partly because of the fact that we rotated this dimension to make it orthogonal to the trustworthiness dimension, as the faces became more exaggerated in the dominance direction, they were more likely to be classified as angry; and as the faces become exaggerated in the submissiveness direction, they were more likely to be classified as fearful (*see* Supporting Information Table 7 in Oosterhof & Todorov, 2008). The original dominance dimension based on dominance judgments was negatively correlated with the trustworthiness dimension and would have been even more sensitive to features resembling emotional expressions, although to a lesser extent than the trustworthiness dimension.

Second, in additional nine studies (five of them reported in Oosterhof & Todorov, 2008), participants were asked to rate the faces on continuous scales on angry/ happy, baby-faced/ mature-faced, and feminine/masculine. We also manipulated the face information available for the judgments. In three of the studies, participants rated the intact faces, in three studies, they rated the faces with their external features masked, and in three they rated the faces with their internal features masked. As shown in Figure 4–3, these studies showed that whereas the trustworthiness dimension was more sensitive to features resembling happy and angry expressions, the dominance dimension was more sensitive to features signaling physical strength. In particular, most of the diagnostic information for the trustworthiness dimension was present in the internal features of the face, whereas most of the diagnostic information for the dominance dimension was present in the external (shape) features of the face.

In principle, the two-dimensional model can represent any social judgment from faces, as we have illustrated with judgments of threat (Oosterhof & Todorov, 2008). Threat judgments are particularly important from a survival point of view (Bar et al., 2006), and these judgments are highly correlated with both trustworthiness and dominance judgments. Threatening faces

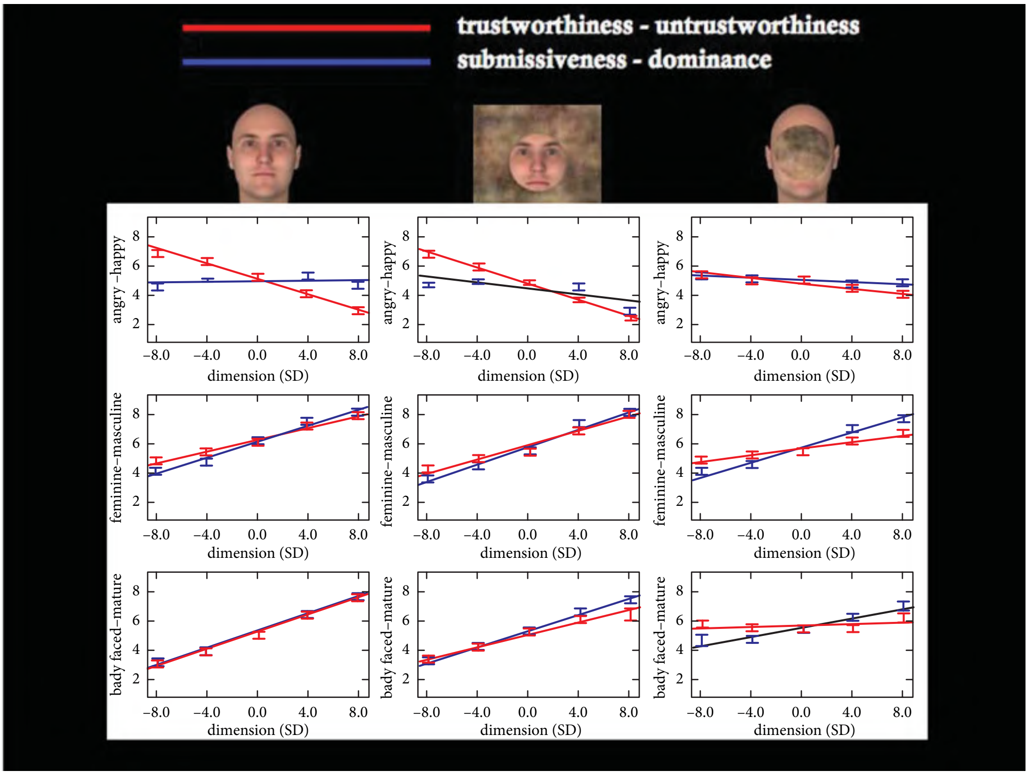


Fig. 4–3 Plots of changes in judgments of expressions of anger/happiness, femininity/masculinity, and facial maturity as a function of trustworthiness and dominance of faces. Expression judgments (1st row) were made on a 9-point scale, ranging from 1 (angry) to 5 (neutral) to 9 (happy). Femininity/masculinity judgments (2nd row) were made on a 9-point scale, ranging from 1 (feminine) to 5 (neutral) to 9 (masculine). Facial maturity judgments (3rd row) were made on a 9-point scale, ranging from 1 (baby-faced) to 5 (neutral) to 9 (mature-faced). Participants made judgments from intact faces (1st column), faces with masked external features (2nd column), and faces with masked internal features (3rd column). Error bars show standard error of the mean. The lines represent the best linear fit. The *x*-axis in the figures represents the extent of face exaggeration in SD units. The direction of the trustworthiness dimension was reversed for these figures to show that the slopes for the change from trustworthy to untrustworthy faces and the change from submissive to dominant faces were similar for facial maturity and femininity/masculinity judgments of intact faces and faces with masked external features.

are both untrustworthy- and dominant-looking. We built a threat dimension in the space defined by the trustworthiness and dominance dimensions by giving equal weights to these two dimensions (1 and –1 for dominance and trustworthiness, respectively; the diagonal in Fig. 4–2 from the 4th to the 2nd quadrant). This dimension was practically identical to a threat dimension based on threat judgments.

Emotion overgeneralization mechanisms

The computer modeling findings suggest that trait judgments are constructed from cues that have evolutionary significance (Zebrowitz, 2004; Zebrowitz & Montepare, 2006, 2008). The primary valence dimension of face evaluation derives from cues resembling expressions of anger and happiness. As Fridlund (1994) has argued, one of the functions of emotional expressions is to signal behavioral intentions. For example, whereas expressions of happiness signal to the perceiver that the person can be approached, expressions of anger signal that the person should be avoided, and there is evidence that angry faces trigger automatic avoidance responses (Adams et al., 2006; Marsh et al., 2005). Thus, consistent with social cognition research suggesting that the valence evaluation of stimuli is directly linked to approach/avoidance behaviors (Chen & Bargh, 1999), the valence evaluation of faces may amount to an approach/avoidance decision. From an evolutionary point of view, the costs of approaching an angry individual are greater than avoiding a happy individual, and this can explain the nonlinearity of trustworthiness judgments. As described above, these judgments were more sensitive to changes at the negative than at the positive end of the trustworthiness dimension. Similarly, threat judgments of faces generated by the threat dimension were more sensitive to changes at the threatening than the nonthreatening end of the dimension.

Consistent with the emotion overgeneralization hypothesis—namely, that similarity of facial features to emotional expressions is attributed to personality traits—previous studies have shown that emotional expressions affect trait judgments from faces (Hess et al., 2000; Knutson, 1996; Montepare & Dobish, 2003). For example, smiling faces are perceived as more trustworthy than neutral faces (Krumhuber et al., 2007) and higher on affiliation, an attribute similar to trustworthiness (Knutson, 1996; Montepare & Dobish, 2003). Moreover, judgments of anger and happiness from emotionally neutral faces are correlated with judgments of trustworthiness (Todorov & Duchaine, 2008) and judgments of affiliation (Montepare & Dobish, 2003).

To provide an extended replication of these findings, we collected emotion judgments of the emotionally neutral faces for which we had already collected judgments on trait dimensions (*see* section A Dimensional Model of Evaluation of Emotionally Neutral Faces). As shown in Figure 4–4a, 55 of the 84 (14 trait × 6 emotion judgments: anger, disgust, fear, sadness, surprise, and happiness) correlations were significant (Said, Sebe, & Todorov, 2009). For example, judgments of happiness were positively correlated with all positive trait judgments and negatively correlated with all negative judgments. The pattern was reversed for judgments of anger. The valence component derived from the trait judgments was strongly correlated with judgments of happiness and anger, moderately correlated with judgments of disgust, and weakly correlated with judgments of sadness (Fig. 4–4b). Faces that were evaluated positively were perceived as happier and more surprised but less angry, less disgusted, less fearful, and less sad. The dominance component was correlated with judgments of anger, surprise, sadness, and fear (Fig. 4–4b). Dominant faces were perceived as angrier, less sad, less fearful, and less surprised than submissive faces.

Although these findings are consistent with the emotion overgeneralization hypothesis, they are also consistent with the hypothesis that these correlations can be accounted for by common semantic properties of emotion and trait judgments rather than by perceptual similarity. For example, expectations about the relation between emotional states (e.g., smiling as an expression of happiness) and personality

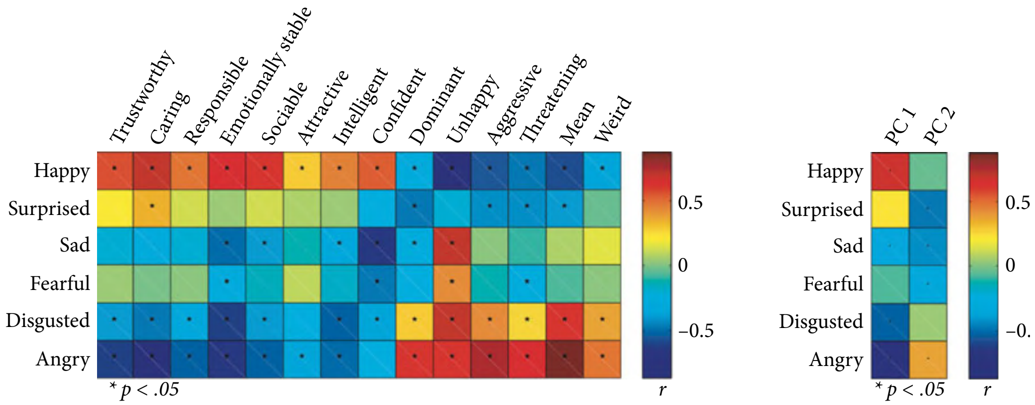


Fig. 4–4 Color maps of correlations between trait and emotion judgments of emotionally neutral faces (a). Traits are ordered by their loadings on the first principal component (PC)—valence—derived from a principal components analysis (PCA) of the trait judgments (*see* the section A Dimensional Model of Evaluation of Emotionally Neutral Faces). Correlations between emotion judgments and the first two PCs—valence and dominance—derived from a PCA of the trait judgments (b).

traits (e.g., sociable) may lead to strong associations between emotion and trait judgments. This hypothesis is consistent with research on implicit personality theory that shows that people hold assumptions about the relationships between various traits (Bruner & Tagiuri, 1954; Cronbach, 1955; Schneider, 1973). In fact, the dimensional structure of the emotion judgments was very similar to the dimensional structure of the trait judgments. The first PC derived from a PCA of the emotion judgments was highly correlated with the valence component of the trait judgments and uncorrelated with the dominance component. In contrast, the second PC of the emotion judgments was more strongly correlated with the dominance component than with the valence component.

To rule out the possibility that the relations between trait and emotion judgments can be accounted entirely for by semantic similarities, we used an emotion classifier to categorize the emotionally neutral faces (Said et al., 2009). Specifically, we used a Bayesian network classifier to detect the subtle presence of features resembling emotions in the faces. The classifier accepts as input a feature vector containing the displacements between automatically chosen landmarks and the same landmarks of a prototypical neutral face and outputs a set of probabilities corresponding to each basic emotion. Because we applied the classifier to neutral faces, the output probabilities were very low. Nevertheless, these probabilities predicted trait judgments from the faces.

The pattern of correlations was similar to the pattern of correlations for emotion and trait judgments, although the correlations were weaker (27 of the 84 probabilities—trait judgments correlations were significant). The probability of classifying faces as happy was positively correlated with all positive trait judgments and negatively correlated with all negative judgments. The probability of classifying faces as angry was positively correlated with judgments of aggressiveness, meanness, unhappiness, and dominance. The valence component was positively correlated with the classifier probabilities of happiness and negatively correlated with the probabilities of anger, disgust, and fear, although only the correlation for happiness reached significance. The dominance component was positively correlated with the classifier probabilities of anger and negatively correlated with the probabilities of surprise and fear.

Although the methods used in Oosterhof and Todorov (2008) and in Said et al. (2009) differed in a number of ways, they converged on similar solutions. Faces with positive valence (trustworthiness) were more likely to be classified as happy and less likely to be classified as angry and disgusted than faces with negative valence. Highly dominant faces were more likely to be classified as angry and less likely to be classified as fearful than highly submissive faces.

To the extent that structural facial features signaling positive valence or trustworthiness are similar to expressions of anger and happiness, it should also be possible to demonstrate that facial features affect the perception of emotional expressions. To test this hypothesis, based on prior trustworthiness judgments, we selected trustworthy and untrustworthy faces and created dynamic stimuli in which the faces expressed either happiness or anger (Oosterhof & Todorov, 2009). Although we added the same amount of emotional intensity to faces, trustworthy faces expressing happiness were perceived as happier than untrustworthy faces. In contrast, untrustworthy faces expressing anger were perceived as angrier than trustworthy faces expressing the same emotion.

We also manipulated changes in trustworthiness during the course of the animation. For example, in incongruent animations, an untrustworthy (or a trustworthy) face gradually morphed into a trustworthy (or an untrustworthy) face. To the extent that trait judgments are an overgeneralization of cues resembling expressions, changes that are in the direction of the expressed emotion (e.g., untrustworthy-to-trustworthy and happiness) should amplify the intensity of the perceived emotion. In contrast, changes in the opposite direction (e.g., untrustworthy-to-trustworthy and anger) should dampen this intensity. As shown in Figure 4–5, this is exactly what we found. For example, when a trustworthy face changed into an untrustworthy face, the same angry expression was perceived as angrier than when an untrustworthy face changed into another untrustworthy face or when there was no change in the identity of the face. Similarly, when an untrustworthy face changed into a trustworthy face, the same angry expression was perceived as less angry than when a trustworthy face changed into another trustworthy face or when there was no change in the identity of the face.

To test for similarities in the neural codes of perceived trustworthiness and expressions of anger and happiness, we used a behavioral adaptation paradigm (Engell, Todorov, & Haxby, in press). The adaptation paradigm has been used to investigate other dimensions of the neural

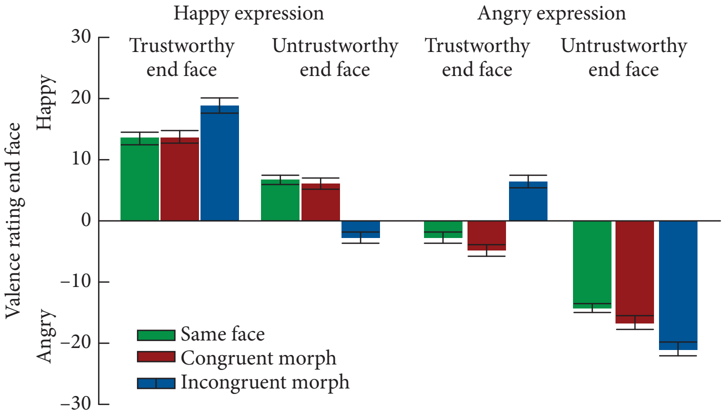


Fig. 4–5 Valence ratings of emotional expressions as a function of the type of emotion, the trustworthiness of the face, and the morphing condition: same face with no change in identity; congruent morph with no change in face trustworthiness but change in identity; and incongruent morph with changes in both face trustworthiness and identity. The ratings were made on a continuous slider ranging from angry to neutral to happy. The error bars show standard errors of the means.

representation of faces, including viewpoint invariance, gender, attractiveness, and expression (e.g., Fox & Barton, 2007; Jeffrey et al., 2006; Rhodes et al., 2006; Webster et al., 1999, 2004). The central tenet of this paradigm is that extended exposure to a stimulus dimension results in fatigue of the neural population that represents the stimulus. Thus, subsequent exposure to a stimulus along the same dimension should result in a perceptual shift away from the adapting stimulus. For example, Webster and colleagues (2004) showed that androgynous faces (i.e., faces that had an equal probability of being categorized by participants as "male" or "female") were seen as distinctly "male" after extended exposure to female faces and as distinctly "female" after extended exposure to male faces.

If trustworthiness evaluation is an overgeneralization of perceiving features resembling angry and happy facial expressions, then we should be able to influence this evaluation by first adapting the neural populations that support the perception of those expressions. In the pre-adaptation stage of the experiment, participants rated the trustworthiness of faces. After the pre-adaptation stage, participants were randomly assigned to one of three adapting conditions: passive viewing of angry, fearful, or happy expressions for 66 seconds. After the adaptation, participants rated the trustworthiness of faces again. The test faces were reduced in size to 80% of the size of the adapter faces to disrupt any low-level adaptation effects. As expected, adaptation to angry faces resulted in higher trustworthiness ratings, whereas adaptation to happy faces resulted in lower trustworthiness ratings. In the control condition of adaptation to fearful faces, trustworthiness ratings were not influenced.

To conclude, several lines of behavioral research provide convergent evidence that trait inferences from emotionally neutral faces are based on resemblance of facial features to emotional expressions. In particular, the evidence suggests that the primary, valence dimension of face evaluation is derived from similarity of facial features to expressions of anger and happiness, expressions that signal potential behavioral intentions.

The role of the amygdala in evaluation of emotionally neutral faces

As noted earlier, the amygdala, a subcortical brain region critical for evaluation of novel stimuli, fear conditioning, and consolidation of emotional memories (Amaral, 2002; Davis & Whalen, 2001; Phelps & LeDoux, 2005; Vuilleumier, 2005), has been implicated in the evaluation of face trustworthiness (Adolphs et al., 1998; Engell et al., 2007; Todorov et al., 2008; Winston et al., 2002). Following the Adolphs et al. (1998) findings from patients with bilateral amygdala damage, in an fMRI study with normal participants, Winston and colleagues showed that the amygdala's response to faces increased as their subjectively perceived trustworthiness decreased (Winston et al., 2002). This was the case independent of whether the evaluation task was explicit (judging the trustworthiness of faces) or implicit (judging the age of faces).

We replicated Winston et al.'s findings, using a single implicit task to rule out the possibility that the performance on implicit evaluation trials was influenced by prior performance on explicit evaluation trials (Engell et al., 2007). Participants were presented with a series of faces in an ostensibly memory task and asked after each block of 11 faces to indicate whether a test face was presented in the block. Although this task did not demand explicit evaluation of faces, as in Winston et al. (2002), the amygdala response to faces increased as their trustworthiness decreased.

We also tested whether the amygdala's response to face trustworthiness was driven by structural properties of the face that signaled untrustworthiness across observers or by idiosyncratic components of trustworthiness judgments. The amygdala's response to faces was better predicted by consensus judgments of trustworthiness—aggregated across a large number of participants separate from the fMRI participants—than by the fMRI participants' individual judgments. When the analysis controlled for the shared variance of individual and consensus judgments, there was little residual variance accounted for by individual judgments in the amygdala.

In a subsequent study, we first built a computer model of face trustworthiness based on behavioral judgments (this work preceded Oosterhof & Todorov, 2008). Second, we generated novel faces based on this model. Third, we used these novel faces in an fMRI study, using the same implicit task as in Engell et al. (2007). As in the previous studies, we found that the right amygdala's response to faces increased as their trustworthiness decreased (Todorov et al., 2008). However, we also found a nonlinear response in the left amygdala so that extremely trustworthy faces evoked a stronger response than faces at the middle of the dimension. This finding is discussed at the end of the section.

Across three different studies, the response to faces in the amygdala was linearly related to judgments of face trustworthiness. However, as described in section A Dimensional Model of Evaluation of Emotionally Neutral Faces, trustworthiness judgments are highly correlated with other social judgments and approximate the valence evaluation of emotionally neutral faces (Fig. 4–1). Because we used the same set of faces in one of our prior fMRI studies (Engell et al., 2007) as in the behavioral studies in which we collected trait judgments (*see* section A Dimensional Model of Evaluation of Emotionally Neutral Faces), we were able to test the hypothesis that the amygdala is involved in general valence evaluation of emotionally neutral faces rather than in evaluation of faces on specific trait dimensions (Todorov & Engell, 2008). According to this hypothesis, face variations on any social dimension (e.g., trustworthiness, attractiveness, aggressiveness) should engage the amygdala to the extent that this dimension has a valence content. In other words, variations on dimensions with clear valence connotations (e.g., trustworthiness and meanness) should engage the amygdala more strongly than variations on dimensions with less clear valence connotations (e.g., dominance).

To test the valence hypothesis, we derived the response to each of the faces in face responsive voxels in the amygdala and then correlated this response with the mean trait evaluations of the faces. Consistent with this hypothesis and as shown in Figure 4–6, the amygdala activation correlated negatively with all judgments on positive traits (e.g., caring, trustworthy, attractive) and positively with all judgments on negative traits (e.g., mean, weird). That is, across trait dimensions, the amygdala responded more strongly to faces that were evaluated negatively. Although all trait judgments (except for dominance) correlated significantly with the amygdala's response, there was considerable variation in the magnitude of the correlations. According to the valence hypothesis, this variation should be predicted by the valence content of the specific judgments.

We used the valence component from the PCA of the trait judgments (*see* section A Dimensional Model of Evaluation of Emotionally Neutral Faces) as a measure of general valence evaluation. This valence component was correlated with both the response in the right and left amygdala (r = –0.50 and –0.48, respectively, p 0.001, Fig. 4–6c & 4–6d). For comparison, the amygdala's response was uncorrelated with the dominance component (r = 0.06 and 0.07, for right and left amygdala, respectively).

We used the variance accounted for by the valence component for each trait judgment as an estimate of the valence content of the trait dimension. For example, the valence component accounted for 90% of the variance of trustworthiness judgments and 9% of the variance of dominance judgments. This variance was strongly correlated with the variance accounted for by each judgment in the amygdala's response to faces (r = 0.90 and 0.79 for right and left amygdala, respectively, $p < 0.001$). The stronger the association of a trait judgment with the valence component, the stronger this judgment engaged the amygdala. Moreover, after controlling for the valence content of the trait judgments, there were no significant relationships between any of the judgments and the amygdala's response (Fig. 4–6b).

We found the same pattern of responses in face responsive regions in temporal and occipital cortices—specifically in the right superior occipital gyrus, bilateral fusiform gyri, and the right middle temporal/occipital gyrus. In all

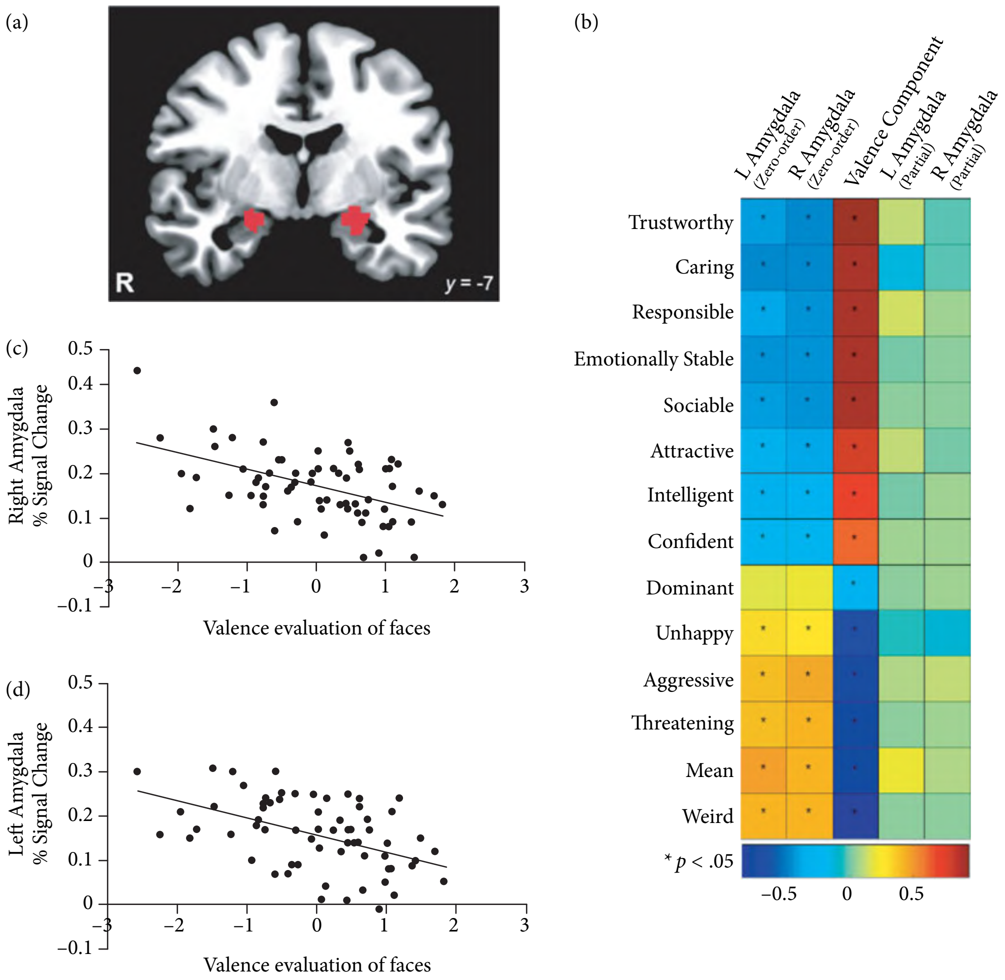


Fig. 4–6 The relation between the amygdala's response to emotionally neutral faces and variations of these faces on trait dimensions. A coronal brain slice showing face responsive voxels in bilateral amygdala (a). An intensity color plot showing correlations between the response in left and right amygdalae to faces and trait judgments of these faces (b). The first two columns show zero-order correlations and the fourth and fifth columns show partial correlations controlling for the valence content of the judgments. The third column shows the correlations between trait judgments and a valence component derived from a principal components analysis of the judgments. The traits are ordered according to their correlations with the valence component (*see* the section A Dimensional Model of Evaluation of Emotionally Neutral Faces). Scatter plots of the amygdala's response to faces (c for right and d for left) and their values on the valence component. Each point represents a face.

these regions, the response to faces was correlated with their valence, and after controlling for the valence content of specific trait judgments, there were no significant relationships between judgments and brain activation. These findings suggest that the valence evaluation of faces recruits a network of perceptual regions in temporal and occipital cortices. Additional analyses have suggested that the response in these regions is modulated by the amygdala. Specifically, controlling for the amygdala's response to faces, the relationship between the activation in these regions and face valence was no longer significant. In contrast, the relationship between the activation in the amygdala and face valence remained significant after controlling for the activation in these regions.

These findings are consistent with the hypothesis that the amygdala amplifies attention to emotionally salient stimuli in perceptual regions (Vuilleumier, 2005). Although such correlational findings cannot be conclusive for a causal influence of the amygdala on perceptual regions in temporal and occipital cortex, Vuilleumier et al. (2004) showed that whereas patients with hippocampal lesions show enhanced responses in regions in occipital and inferotemporal cortex to emotionally salient but unattended stimuli, patients with amygdala lesions do not show such enhanced responses. These regions included the same regions observed in our study. In addition, anatomical evidence from tracing studies of the macaque brain shows that the projections from the amygdala to visual cortex are more extensive than those from visual cortex to the amygdala (Amaral et al., 2003). Whereas the amygdala receives visual input only from temporal visual areas, it projects to multiple areas in both temporal and occipital visual areas, including early visual areas.

The findings suggest that the amygdala automatically evaluates novel faces along a general valence dimension and that it modulates a face responsive network of regions in occipital and temporal cortices recruited for this evaluation. The extent to which the amygdala is engaged in tracking variations of faces on social dimensions is a function of the valence content of these dimensions. Given the high correlation between trustworthiness judgments and valence evaluation of faces (Fig. 4–1a), it is not surprising that previous studies have found that the amygdala is engaged in the evaluation of face trustworthiness. However, in light of the current findings, it would be misleading to describe the amygdala's response to emotionally neutral faces as driven by their trustworthiness. In terms of practical implications, it would often be unfeasible to collect multiple social judgments of faces to estimate their valence evaluation, although some of our analyses suggest that a robust estimation of face valence can be achieved with as few as five different social judgments (Oosterhof & Todorov, 2008). If, in fact, it is unfeasible to collect multiple judgments, then it would be best to collect judgments of trustworthiness as an *approximation* of general valence evaluation.

As shown in Figure 4–6c and 4–6d, the response of the amygdala to face valence was linear. However, there have been three recent studies—two coming from our lab—reporting a nonlinear amygdala response to face trustworthiness (Said et al., 2009; Todorov et al., 2008) and face attractiveness (Winston et al., 2007). Specifically, as noted for the left amygdala in one of our prior studies (Todorov et al., 2008), the activation was stronger to faces at the extremes of the dimensions than to faces at the middle of the dimension.

The nonlinear responses to face trustworthiness are broadly consistent with the emotion overgeneralization hypothesis (section Emotion Overgeneralization Mechanisms). As our modeling and behavioral findings showed (sections on Computer Modeling of Face Trustworthiness and Face Dominance and Emotion Overgeneralization Mechanisms), variations on the dimension of trustworthiness can be understood in terms of similarity to expressions of happiness on the positive extreme of the dimension and expressions of anger on the negative end. Given that a number of functional neuro-imaging studies have found a stronger amygdala response to happy than to neutral faces (e.g., Breiter et al., 1996; Pessoa et al., 2006; Winston, O'Doherty, & Dolan, 2003; Yang et al., 2002), it should be possible to observe a nonlinear response to face trustworthiness with elevated responses to both extremely trustworthy and untrustworthy faces.

Even if this is the case, one should be able to specify the conditions under which the amygdala's response to face valence is linear and the conditions under which the response is nonlinear. There are at least two hypotheses about these conditions. According to the first hypothesis, the nature of the evaluation—implicit versus explicit—may be critical. In contrast to the study by Engell et al. (2007), participants in Said et al.'s study explicitly evaluated the faces on trustworthiness, and this may have biased attention to extreme faces. In a recent study, Cunningham et al. (2008) observed similar quadratic responses in the amygdala in a valence evaluation task of famous people. When participants focused on the positivity of the evaluation, the response was enhanced to positive stimuli; when they focused on the negativity, the response was enhanced to negative stimuli.

However, this hypothesis cannot account for all of the data. In Todorov et al. (2008), the task was the same as the task used in Engell et al. According to the second hypothesis, the range of face valence used in a particular study may determine the nature of the amygdala's response. For wider ranges of face valence, the response may be quadratic. For example, we compared the trustworthiness of the faces used in Todorov et al. (2008) and the faces used in Engell et al. (2007) in our computer model of face trustworthiness (Oosterhof & Todorov, 2008). The range of trustworthiness in the former study was from –3.26 to 2.64 in SD units, whereas the range in the latter study was from –1.79 to 1.53. The range for the faces used in Said et al.'s study, in which participants explicitly evaluated the faces, was from –2.71 to 1.37. Studies on attractiveness typically use extreme faces (O'Doherty et al., 2003; Winston et al., 2007), and given the high correlation between attractiveness and face valence, this can lead to nonlinear responses in the amygdala as observed by Winston et al. (2007).

Both of these hypotheses, as well as linear and nonlinear responses in the amygdala, are consistent with a common attentional mechanism according to which the amygdala biases attention toward stimuli that are of current motivational significance to the person (Cunningham et al., 2008; LaBar et al., 2001; Vuilleumier, 2005). Interestingly, early studies in social cognition showed that allocation of attention to social stimuli exhibits nonlinear quadratic response to people as a function of their extremeness rather than their valence (Fiske, 1980), and more recent studies have shown that evaluative processes are context-dependent (Ferguson & Bargh, 2004).

As argued in section Emotion Overgeneralization Mechanisms, valence evaluation of faces may be in the service of approach/avoidance decisions. This is consistent with findings that macaque monkeys with bilateral amygdala lesions exhibit uninhibited approach behaviors during social interactions (Emery et al., 2001) and with theories that posit that one of the primary functions of the amygdala is to provide continuous vigilance by evaluating objects and agents prior to interacting with them (Amaral, 2002; Whalen, 1998). Evaluation processes in the amygdala may not only enhance attention and processing of stimuli in perceptual areas (Anderson & Phelps, 2001; Vuilleumier et al., 2004) but may also influence approach/avoidance decisions via interactions with orbital frontal cortex (Baxter et al., 2000).

Beyond facial appearance: impressions from behaviors

Before the widespread use of fMRI to study the neural basis of social cognition, Leslie Brothers wrote that "the visual appearance of a face in social cognition is analogous to a stream of speech in linguistic processing: the face stimulus is immediately and obligatorily transformed into the representation of a person (with dispositions and intentions) before having access to consciousness." (Brothers, 1990, p. 35). These prescient insights are consistent with recent behavioral (Todorov & Uleman, 2004) and neuro-imaging studies (Chapter 3; Todorov, Gobbini, Evans, & Haxby, 2007). As Gobbini and her colleagues have shown, faces of significant others activate a network of regions implicated in social cognition such as the medial prefrontal cortex (mPFC) and precuneus (Chapter 3; Gobbini & Haxby, 2007; Gobbini, Leibenluft, Santiago, & Haxby, 2004).

These effects are based on prior person knowledge rather than on facial appearance. There is a long tradition of research in social psychology showing that people form impressions from observing the behaviors of other people (e.g., Gilbert & Malone, 1995; Jones & Davis, 1965; Trope, 1986; for a review, *see*

Gilbert, 1998). A number of studies on spontaneous trait inferences (STIs) from behaviors have demonstrated that such inferences are associated with the faces that accompanied the behaviors (Carlston & Skowronski, 1994; Carlston, Skowronski, & Sparks, 1995; Goren & Todorov, 2009; Todorov & Uleman, 2002, 2003, 2004). Importantly, in our STI studies, we randomly assigned behaviors to faces to avoid effects of facial appearance on inferences and subsequent judgments.

To study whether rapidly acquired person knowledge affects the neural representation of faces, we conducted an fMRI study (Todorov et al., 2007) modeled after our behavioral paradigm in which faces are presented with single behavioral descriptions for a few seconds. In the first stage of the study, participants familiarized themselves with the faces and behaviors. In the second stage, they were presented with the faces that were associated with behaviors intermixed with novel faces. Although the task was perceptual—deciding whether each face was the same as the preceding one—and did not demand retrieval of person knowledge, the rapidly acquired person knowledge modulated the response to faces in a number of brain regions. Specifically, faces that were associated with behaviors evoked a stronger response in the mPFC and the STS than novel faces. Moreover, the type of behaviors associated with the faces affected the response to the faces. For example, faces associated with disgusting behaviors evoked a stronger response in the anterior insula, a region implicated in the processing of disgust related stimuli (Calder et al., 2000; Philips et al., 1997), than faces associated with aggressive behaviors. These findings are consistent with Brothers' hypothesis that person knowledge is automatically retrieved in the process of face perception.

From an adaptive point of view, people should be able to rapidly learn about other people and overwrite initial impressions. The robustness of the learning process is demonstrated by findings that person learning *(1)* occurs after minimal time exposure to faces and behaviors; *(2)* is relatively independent of availability of cognitive resources; *(3)* is independent of explicit goals to form impressions; and *(4)* subsequent effects on perception and judgments are independent of explicit memory for the behaviors (Todorov & Uleman, 2003).

Findings from studies of patients with brain lesions are consistent with the idea of robust person learning mechanisms (Johnson et al., 1985; Tranel & Damasio, 1993; Todorov & Olson, 2008). In a particularly striking case of brain damage, Tranel and Damasio described a patient (Boswell) with an extensive damage in the medial temporal lobe and orbitofrontal cortex. Boswell had dense amnesia, did not recognize the faces of caregivers, and did not even show increased galvanic skin response to familiar faces, an index of implicit face processing. However, if consistently treated nicely by a caregiver, Boswell had a reliable preference for her face in forced choice preference tasks.

This case is extreme, but research with Korsakoff patients has also showed that they can acquire and preserve affective responses to people's faces despite lack of explicit memory (Johnson et al., 1985). Johnson et al. presented such patients with two pictures and described one of the people as bad (e.g., "… stole a car … robbed an old man who lived in the neighborhood") and the other as good (e.g., " … joined the Navy … saved a fellow sailor"). The patients reliably preferred the good person despite lack of memory for the origin of these impressions.

Recently, in a conceptual replication of Johnson et al., we studied how inferences from facial appearance and behavioral descriptions were integrated in person impressions (Todorov & Olson, 2008). Normal participants and three patients with amnesia caused by lesions in the hippocampus were presented with trustworthy- and untrustworthy-looking faces paired with trustworthy and untrustworthy behaviors. After the learning stage of the experiment, participants were asked to judge the faces on a number of trait dimensions. One of the patients with a localized lesion in the hippocampus showed excellent learning just as young and older control participants did. Faces associated with positive behaviors were judged more positively than faces associated with negative behaviors, and this learning effect was stronger than the effect of facial appearance on judgments. The other two patients, whose lesions extended into the left amygdala and left temporal pole, showed little evidence of learning. At the same time, all patients showed effects of facial appearance on judgments similar to the effects observed for prosopagnosics (Todorov & Duchaine, 2008). These findings suggest that the hippocampus may not be necessary for forming of affective associations with faces. Other structures in the medial temporal lobe like the amygdala may be critical for this process (Somerville, Wig, Whalen, & Kelley, 2006).

The findings show that learning can overwrite initial impressions based on facial appearance. However, at present, we lack models specifying how person representations are dynamically updated in the brain. One of the most important tasks for future research is to specify models of how different sources of person information are integrated in coherent person representations.

Conclusions and Outstanding Questions

People rapidly form impressions of other people based on minimal information. In this chapter, I focused on the processes underlying evaluation of faces. Although people evaluate faces on multiple trait dimensions, these evaluations are highly correlated with each other. Findings from data-driven methods suggest that these evaluations can be represented within a two-dimensional space defined by valence and dominance evaluation of faces (section A Dimensional Model of Evaluation of Emotionally Neutral Faces). Computer modeling findings suggest that whereas valence evaluation is based on facial cues resembling emotional expressions signaling approach/avoidance behavior, dominance evaluation is based on cues signaling physical strength (section Computer Modeling of Face Trustworthiness and Face Dominance). Additional behavioral and computer modeling studies provide convergent evidence that evaluation of emotionally neutral faces is rooted in adaptive mechanisms for inferring emotional states and corresponding behavioral intentions (section Emotion Overgeneralization Mechanisms). The two-dimensional model provides a unifying framework for the study of face evaluation. In light of this framework, a re-analysis of functional neuro-imaging data from an implicit face evaluation paradigm has showed that the amygdala *(1)* is engaged in general valence evaluation rather than in specific trait evaluations of faces and *(2)* modulates the activity in perceptual areas in temporal and occipital cortices (section The Role of the Amygdala in Evaluation of Emotionally Neutral Faces). Finally, I reviewed initial evidence about how person representations are updated in light of new knowledge (section Beyond Facial Appearance: Impressions From Behaviors).

Although we have made progress in identifying some of the key phenomena in face evaluation and the key brain regions involved in this evaluation, there are a number of outstanding questions. First, although people do agree in making judgments from facial appearance, there are also individual differences in these judgments (Engell et al., 2007; Hönekopp, 2006). Understanding these differences may be critical for understanding perceptual learning and top-down effects on social perception. Two likely sources of idiosyncratic contributions to judgments of novel faces are self-resemblance (e.g., DeBruine, 2002, 2005) and resemblance to faces of familiar people. This latter source is directly related to how person knowledge can affect evaluation of novel faces. At present, there are no compelling tests of this hypothesis. Similarly, we do not know what neural regions subserve idiosyncratic contributions to person judgments. To begin addressing these questions, we need statistical models that can estimate consensus and idiosyncratic contributions to these judgments.

Second, the dimensional model of face evaluation was designed as a general model of implicit face evaluation and may be most applicable to situations where no specific evaluative context is provided (e.g., Engell et al., 2007). Specific judgmental dimensions (e.g., variance in competence judgments not shared with the valence component) may be prominent in contexts that make these dimensions relevant. For example, in electoral decisions, voters believe that competence is the most important attribute for a politician and evaluations of competence but not trustworthiness predict electoral success (Todorov et al., 2005). Similarly, in mating decisions, physical attractiveness could trump evaluations on other dimensions including trustworthiness (DeBruine, 2005). Such decisional contexts may change the nature of brain responses and recruit specific sets of brain regions. As argued in the section The Role of the Amygdala in Evaluation of Emotionally Neutral Faces, the nature of the evaluation task may be critical for the shape of the amygdala's response to faces.

Third, according to the emotion overgeneralization hypothesis, the same neural systems should underlie face evaluation on trait dimensions and perception of emotional expressions. The direct tests of this hypothesis remain to be conducted. Multi-voxel pattern analysis methods (Norman, Polyn, Detre, & Haxby, 2007) and methods designed to measure adaptation effects in fMRI designs (Aguirre, 2007) would be particularly useful to test these hypotheses. For example, if overgeneralization underlies evaluation of emotionally neutral faces, pattern classifiers trained to detect the pattern of neural activation associated with the perception of expressions (Said, Moore, Engell, Todorov, & Haxby, 2010) should be able to detect similar patterns of activation during the perception of neutral faces that vary on trait dimensions associated with the particular expressions (e.g., dominance and anger).

Fourth, although several studies have implicated the amygdala as one of the key regions in face evaluation (section The Role of the Amygdala in Evaluation of Emotionally Neutral Faces), we know very little about the interactions between the regions involved in this evaluation. The amygdala is in the perfect anatomical position—with connections to orbitofrontal, temporal, and occipital cortices—to serve as an interface between perceptual and decision making regions in the brain. We need better mechanistic models of the interactions between these regions based on both anatomical evidence and causal modeling of their dynamic interactions.

Addressing all these questions will require well-defined behavioral models that constrain and guide research on the neural basis of social cognition. We have tried to follow this research strategy in our lab.

References

Aguirre, G. K. (2007). Continuous carry-over designs for fMRI. *Neuroimage, 35*, 1480–1494.

Adams, R. B., Ambady, N., Macrae, N., & Kleck, R. E. (2006). Emotional expressions forecast approach-avoidance behavior. *Motivation & Emotion, 30*, 179–188.

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current Opinion in Neurobiology, 12*, 169–177.

Adolphs, R. (2003). Cognitive neuroscience of human social behavior. *Nature Reviews Neuroscience, 4*, 165–178.

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. R. (1995). Fear and the human amygdala. *The Journal of Neuroscience, 15*, 5879–5891.

Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology, 55*, 387–395.

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences, 4*, 267–278.

Amaral, D. G. (2002). The primate amygdala and the neurobiology of social behavior: implications for understanding social anxiety. *Biological Psychiatry, 51*, 11–17.

Amaral, D. G., Behniea, H., & Kelly, J. L. (2003). Topographic organization of projections from the amygdala to the visual cortex in the macaque monkey. *Neuroscience, 118*, 1099–1120.

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

Ambady, N., Hallahan, M., & Rosenthal, R. (1995). On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology, 69*, 518–529.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychological Bulletin, 111*, 256–274.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258–290.

Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences of the USA, 104*(46), 17,948–17,953.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion, 6*, 269–278.

Baxter, M. G., Parker, A., Lindner, C. C. C., Izquierdo, A. D., & Murray, E. A. (2000). Control of response selection by reinforcer value requires interaction of amygdala and orbital prefrontal cortex. *Journal of Neuroscience, 20*, 4311–4319.

Bentin, S., Degutis, J. M., D'Esposito, M., & Robertson, L. C. (2007). Too many trees to see the forest: Performance, event-related potential, and functional magnetic resonance imaging manifestations of integrative congenital prosopagnosia. *Journal of Cognitive Neuroscience, 19*, 132–146.

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science, 15*, 674–679.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 187–194.

Blanz, V., & Vetter, T. (2003). Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*, 1063–1074.

Breiter, H. C., Etcoff, N. L., Whalen, P. J., et al. (1996). Response and habituation of the human amygdala during visual processing of facial expression. *Neuron, 17*, 875–887.

Brothers, L. (1990). The social brain: a project for integrating primate behavior and neurophysiology in a new domain. *Concepts in Neuroscience, 1*, 27–51.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305–327.

Bruner, J. S., & Tagiuri, R. (1954). The perception of people. In G. Lindzey (Ed.), *Handbook of Social Psychology*, Vol. 2. Cambridge, MA: Addison-Wesley.

Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience, 3*, 1077–1078.

Calder, A. J., Lawrence, A. D., & Young, A. W. (2001). Neuropsychology of fear and loathing. *Nature Reviews Neuroscience, 2*, 352–363.

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience, 6*, 641–651.

Carlston, D. E., & Skowronski, J. J. (1994). Saving in the relearning of trait information as evidence for spontaneous inference generation. *Journal of Personality and Social Psychology, 66*, 840–856.

Carlston, D. E., Skowronski, J. J., & Sparks, C. (1995). Savings in relearning: II. On the formation of behavior-based trait associations and inferences. *Journal of Personality and Social Psychology, 69*, 420–436.

Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin, 25*, 215–224.

Cronbach, L. J. (1955). Processes affecting scores on understanding others and assuming "similarity." *Psychological Bulletin, 52*, 177–193.

Cunningham, W. A., Van Bavel, J. J., & Johnsen, I. R. (2008) Affective flexibility: evaluative processing goals shape amygdala activity. *Psychological Science, 19*, 152–160.

Damasio, A. R., Tranel, D., & Damasio, H. (1990). Face agnosia and the neural substrates of memory. *Annual Review of Neuroscience, 13*, 89–109.

Davis, M., & Whalen, P. J. (2001). The amygdala: vigilance and emotion. *Molecular Psychiatry, 6*, 13–34.

DeBruine, L. M. (2002). Facial resemblance enhances trust. *Proceedings of the Royal Society B, 269*, 1307–1312.

DeBruine, L. M. (2005). Trustworthy but not lustworthy: context-specific effects of facial resemblance. *Proceedings of the Royal Society B, 272*, 919–922.

Duchaine, B. C., Parker, H., & Nakayama, K. (2003). Normal emotion recognition in a developmental prosopagnosic. *Perception, 32*, 827–838.

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychological Science, 17*, 383–386.

Emery, N. J., Capitanio, J. P., Mason, W. A., Machado, C. J., Mendoza, S. P., & Amaral, D. G. (2001). The effects of bilateral lesions of the amygdala on dyadic social interactions in rhesus monkeys (Macaca mulatta). *Behavioral Neuroscience, 115*, 515–544.

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: automatic coding of face properties in human amygdala. *Journal of Cognitive Neuroscience, 19*, 1508–1519.

Engell, A., Todorov, A., & Haxby, J. (In press). Common neural mechanisms for the evaluation of facial trustworthiness and emotional expressions as revealed by behavioral adaptation. *Perception*.

Fiske, S. T. (1980). Attention and weight in person perception: the impact of negative and extreme behavior. *Journal of Personality and Social Psychology, 38*, 889–906.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences, 11*, 77–83.

Ferguson, M. J., & Bargh, J. A. (2004). Liking is for doing: the effects of goal pursuit on automatic evaluation. *Journal of Personality and Social Psychology, 87*, 557–572.

Fox, C. J., & Barton, J. J. (2007). What is adapted in face adaptation? The neural representations of expression in the human visual system. *Brain Research, 1127*, 80–89.

Fridlund, A. J. (1994). *Human Facial Expression: An Evolutionary View*. San Diego, CA: Academic Press.

Gilbert, D. T. (1998). Ordinary personology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology*, Vol. 2 (pp. 89–150). New York: McGraw-Hill.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*, 21–38.

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia, 45*, 32–41.

Gobbini, M. I., Leibenluft, E., Santiago, N., & Haxby, J. V. (2004). Social and emotional attachment in the neural representation of faces. *NeuroImage, 22*, 1628–1635.

Goren, A., & Todorov, A. (2009). Two faces are better than one: eliminating false trait associations with faces. *Social Cognition, 27*, 222–248.

Hall, C. C., Goren, A., Chaiken, S., & Todorov, A. (2009). Shallow cues with deep effects: trait judgments from faces and voting decisions. In E. Borgida, J. L. Sullivan, & C. M. Federico (Eds.), *The Political Psychology of Democratic Citizenship* (pp. 73–99). London: Oxford University Press.

Hamermesh, D. & Biddle, J. (1994). Beauty and the labor market. *The American Economic Review, 84*, 1174–1194.

Hassin, R., & Trope, Y. (2000). Facing faces: studies on the cognitive aspects of physiognomy. *Journal of Personality and Social Psychology, 78*, 837–852.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*, 2425–2430.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*, 223–233.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2002). Human neural systems for face recognition and social communication. *Biological Psychiatry, 51*, 59–67.

Haxby, J. V., Ungerleider, L. G., Clark, V. P., Schouten, J. L., Hoffman, E. A., & Martin, A. (1999). The effect of face inversion on activity in human neural systems for face and object perception. *Neuron, 22*, 189–199.

Hess, U., Blairy, S., & Kleck, R. E. (2000). The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation. *Journal of Nonverbal Behavior, 24*, 265–283.

Hoffman, E., & Haxby, J. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience, 3*, 80–84.

Hönekopp, J. (2006). Once more: is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 199–209.

Humphreys, K., Avidan, G., & Behrmann, M. (2007). A detailed investigation of facial expression processing in congenital prosopagnosia as compared to acquired prosopagnosia. *Experimental Brain Research, 176*, 356–373.

Jeffrey, L., Rhodes, G., & Busey, T. (2006). View-specific coding of face shape. *Psychological Science, 17*, 501–505.

Johnson, M. K., Kim, J. K., & Risse, G. (1985). Do alcoholic Korsakoff's Syndrome patients acquire affective reactions? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 22–36.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: the attribution process in person perception. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 2, pp. 219–266). New York: Academic 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.

Kim, M. P., & Rosenberg, S. (1980). Comparison of two structural models of implicit personality theory. *Journal of Personality and Social Psychology, 38*, 375–389.

Knutson, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior, 20*, 165–181.

Krumhuber, E., Manstead, A. S. R., Kappas, A., Cosker, D., Marshall, D., & Rosin, P. L. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion, 7*, 730–735.

LaBar, K. S., Gitelman, D. R., Parrish, T. B., Kim, Y.-H., Nobre, A. C., & Mesulam, M.-M. (2001). Hunger selectively modulates corticolimbic activation to food stimuli in humans. *Behavioral Neuroscience, 115*, 493–500.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin, 126*, 390–423.

Little, A. C., Burriss, R. P., Jones, B. C., & Roberts, S. C. (2007). Facial appearance affects voting decisions. *Evolution and Human Behavior, 28*, 18–27.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska Directed Emotional Faces*. Psychology section, Department of Clinical Neuroscience, Karolinska Institute, Stockholm, Sweden.

Macrae, C. N., Quinn, K. A., Mason, M. F., & Quadflieg, S. (2005). Understanding others: the face and person construal. *Journal of Personality and Social Psychology, 89*, 686–695.

Marsh, A. A., Ambady, N., & Kleck, R. E. (2005). The effects of fear and anger facial expressions on approach- and avoidance-related behaviors. *Emotion, 5*, 119–124.

Mazur, A., Mazur, J., & Keating, C. (1984). Military rank attainment of a West Point class: effects of cadets' physical features. *American Journal of Sociology, 90*, 125–150.

McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997) Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neurosciences, 9*, 605–610.

Montepare, J. M., & Dobish, H. (2003). The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of Nonverbal Behavior, 27*, 237–254.

Montepare, J. M., & Zebrowitz, L. A. (1998). Person perception comes of age: the salience and significance of age in social judgments. *Advances in Experimental Social Psychology, 30*, 93–161.

Morris, J. S., Frith, C. D., Perrett, D. I., Rowland, D., Young, A.W., Calder, A. J., & Dolan, R. J. (1996). A differential neural response in the human amygdala to fearful and happy expressions. *Nature, 383*, 812–815.

Mueller, U., & Mazur, A. (1996). Facial dominance of West Point cadets as a predictor of later military rank. *Social Forces, 74*, 823–850.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2007). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences, 10*, 424–430.

O'Doherty, J., Winston, J., Critchley, H., Perrett, D., Burt, D. M., & Dolan, R. J. (2003). Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia, 41*, 147–155.

Olivola, C. Y., & Todorov, A. (2010). Fooled by first impressions? Re-examining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology, 46*, 315–324.

Olson, I. R., & Marshuetz, C. (2005). Facial attractiveness is appraised in a glance. *Emotion, 5*, 498–502.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the USA, 105*, 11087–11092.

Oosterhof, N. N., & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion, 9*, 128–133.

Osgood, C. E., Suci, G. I., & Tennenbaum, P. H. (1957). *The Measurement of Meaning.* Urbana: University of Illinois Press.

Pessoa, L., Japee, S., Sturman, D., & Underleider, L. G. (2006). Target visibility and visual awareness modulate amygdala responses to fearful faces. *Cerebral Cortex, 16*, 366–375.

Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage, 16*, 331–348.

Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: from animal models to human behavior. *Neuron, 48*, 175–187.

Phillips, M. L., Young, A. W., Senior, C., et al. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature, 389*, 495–498.

Puce, A., Allison, T., Benit, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation of humans viewing eye and mouth movements. *Journal of Neuroscience, 18*, 2188–2199.

Rhodes, G., Jeffery, L., Watson, T. L., Clifford, C. W. G., & Nakayama, K. (2003). Fitting the mind to the world: face adaptation and attractiveness aftereffects. *Psychological Science*, 14, 558–566.

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9*, 283–294.

Said, C. P., Baron, S., & Todorov, A. (2009). Nonlinear amygdala response to face trustworthiness: Contributions of high and low spatial frequency information. *Journal of Cognitive Neuroscience, 21*, 519–528.

Said, C. P., Moore, C. D., Engell, A. D., Todorov, A., & Haxby, J. V. (2010). Distributed representations of dynamic facial expressions in the superior temporal sulcus. *Journal of Vision, 10*(5), 1–12.

Said, C., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion, 9*, 260–264.

Schneider, D. J. (1973). Implicit personality theory: a review. *Psychological Bulletin, 79*, 294–309.

Singular Inversions (2006). *FaceGen 3.1 Ful l SDK Documentation*. http://facegen.com. Date last accessed July 30, 2008.

Somerville, L. H., Wig, G. S., Whalen, P. J., & Kelley, W. M. (2006). Dissociable medial temporal lobe contributions to social memory. *Journal of Cognitive Neuroscience, 18*, 1253–1265.

Todorov, A. (2008). Evaluating faces on trustworthiness: an extension of systems for recognition of emotions signaling approach/ avoidance behaviors. In A. Kingstone & M. Miller (Eds.), The Year in Cognitive Neuroscience 2008, *Annals of the New York Academy of Sciences, 1124*, 208–224.

Todorov, A., Baron, S., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: a model based approach. *Social, Cognitive, & Affective Neuroscience, 3*, 119–127.

Todorov, A., & Duchaine, B. (2008). Reading trustworthiness in faces without recognizing faces. *Cognitive Neuropsychology*, 25, 395–410.

Todorov, A., & Engell, A. (2008). The role of the amygdala in implicit evaluation of emotionally neutral faces. *Social, Cognitive, & Affective Neuroscience, 3*, 303–312.

Todorov, A., Gobbini, M. I., Evans, K. K, & Haxby, J. V. (2007). Spontaneous retrieval of affective person knowledge in face perception. *Neuropsychologia, 45*, 163–173.

Todorov, A., & Olson, I. (2008). Robust learning of affective trait associations with faces when the hippocampus is damaged, but not when the amygdala and temporal pole are damaged. *Social, Cognitive, & Affective Neuroscience, 3*, 195–203.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*, 1623–1626.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27*, 813–833.

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actor's faces: evidence from a false recognition paradigm. *Journal of Personality and Social Psychology, 83*, 1051–1065.

Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actor's faces. *Journal of Experimental Social Psychology, 39*, 549–562.

Todorov, A., & Uleman, J. S. (2004). The person reference process in spontaneous trait inferences. *Journal of Personality and Social Psychology, 87*, 482–493.

Tranel, D., & Damasio, A. R. (1993). The covert learning of affective valence does not require structures in hippocampal system or amygdala. *Journal of Cognitive Neuroscience, 5*, 79–88.

Tranel, D., Damasio, A. R., & Damasio, H. (1988). Intact recognition of facial expression, gender, and age in patients with impaired recognition of face identity. *Neurology, 38*, 690–696.

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review, 93*, 239–257.

Uleman, J. S., Blader, S., & Todorov, A. (2005). Implicit impressions. In R. Hassin, J. S. Uleman, & J. A. Bargh (Eds.), *The New Unconscious* (pp. 362–392). New York: Oxford University Press.

Uleman, J. S., Newman, L. S., & Moskowitz, G. B. (1996). People as flexible interpreters: evidence and issues from spontaneous trait inference. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 28, pp. 211–279). San Diego, CA: Academic Press.

Vuilleumier, P. (2005). How brains beware: neural mechanisms of emotional attention. *Trends in Cognitive Sciences, 9*, 585–594.

Vuilleumier, P., Richardson, M., Armony, J. L., Driver, J., & Dolan, R. J. (2005). Distant influences of amygdala lesion on visual cortical activation during emotional face processing. *Nature Neuroscience, 7*, 1271–1278.

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature, 428*(6982), 557–561.

Webster, M. A., & MacLin, O. H. (1999). Figural aftereffects in the perception of faces. *Psychonomic Bulletin & Review, 6*(4), 647–653.

Whalen, P. J. (1998). Fear, vigilance, and ambiguity: initial neuroimaging studies of the human amygdala. *Current Directions in Psychological Science, 7*, 177–188.

Wiggins, J. S. (1979). A psychological taxonomy of trait descriptive terms: the interpersonal domain. *Journal of Personality and Social Psychology, 37*, 395–412

Wiggins, J. S., Philips, N., & Trapnell, P. (1989). Circular reasoning about interpersonal behavior: evidence concerning some untested assumptions underlying diagnostic classification. *Journal of Personality and Social Psychology, 56*, 296–305.

Willis, J., & Todorov, A. (2006). First impressions: making up your mind after 100 ms exposure to a face. *Psychological Science, 17*, 592–598.

Winston, J., O'Doherty, J., & Dolan, R. J. (2003). Common and distinct neural responses during direct and incidental processing of multiple facial emotions. *Neuroimage, 20*, 84–97.

Winston, J., O'Doherty, J., Kilner, J. M., Perrett, D. I., & Dolan, R. J. (2007). Brain systems for assessing facial attractiveness, *Neuropsychologia, 45*, 195–206.

Winston, J., Strange, B., O'Doherty, J., & Dolan, R. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of face. *Nature Neuroscience, 5*, 277–283.

Wiggins, J. S., Philips, N., & Trapnell, P. (1989). Circular reasoning about interpersonal behavior: evidence concerning some untested assumptions underlying diagnostic classification. *Journal of Personality and Social Psychology, 56*, 296–305.

Yang, T. T., Menon, V., Eliez, S., et al. (2002). Amygdalar activation associated with positive and negative facial expressions. *NeuroReport, 13*, 1737–1741.

Zebrowitz, L. A. (1999). *Reading Faces: Window to the Soul?* Boulder, CO: Westview Press.

Zebrowitz, L.A. (2004). The origins of first impressions. *Journal of Cultural and Evolutionary Psychology, 2*, 93–108.

Zebrowitz, L. A., & McDonald, S. M. (1991). The impact of litigants' babyfaceness and attractiveness on adjudications in small claims courts. *Law and Behavior, 15*, 603–623.

Zebrowitz, L. A. & Montepare, J. M. (2006). The ecological approach to person perception: evolutionary roots and contemporary offshoots. In M. Schaller, J.A. Simpson, & D.T. Kenrick. *Evolution and Social Psychology*, New York: Psychology Press.

Zebrowitz, L. A. & Montepare, J. M. (2008). Social psychological face perception: why appearance matters. *Social and Personality Psychology Compass, 2*, 1497–1517.

CHAPTER 5
Social Neuroscience and the Representation of Others: Commentary

James V. Haxby

The representation of others is a central problem that brings social neuroscience and cognitive neuroscience together. The four reviews in this section by Gobbini; Jenkins and Mitchell; Todorov; and Zaki and Ochsner present an overview of the current state of our understanding for the neural systems that participate in the representation of others and highlight the major themes and issues that characterize this area of investigation. In this commentary, I would like to address why this particular problem is of great interest to both social and cognitive neuroscientists.

The representation of others is a difficult problem in terms of the inherent cognitive demands and the level of abstraction that investigators and theorists must deal with for developing good models. At a very young age, infants distinguish between two classes of entities in the world: objects and agents (Mandler, 1992). Objects move in response to external forces, whereas agents generate their actions. Object movement, therefore, mostly can be understood in terms of observable external forces. Understanding the behavior of agents, on the other hand, requires a representation of unobservable inner states that initiate, direct, and motivate that behavior. These unobservable inner states are inferred from the behavior of agents—a much more difficult problem than understanding the physical interactions among objects.

The *social brain* appears to be organized for the representation of agents and the inner states of agents that underlie their behavior. The social brain hypothesis has two principle forms. The first social brain hypothesis proposes that the human brain has evolved into its current form—in terms of size and functional organization—to handle the cognitive demands that are posed by living in large and complex social groups (Brothers, 1990; Dunbar, 1998). Under this hypothesis, the human brain is a social brain. The importance and sophistication of neural systems for cognitive tasks that are not necessarily social, such as tool use and abstract and analytic thought (e.g., logic and mathematics), are not overlooked, but the principal evolutionary pressure is thought to be in the realm of social cognition. The second social brain hypothesis proposes that the human brain contains systems that are specialized for social cognition that are relatively independent of the systems for nonsocial cognition (e.g., Jenkins & Mitchell, this volume). These two hypotheses are not mutually exclusive, of course. The human brain may have become a social brain via the evolution of specialized neural systems for social cognition. This second hypothesis, however, provides a better framework for discussing the relations between neural systems for social cognition relative to neural systems for nonsocial cognition. In this commentary, I will refer to the social brain in terms of the second hypothesis—namely, as the parts of the human brain that play a major role in social cognition.

The chapters by Gobbini, Jenkins and Mitchell, Todorov, and Zaki and Ochsner in this volume address different issues in the neural representation of others and the organization of the social brain. Jenkins and Mitchell address some key issues about the status of neural systems for social cognition relative to systems for nonsocial cognition. They suggest that social cognition recruits brain regions that are distinct from those that mediate nonsocial cognition and that the representation of others involves functions that are fundamentally different from those that are involved in nonsocial cognition. The chapters by Todorov and Gobbini address how face perception leads rapidly to access to person information. This information can be inferred from physical appearance (Todorov), can be based on trait inferences from minimal behaviors (Todorov), or can be based on long-term familiarity (Gobbini). The chapter by Zaki and Ochsner attempts to introduce differences between bottom-up automatic processes and top-down controlled processes in mentalizing, especially the effect of context.

What is the best way to describe the set of brain regions that are recruited for the representation of others and how are they organized? I propose that the social brain can be understood in terms of distributed, large-scale, partially overlapping neural systems for different aspects of social cognition. I would like to focus on four systems that play a major role in the representation of others. I will refer to these systems as the agent perception system, the action understanding system, the person knowledge system, and the emotion processing system.

The first social brain system mediates the perception of agent form and motion. In the visual domain, subsectors of the lateral occipital area (LO), the lateral part of the fusiform gyrus (FG), and the posterior superior temporal sulcus (pSTS) are the principal components of this system (Kanwisher et al., 1997; McCarthy et al., 1997; Haxby et al., 1999, 2000; Downing et al., 2001, 2003; Grossman et al., 2000; Grossman & Blake, 2002; Beauchamp et al., 2003; Chao et al., 1999; Gobbini et al., 2007). Although these areas are generally considered visual extrastriate cortex, considerable evidence suggests that they mediate more abstract representations that are supramodal (Pietrini et al., 2004; Beauchamp et al., 2004).

The second social brain system mediates the understanding of actions, including the activation of motor representations of perceived actions and inferring the intentions and goals implied by perceived actions. The action understanding system is based on research on mirror neurons, which were discovered in the monkey by Rizzolatti and colleagues (Gallese et al., 1996; Rizzolatti et al., 2001). In the monkey, mirror neurons are defined as those that respond to both the perception and execution of specific actions. The specificity to particular actions is an important criterion that demonstrates that these cells don't simply respond in a global way during perception and action. Work on the human mirror neuron system (hMNS) cannot use this criterion because it relies on characterizing the tuning function of individual cells. Consequently, the hMNS is defined as brain regions that are active during both the perception and execution of actions. Whereas mirror neurons have been detected only in premotor and inferior parietal cortex in the monkey, investigations in humans detect hMNS like activity in the posterior STS as well.

The third social brain system mediates the representation of person knowledge. This domain includes the representation of the mental states of others—theory of mind (ToM) (Frith & Frith, 1999, 2006; Saxe & Powell, 2006; Mitchell et al., 2002) and the representation of the enduring traits of others (Mitchell et al., 2002; Jenkins & Mitchell, this volume; Gobbini & Haxby, 2007; Gobbini, this volume; Todorov, this volume). The principal components of the person knowledge system are the medial prefrontal cortex (MPFC) and the parietal-temporal junction (TPJ). Anterior temporal cortex (ATC) and the posterior cingulate/precuneus (PCC/PC) are sometimes also implicated in this system, although they appear to play less central roles that are associated more with memory and biographical knowledge.

The fourth social brain system is more of a conglomerate of systems that mediate processing different emotional states. The representation of others involves both the representation of others' emotional states as well as one's own emotional responses to others (e.g., Gobbini, this volume). It has been proposed that the representation of others' emotional states involves mirroring those states by activating the same systems that are involved in experiencing those emotions (Wicker et al., 2003; Singer et al., 2004). Major components of the emotion processing system include the amygdala (Whalen et al., 1998; Vuilleumier, 2005; Engell et al., 2007; Todorov, this volume), the anterior insula (Phillips et al., 1997; Wicker et al., 2003; Singer et al., 2004), and the superior anterior cingulate (Singer et al., 2004).

As is clear from this thumbnail summary of the constituent systems that make up the social brain, the representation of others is a large domain that involves high-level and complex processes. In many ways, the social brain appears to be distinct from the neural systems for non-social cognition, as suggested by Mitchell and Jenkins (this volume). It is composed of distinct brain regions and operates on information that poses different computational challenges from those posed by nonsocial information. Some parts of the social brain, especially the person knowledge system, overlap extensively with the "default mode" or "intrinsic" system (Gusnard & Raichle, 2001; Golland et al., 2007), suggesting to Jenkins and Mitchell (this volume) that social cognition is not just served by different brain regions from those for nonsocial cognition but that the processes for social cognition are somehow "special." As discussed above, social cognition is special because it involves understanding the behavior of agents as compared to objects. Jenkins and Mitchell suggest that it is also special because it is associated with the default or "standby" state of the awake brain.

The cognitive functions associated with the social brain are of great interest to cognitive neuroscientists, and this interest reflects a convergence of the fields of cognitive neuroscience and social neuroscience. Social neuroscientists have found that understanding the neural systems that mediate social cognition helps to understand social cognition. Cognitive neuroscience is interested in the same systems, reflecting increasing interest in high-level representation and processing. Much of social neuroscience has focused on face perception. The neural systems that mediate face perception were first investigated by cognitive neuroscientists as an instance of high-level visual representation (e.g., Haxby et al., 2000; Kanwisher & Yovel, 2006). Similarly, biological motion is intensely investigated by cognitive neuroscientists because it involves high-level representations of complex motion that are optimized because of biological relevance (Johanson, 1972; Bonda et al., 1996; Grossman et al., 2000; Grossman & Blake, 2002; Beauchamp et al., 2002, 2003). The neural systems for action understanding are of interest to cognitive neuroscientists because they represent integration of sensory and motor representations (Rizzolatti et al., 2001) and control system theory (Wolpert et al., 2003). Emotion influences cognition and vice versa, making understanding the emotion processing system essential for understanding basic cognitive processes like memory and attention (Vuilleumier, 2005; Dolcos, Bar, & Cabeza, 2004). These two communities, cognitive neuroscience and social neuroscience, are sometimes difficult to distinguish on the topics where they intersect.

The emergence of social neuroscience as a field, therefore, does not reflect just the discovery that social cognition has a special or separate status in the brain. Rather, it reflects the realization that the human brain evolved to handle the difficult cognitive problems that are posed by being a highly social animal. Primary among these is the problem of understanding the behavior of agents. Both the cognitive neuroscience and the social neuroscience communities are passionately interested in understanding how neural systems that mediate the representation of agents—perception of agent form and motion, action understanding, person knowledge, and emotion—are organized and the processes and neural codes that are involved.

References

Beauchamp, M.S., Lee, K.E., Haxby, J.V., & Martin, A. (2003). FMRI responses to video and point-light displays of moving humans and manipulable objects. *Journal of Cognitive Neuroscience*, 15, 991–1001.

Beauchamp, M.S., Lee, K.E., Argall, B.D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, 41, 809–823.

Bonda, E., Petrides, M., Ostry, D., & Evans, A. (1996). Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *Journal of Neuroscience*, 16, 3737–3744.

Brothers, L. (1990). The social brain: A project for integrating primate behavior and neurophysiology in a new domain. *Concepts in Neuroscience*, 1, 27–51.

Chao, L.L., Martin, A., & Haxby, J.V. (1999). Are face-responsive regions selective only for faces? *Neuroreport*, 10, 2945–2950.

Dolcos, F., Bar, K., & Cabeza, R. (2004). Interaction between the amygdala and the medial temporal lobe memory system predicts better memory for emotional events. *Neuron*, 10, 855–863.

Downing, P.E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, 293, 2470–2473.

Dunbar, R.I. (1998). The social brain hypothesis. *Evolutionary Anthropology*, 178–190.

Engell, A.D., Haxby, J.V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in human amygdala. *Journal of Cognitive Neuroscience*, 19, 1508–1519.

Frith, C.D., & Frith U. (1999). Interacting minds—a biological basis. *Science*, 286, 1692–1695.

Frith, C.D., & Frith, U. (2006). How we predict what other people are going to do. *Brain Research*, 1079, 36–46.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119, 593–609.

Gobbini, M.I., & Haxby, J.V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia*, 45, 32–41.

Gobbini, M.I., Koralek, A.C., Bryan, R.E., Montgomery, K.J., & Haxby, J.V. (2007). Two takes on the social brain: A comparison of theory of mind tasks. *Journal of Cognitive Neuroscience*, 19, 1803–1814.

Golland, Y., Bentin, S., Gelbard, H., et al. (2007). Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation. *Cerebral Cortex*, 17, 766–777.

Grossman, E., Donnelly, M., Price, R., et al. (2000). Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience* 12, 711–720.

Grossman, E.D., & Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron*, 35, 1167–1175.

Gusnard, D.A., & Raichle, M. E. (2001). Searching for a baseline: Functional imaging and the resting human brain. *Nature. Reviews Neuroscience*, 10, 685–694.

Haxby, J.V., Hoffman, E.A., & Gobbini, M.I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, 4, 223–233.

Haxby, J.V., Ungerleider, L.G., Clark, V.P., Schouten, J.L., Hoffman, E.A., & Martin, A. (1999). The effect of face inversion on activity in human neural systems for face and object perception. *Neuron*, 22, 189–199.

Johansson, G. (1973). Visual percepti on of bi ological motion and a model for its analysis. *Perception and Psychophysics*, 14, 201–211.

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.

Kanwisher, N., & Yovel, G (2006) The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B*, 361, 2109–2128.

Mandler, J.M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review*, 4, 587–604.

McCarthy, G., Puce, A., Gore, J.C., & Allison, T. (1997) Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neurosciences*, 9, 605–610.

Mitchell, J.P., Heatherton, T.F., & Macrae, C.N. (2002). Distinct neural systems subserve person and object knowledge. *Proceeding of the National Academy of Sciences, USA*, 99, 15,238–15,243.

Phillips, M.L., Young, A.W., Senior, C., et al. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature*, 389, 495–498.

Pietrini, P., Furey, M.L., Ricciardi, E., et al. (2004). Beyond sensory images: Object-based representation in the human ventral pathway. *Proceedings of the National Academy of Sciences, USA*, 101, 5658–5663.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2, 661–670.

Saxe, R., & Powell, L.J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, 17, 692–699.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R.J., & Frith, C.D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, 303, 1157–1162.

Vuilleumier, P. (2005). How brains beware: Neural mechanisms of emotional attention. *Trends in Cognitive Science*, 9, 585–594.

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.

Wicker, B., Keysers, C., Plailly, J., Royet, J.P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in my insula: The common neural basis of seeing and feeling disgust. *Neuron*, 40, 655–664.

Wolpert, D.M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philos Trans R Soc Lond B Biological Science*, 358, 593–602.

This page intentionally left blank

PART II

Understanding and Representing Social Groups

This page intentionally left blank

CHAPTER 6
Perceiving Social Category Information from Faces: Using ERPs to Study Person Perception

Tiffany A. Ito

The human ability to perceive faces is particularly impressive when considered in the context of the many different social inferences we perform, as well as the high degree of accuracy and relatively little effort with which they are typically made. Even very brief glimpses at a person's face may allow us to gain information relevant to determining his or her emotional state, personality characteristics, and identity. Face perception has also been recognized as supporting inferences about social category membership, with information about an individual's race, gender, and age usually easily determined from his or her face. All of these inferences are integral to social perception, but it is the latter category of inferences that are of particular interest in this chapter. This is not to say that determining whether someone is happy versus angry or telling the difference between Brad Pitt and your grandfather are unimportant, but the ability of even very brief exposures to individuals from certain social groups to activate negative stereotypes and evaluations makes the processing of social category information from faces a particularly interesting topic of study. Moreover, these questions are increasingly being examined with neuroscience measures, allowing integration of new methods with extant findings and theories. The purpose of this chapter is, therefore, to review the line of research we have pursued using event-related brain potentials (ERPs) to study social perception, focusing particularly on the perception of race and gender cues.

Using ERPs to study social perception

ERPs are changes in electrical brain activity occurring in response to discrete events such as stimulus presentation or execution of a behavioral response. They can be recorded noninvasively from the surface of the scalp and are thought to reflect summated postsynaptic potentials from large sets of synchronously firing neurons in the cerebral cortex (Fabiani, Gratton, & Coles, 2000). The recorded electrical waveform is a time by voltage function composed of a series of positive and negative deflections.[1] Time-locked deflections in the waveform are referred to as *components*. The importance of ERPs to the study of psychological processes derives from the association of individual ERP components with distinct information processing operations (Gehring, Gratton, Coles, & Donchin, 1992). Component amplitude is thought to reflect the extent to which the associated psychological operation has been engaged, and latency of the component's peak is thought to reflect the point in time by which the operation has been completed.

Several factors make ERPs attractive for the study of social perception. First, ERPs have been examined in response to a wide range of

[1] Polarity of the signal is determined by the polarity of the electrical potential at that scalp location at that point in time relative to the reference electrode(s).

psychological operations, providing a large corpus of research from which to draw in linking observed electrical activity to its assumed underlying psychological meaning. The long history of using ERPs to study cognitive processes has resulted in the association of many different components with various cognitive operations such as aspects of attention and memory. Many of these same psychological processes are of relevance in understanding social behavior. For example, components sensitive to covert orienting processes (e.g., N100, P200, and N200) may be used to assess attention to social cues. Similarly, a number of components have been associated with behavioral control processes (e.g., N200, N400, the negative slow wave) and can be used to examine how social behavior is regulated, and components associated with attitudes and affective processes (e.g., the P300) can be used to understand attitudes and affective responses toward other people. In other cases, components uniquely associated with social processes have been identified (e.g., structural encoding of conspecifics has been associated with the N170).[2]

The high temporal resolution of ERPs (on the order of milliseconds) is also a benefit because it allows them to provide access to processes that occur very quickly after stimulus onset, facilitating inferences about the time-course and ordering of mental operations. Because of this temporal resolution, ERPs also provide access to complex mental phenomena about which participants might be unaware. As the research reviewed in this chapter will illustrate, perceiving other people is likely to be composed of multiple information processing operations, and there is no reason to expect that perceivers are explicitly aware of all aspects of this process. In addition, not having to rely on conscious awareness or a willingness to accurately report internal states makes ERPs useful in measuring socially sensitive topics such as processes related to stereotyping and prejudice. Finally, although ERPs have a lower spatial resolution than techniques such as fMRI and PET, the scalp distribution of observed activity can be used to obtain estimates of neuro-anatomical location.

Based on these considerations, we have used ERPs to study several aspects of person perception. In this chapter, we review the research we have done to address three main issues: how social category information relevant to classifying individuals into meaningful social groups is perceived, how this perception relates to implicit stereotyping and prejudice, and how social category effects may be regulated.

Social categorization

It is generally assumed that social category information is processed automatically, at least for certain dimensions such as race, gender, and age (Bodenhausen & Macrae, 1998; Brewer, 1988; Fiske & Neuberg, 1990). Although this assumption has important implications for the impressions we form about other people, past research has not always been able to examine it directly or to examine its more implicit aspects. This is because past studies either have typically relied on indirect information, using the fact that a stereotype has been activated or prejudice has been displayed as evidence that categorization has occurred (e.g., Macrae, Bodenhausen, Milne, Thorn, & Castelli, 1997), or have measured response latencies in making explicit categorizations (e.g., Stroessner, 1996; Zarate, Bonilla, & Luevano, 1995; Zarate & Smith, 1990). These types of measures clearly provide information about aspects of social categorization, but assessments of explicit categorization decisions or of other aspects of person perception, such as stereotyping, may leave unaddressed other aspects of category perception, especially early perceptual aspects.

The features discussed in the prior section make ERPs well-suited to examining issues related to social categorization. In several studies, we have done so by showing participants pictures of individuals who differ in race and gender while recording ERPs. In terms of race,

[2] Although relevant to understanding the broader process of social perception, N170 research will not be reviewed in this chapter because it is not clear whether social category information influences the initial face encoding thought to be reflected in the N170 (for a discussion, see Ito & Urland, 2005; Ito & Bartholow, 2009).

the two target groups we have studied most often are Blacks and Whites. In one study (Ito & Urland, 2003), participants (who were primarily White) saw faces of Black and White males and females. To ensure that we were assessing processes related to social categorization, all participants were explicitly instructed to attend to social category information; participants were randomly assigned to explicitly categorize the faces in terms of either race or gender.

The resulting waveforms revealed four distinct deflections: a negative-going component with a mean latency of 122 milliseconds after face onset, a positive-going component with a mean latency of 176 milliseconds, a negative-going component with a mean latency of 256 milliseconds, and a positive-going component with a mean latency of 485 milliseconds. Examples of the ERP waveforms from Ito and Urland (2003) can be seen in Figures 6–1 and 6–2. We refer to these components based on their polarity and latency as the N100, P200, N200, and P300[3], respectively. Several differences to Black versus White and male versus female faces were observed in these components.

Consistent with assumptions that social category information is processed automatically, race quickly modulated attention, showing effects in the N100 component. N100s were larger to Blacks than Whites. This continued into the next component, the P200, with P200s larger to Blacks than Whites. Both effects can be seen in Figure 6–1, panel A. Target gender did not affect N100 amplitude but did show effects in the P200, where responses were larger to males than females (*see* Fig. 6–1, panel B). In the third temporally occurring component, the N200, the direction of both race and gender effects was reversed, with N200s larger to Whites and females than Blacks and males, respectively. Given the association of these early components with attentional selection (e.g., Czigler & Geczy, 1996; Eimer, 1997; Kenemans, Kok, & Smulders, 1993; Naatanen & Gaillard, 1983; Wijers, Mulder, Okita, Mulder, & Scheffers, 1989), these results suggest initially greater attention to Blacks and males but subsequently greater attention to Whites and females.

We have speculated that the initially greater attention to Blacks from our largely White sample of participants may reflect a course vigilance effect, with participants initially orienting to stimuli that are novel and/or associated with more negativity from either personal or cultural beliefs. Similarly, males may be more strongly associated with power and agency, initially triggering greater vigilance than female faces. The time-course of these effects is consistent with other ERP studies showing affective modulation as early as 100 milliseconds (Pizzagalli, Regard, & Lehmann, 1990; Smith, Cacioppo, Larsen, & Chartrand, 2003). Because these faces are being viewed in a passive viewing context, initial vigilance processing would not reveal any continued threat to self. As processing continues, we think attention therefore begins to orient to categories of individuals typically associated with deeper processing. For race, in-group members are typically processed more deeply than out-group members (e.g., Anthony, Cooper, & Mullen, 1992; Levin, 2000), which is consistent with the larger N200s to Whites. Similarly, when there are differences in depth of processing for male and female targets, they are more often in the direction of greater attention to females (Lewin & Herlitz, 2002; McKelvie, 1981; McKelvie, Standing, St. Jean, & Law, 1993; O'Toole et al., 1998). Again, speculation that the N200 reflects depth of processing is consistent with extant ERP research, such as studies showing larger N200s to familiar as compared to unfamiliar faces (Bentin & Deouell, 2000) and one's own face as compared to faces of strangers (Tanaka, Curran, Porterfield, & Collins, 2006).

In addition to components sensitive to attention, Ito and Urland (2003) also examined the P300, a component thought to reflect updates to working memory that serve to maintain an accurate mental model of the external environment (Donchin, 1981). P300 amplitude typically increases as a function of

[3] Although the P300 had a latency longer than 300 milliseconds, we use the P300 name because its scalp distribution and response to psychological processes mirrors the classic P300 component. The N100, P200, and N200 were typically largest at frontal and central scalp sites while the P300 had a parietal-maximal distribution.

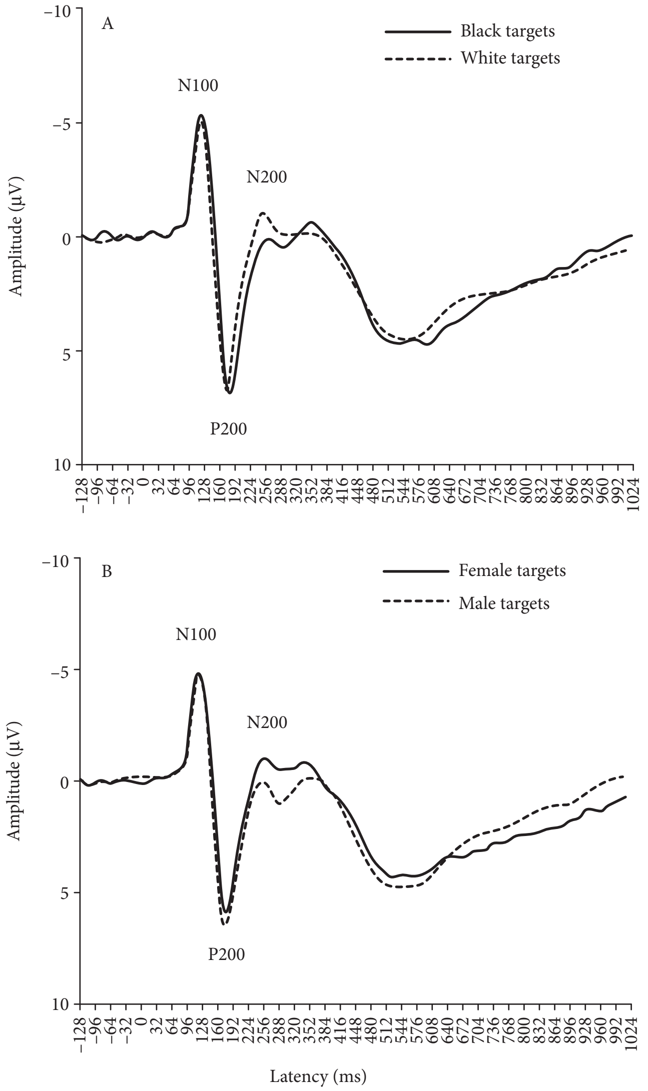


Fig. 6–1 N100, P200, and N200 responses to faces of different races and genders. Panel A shows responses as a function of target race, and Panel B shows responses as a function of target gender. Waveforms are from the central scalp area (electrode Cz). Reprinted with permission from Ito, T.A., & Urland, G.R. (2003). Race and gender on the brain: Electrocortical measures of attention to race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology, 85,* 616–626. Copyright American Psychological Association.

the discrepancy between a given stimulus and preceding stimuli along salient dimensions. To allow an examination of how social category information affects working memory, stimuli were systematically varied so that responses could be analyzed not only in terms of the race and gender of the target picture but also in terms of the race and gender of the faces that preceded it. Consistent with past P300 research (e.g., Donchin, 1981), P300s were larger when a target individual's social category membership differed from preceding individuals on the *task-relevant* dimension (as compared to when it matched the preceding faces; *see* Fig. 6–2).

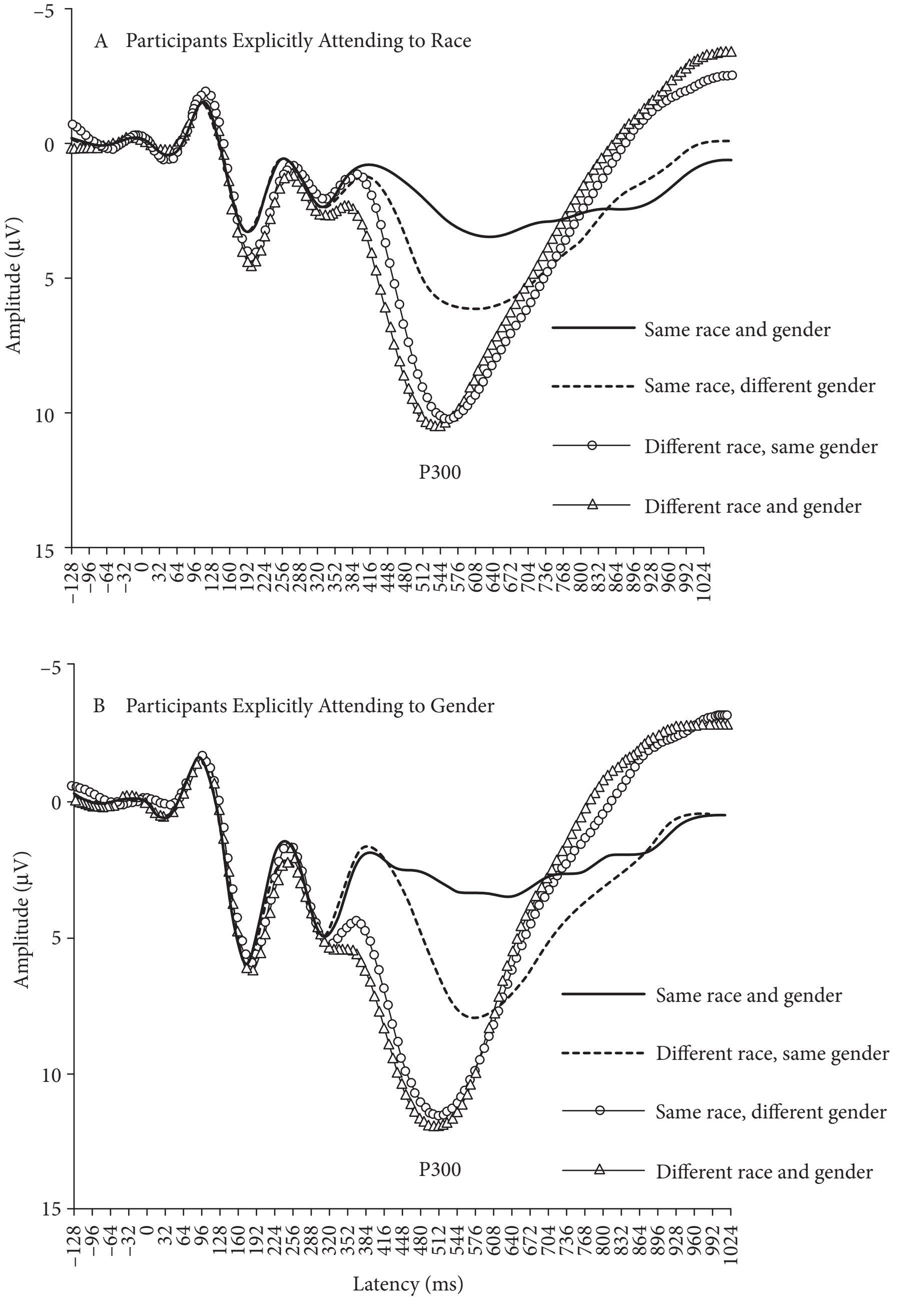


Fig. 6–2 P300 responses as a function of whether the face was of the same race and gender as the faces that immediately preceded it. Panel A shows responses from participants who were explicitly categorizing faces in terms of their race. Panel B shows responses from participants who were explicitly categorizing faces in terms of their gender. Four trial types were possible: *(1)* current face was of the same race and gender as preceding faces, *(2)* current face was the same race but a different gender than preceding faces, *(3)* current face was the same gender but a different race than preceding faces, and *(4)* current face was both a different race and gender than preceding faces. Waveforms are from the parietal scalp area (electrode Pz). Reprinted with permission from Ito, T.A., & Urland, G.R. (2003). Race and gender on the brain: Electrocortical measures of attention to race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology, 85,* 616–626. Copyright American Psychological Association.

For example, for participants categorizing faces in terms of *gender*, P300s were larger to a male face presented after a gender-incongruent female than congruent male face. This can be seen in panel B of Figure 6–2. Comparable effects occurred for participants categorizing faces in terms of race, as seen in panel A. This result confirms past P300 research in showing that working memory processes are sensitive to the social category dimension along which explicit categorization was occurring. In addition, as was expected based on the implicit attentional effects seen in the earlier components, implicit working memory effects were also seen. P300 amplitude increased when a target picture differed from the individuals pictured in preceding pictures along the *task-irrelevant* dimension. For example, for participants categorizing faces in terms of *gender*, P300s were larger to a Black face presented after a racially incongruent White face relative to a race-congruent Black face.

Effects of Perceivers' Goals

The preceding results clearly suggest that race and gender information are both quickly encoded. Moreover, the effects of social category were relatively obligatory, occurring even when participants were not explicitly attending to that dimension. We have addressed this issue more directly in subsequent studies by examining attention to race and gender under task instructions designed to direct explicit attention away from the social category information. In one study, the effects of focusing attention at a level either more shallow or deep than the social category were examined (Ito & Urland, 2005). A more shallow level of processing was encouraged by having some participants perform a visual-feature detection task where they monitored the stimuli for the presence or absence of a white dot on each face. This task focuses attention away from the social nature of the stimulus person and has been associated with the attenuation of stereotype activation (Macrae et al., 1997). Other participants were encouraged to adopt a deeper level of processing by performing an individuating task, judging whether each individual they saw would like various kinds of vegetables. This task is also successful in attenuating stereotype activation as well as decreasing differences in amygdala activation to racial out-group versus in-group faces that are thought to reflect greater negativity toward the out-group (Wheeler & Fiske, 2005). A second study examined the effects of another individuating task by having participants make introversion/extraversion judgments about each individual. This task was chosen because it was assumed to be engaging and easy for participants to perform, thereby easily directing attention away from social category cues and encouraging more person-based than category-based encoding.

Even with these very different processing goals, results indicated that race and gender information were still encoded early in processing. ERP effects were very similar to those obtained by Ito and Urland (2003), when participants were explicitly attending to race and gender; P200s were larger to Blacks and males, N200s were larger to Whites and females, and the P300 was sensitive to the match between a target's race and gender and the race and gender of preceding faces. In other words, results looked very similar to those shown in Figures 6–1 and 6–2. One meaningful difference did occur in the N100, where the N100 race differences previously obtained were attenuated. Whereas N100s were larger to Blacks when participants explicitly attended to race or gender (Ito & Urland, 2003), race did not affect the N100 when participants attended to dots or performed the vegetable preference task. It is worth noting that stimulus presentation was more complex in the latter studies. The dot task required placing dots on some of the faces, and the vegetable task required presenting the name of the vegetable about which the preference judgment was to be made before each face. This makes it difficult to determine whether the difference in N100 results reflects a processing or stimulus effect. It is notable that stimulus presentation for participants performing the introversion/extraversion task was identical to the prior studies, and N100s in that study were larger to Blacks than Whites. This suggests that increased visual complexity more so than level of processing may have been responsible for the slowing of race effects from the N100 to the P200 in the dot and vegetable task conditions.

Perception of racially ambiguous faces

We think the preceding studies aid our understanding of how sensitive perceivers are to social category information. As an extension of this work, we have also examined the interesting situation created by the growing population of multiracial individuals. Specifically, whereas the stimuli in our initial studies were chosen to be as unambiguous as possible with respect to their race and gender, the rapid increase in the number of multiracial individuals in the United States (and elsewhere) means that perceivers increasingly encounter individuals with more variable racial cues. To understand race perception in this context, while also gaining additional information on the perception of race in general, we have conducted a series of studies recording ERPs as White participants view digitally morphed photos of Asian and White males, and Black and White males (Willadsen-Jensen & Ito, 2006). The digital morphing process produces realistic photos of faces possessing features intermediate to the two "parent" racial groups. To examine responses to faces that were maximally racially ambiguous, morphs that were a 50%–50% blend of an Asian and a White face, or a Black and a White face, were created. Pilot testing determined that the faces were subjectively perceived as falling between the two racial extremes used to create them and not simply perceived as some other racial group. ERPs were then recorded as White participants viewed the racially ambiguous Asian-White morphs as well as unambiguously Asian and White faces in one study, and the racially ambiguous Black-White morphs as well as unambiguously Black and White faces in another study. Participants performed an explicit race categorization task, choosing between *Asian* and *White* in the first study, and *Black* and *White* in the second study.

Focusing first on the responses to the unambiguous faces, P200s were larger to Asians and Blacks than to Whites, and N200s were larger to Whites than Asians and Blacks. This both replicates past findings with Black and White targets and also extends the effects to another racial target group (i.e., Asians). Across multiple studies and target racial groups, then, we find that our mostly White participants show larger P200s to racial out-group members but larger N200s to racial in-group members. These effects can be seen for the study in which Black, White, and Black-White faces were seen in Figure 6–3. As can be seen in Figure 6–3, an interesting pattern was obtained for the racially ambiguous faces whereby P200 and N200 responses were indistinguishable from responses to Whites in both studies. At the same time, responses to racially ambiguous (and White) faces differed from the responses to Asians and Blacks.

It was not until the P300, peaking at around 500 milliseconds, that responses to Whites and the racially ambiguous faces diverged. This was revealed by showing participants a majority of White faces in one block and a majority of Asian or Black faces in the other block (depending on study). Recall that the P300 is sensitive to incongruities along salient dimensions, including race. As we expected, we replicated racial-incongruity effects for the *unambiguous* White, Black, and Asian faces. Using the study in which Black, White, and Black-White faces were shown as an example, this manifested as increased P300s to incongruent Black faces when they were seen in the block in which primarily White faces were shown (Fig. 6–3, panel A). Similarly, P300s were increased to incongruent White faces when they were seen in the block in which primarily Black faces were shown (Fig. 6–3, panel B). Our primary interest, of course, was responses to the racially ambiguous faces. In the block in which a majority of Black faces were seen, P300s were also increased to the racially ambiguous faces (Fig. 6–3, panel B). This continues the pattern seen in the P200 and N200 of different responses to the Black and racially ambiguous faces. But note that in the block in which a majority of White faces were seen (Fig. 6–3, panel A), P300s were now different to the White and racially ambiguous faces. Thus, for the first time in processing, perceivers were differentiating between in-group

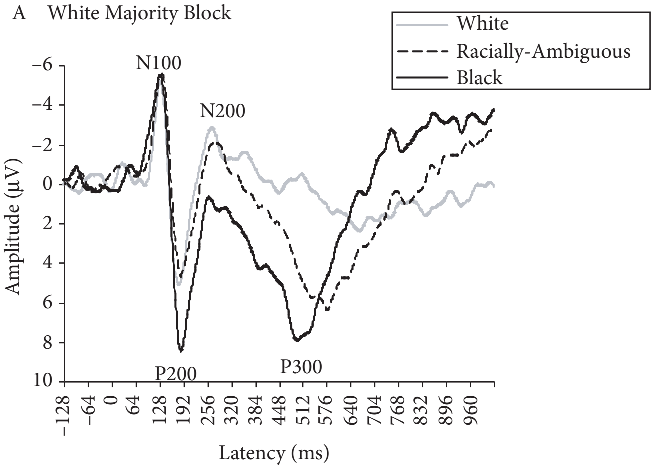


Fig. 6–3 Responses to Black, White, and racially ambiguous faces. In Panel A, faces were seen in the context of primarily White faces. In Panel B, faces were seen in the context of primarily Black faces. Waveforms are from the central scalp area (electrode Cz). Reprinted with permission from Willadsen-Jensen, E.C. & Ito, T.A. (2006). Ambiguity and the timecourse of racial perception. *Social Cognition, 24,* 580–606. Copyright Guilford Press.

and the racially ambiguous faces. Participants' explicit categorization decisions also differentiated the racially ambiguous faces from both the faces of Whites and Asians or Blacks, indicating subjective ambiguity in explicit categorization (e.g., Black-White morphs were categorized as White 50% of the time).

As noted, the participants in these studies were White. The similarity of the P200 and N200 responses between White and racially ambiguous faces could therefore reflect an assimilation of the racially -ambiguous individuals to the in-group. This is interesting in light of later effects obtained in the P300 and explicit categorization responses, where the racially ambiguous faces were perceived in a manner consistent with their objective status as 50%–50% blends between two racial groups. We think this pattern of results demonstrates that the processing of social category information is initially more gross than fine-grained. Although the participants in these studies eventually explicitly perceived the faces as belonging to neither the in-group nor out-group, the overlap in physical features between the in-group and the racially ambiguous faces appears to have lead the racially ambiguous faces to be initially processed in a manner indistinguishable from in-group faces. The interval where racially ambiguous faces switch from being processed similarly to in-group members (the N200) to being differentiated from both in-group and out-group members (the P300) may signal the point at which more finely tuned processing occurs.

Are effects the result of social category cues?

It is worth asking whether the effects discussed to this point are specific to faces, and to social category cues *per se* or whether they may result from other perceptual differences unrelated to social category information. For example, most of the studies used color stimuli. We think this is important for experimental realism because typical face-to-face human interactions occur in color. Nevertheless, it may be that faces from different social categories differ in low-level perceptual features like luminance and that these differences are driving the ERP effects we have obtained. One might worry that this could be especially likely for the race effects because basic perceptual features like luminance seem especially different for White and Black faces. We do not think this would make the effects uninteresting because, as we noted, typical human interaction includes these perceptual features, but it would suggest that the effects did not specifically result from social category information.

There are several findings that argue against this interpretation. The first is that differences are seen not only between Black and White faces but also between males and females, where the perceptual differences between categories seem smaller (Ito & Urland, 2003; 2005). Second, our race effects are the same when we use color images and when we use grayscale images that have been equated for luminance (Ito & Urland, 2003). Moreover, as the studies with the racially ambiguous faces indicate, similar effects are obtained for racial out-groups other than Blacks (i.e., Asians) for whom perceptual contrast to in-group White faces is likely smaller (Willadsen-Jensen & Ito, 2006). Finally, Kubota and Ito (2007) demonstrated that race effects were absent when the face nature of the stimuli was obscured, but all other physical differences between the stimuli were maintained. This was achieved by inverting and blurring the faces. This made them difficult to identify as faces but retained many of their physical features such as color and luminance. Participants viewed the stimuli while indicating whether they were presented to the left or right of fixation. Importantly, there were no significant race effects with these stimuli. In addition, the morphology of the waveforms was quite different than what is obtained when participants can clearly tell they are viewing faces. Together, these findings provide strong converging evidence that the effects obtained in response to differences in race and gender result from the perception of social category information, and not caused by other perceptual effects that may covary with social category.

Activation of stereotypes and prejudice

The relatively effortless degree to which social category information is processed implies that associated stereotypes and prejudices can be easily activated when encountering a group exemplar. Consistent with this, implicit stereotyping and prejudice have been demonstrated in a range of contexts (e.g., Dovidio, Kawakami, Johnson, Johnson, & Hayward, 1997; Fazio, Jackson, Dunton, & Williams, 1995; Greenwald, McGhee, & Schwartz, 1998; Macrae et al., 1997; Payne, 2001). At the same time, these effects are known to vary as a function of individual factors such as the need for self-enhancement or motivations to control prejudice (e.g., Amodio, Harmon-Jones, & Devine, 2003; Hausmann & Ryan, 2004; Sinclair & Kunda, 1999). Given the attentional differences observed to different social groups in our earlier studies, we wondered whether individual differences in these lower-level, more perceptual processes would also relate to implicit bias.

We first examined this in the context of stereotype activation, assessing whether the degree to which perceivers differentiated Black faces from White faces relatively early in perception affected the degree to which stereotypes were implicitly activated (Ito & Urland, 2006). Participants completed a sequential priming task in which they made decisions of target objects that were primed by faces of Black and White males (c.f., Judd, Blair, & Chapleau, 2004; Payne, 2001). Participants saw two different blocks of trials that were designed to examine implicit negative associations with Blacks in slightly different ways. In one block of trials, participants classified pictures of guns and insects. A stronger association between Blacks and concepts related to threat and danger has been demonstrated in a range of contexts (e.g., Correll, Park, Judd, & Wittenbrink, 2004; Payne, 2001). This block, therefore, compared responses to a class of objects associated with a negative aspect of the cultural stereotype of Blacks (guns) with a category that is also negative in valence but is not racially stereotypical (insects; Judd et al., 2004). In the other block of trials, participants classified pictures of guns and images associated with sports. The latter category contained pictures associated with basketball and football, both sports that are considered relatively stereotypical of Blacks (Judd et al., 2004). This block, therefore, allowed a comparison between the same negative aspect of the stereotype and a category that is also racially stereotypical but positive in valence.

ERPs recorded to the face primes replicated the results obtained in our prior social categorization studies—that is, P200s were larger to Blacks and N200s were larger to Whites.[4] We also obtained the expected priming effects. For example, when the target was a gun, participants were faster to classify it following a Black than White prime. Of greater interest was whether the magnitude of these stereotyping effects was predicted by magnitude of the ERP race effects when viewing the face primes. In both blocks, the degree to which N200s were greater for Whites than Blacks predicted greater bias in response latency. Said differently, as the degree to which individuals showed attentional differences in the N200 favoring Whites over Blacks increased, the degree to which Blacks were associated with threat and danger as compared to non-stereotypical negative stimuli (insects) or stereotypical positive stimuli (basketball and football) increased. As noted earlier, N200 effects have been associated with deeper processing in a range of face perception studies (Bentin & Deouell, 2000; Tanaka et al., 2006). This suggests that the degree to which Whites are processed more deeply than Blacks is associated with stronger racial stereotype activation effects showing a functional relationship between early differences in attention as a function of race during the processing of the face prime and the magnitude of stereotypical and evaluative bias elicited by the face. Our studies in this area are just beginning, but we think the results expand our understanding of the types of individual differences that moderate aspects of stereotyping and prejudice. Whereas prior investigations have focused on what could be considered higher-order construct (e.g., aspects of motivation), the study reviewed here shows that attentional effects occurring fairly early in processing also predict bias.

Behavioral regulation

We have reviewed studies examining perceptual processes associated with perceiving members of different social groups. We view these processes of interest in their own right but also because of their relevance to behavior. As just described, the study in the last section demonstrating a relationship between differences in attention to members of different racial groups and the beliefs they activate suggest a way in which these perceptual differences could relate to behavior; bigger differences in attention as a function of race could be associated with bigger differences in the semantic and evaluative associations activated by group exemplars, which in turn could influence how the perceiver behaves toward a group member. The prior studies, however, did not directly examine behavioral outcomes. Although we have not done so in the context of studies in which participants simply view pictures of faces, we have done so using a slightly more complex task. Specifically, we have employed a paradigm developed by Correll et al. (2002) in which participants see a variable number of background scenes for short but variable durations on each trial. At some point, a person appears, holding either a handgun or a similarly

[4] In this study, the N100 race effect was not significant, but it was in the same direction as prior studies, with directionally larger N100s to Blacks.

sized object such as a wallet or cell phone. Participants are instructed to "shoot" anyone holding a gun and to press another button to "not shoot" anyone who is unarmed. Corell et al. found that the target's race affected responses—participants were faster and more accurate to make shoot responses to armed Black than White targets but were faster and more accurate to make not-shoot responses to unarmed White, rather than Black, targets.

Our interest was in using ERP responses to consider behavior in this task from the perspective of behavior regulation. As a guide, we used general models of behavioral regulation that specified a two-part behavioral control system (e.g., Botvinick, Carter, Braver, Barch, & Cohen, 2001; Carter et al., 1998). One component is thought to continuously and preconsciously monitor for conflict in activated representations during ongoing information processing. The outcome of this monitoring is used by the second, regulatory part of the system, which responds to detected conflict with the implementation of higher-order cognitive control.

Making the decisions required by the shooter task is likely to activate multiple representations. In addition to the information explicitly relevant to determining whether someone is holding a weapon, the speed and ease with which race is encoded suggests that race categorization is also occurring, which in turn can activate beliefs associated with racial groups. Because of cultural stereotypes more strongly associating Blacks than Whites with concepts related to violence and threat (Correll et al., 2002), perceiving a Black individual may more strongly activate beliefs that facilitate making a shoot response. Putting these two features together suggests that conflict monitoring processing should detect less conflict between considering a shoot response and perceiving a Black target as compared to a White target.

To determine whether such differences in the operation of the behavior regulation system occur as a function of target race and whether any such differences are related to task performance, we recorded ERPs as participants completed the shooter task (Correll, Urland, & Ito, 2006). We were particularly interested in the N200 component, which is associated with the detection of conflict when behavior is successfully regulated (e.g., when correct behavior is implemented) (Nieuwenhuis, Yeung, van den Wildenberg, & Ridderinkhof, 2003).[5] The association of the N200 with conflict detection is supported by source modeling. Neuro-imaging research implicates the anterior cingulate cortex (ACC) and areas of prefrontal cortex (PFC) in conflict monitoring and cognitive control, respectively (e.g., Botvinick et al., 2001; Carter et al., 1998; MacDonald, Cohen, Stenger, & Carter, 2000), and source modeling of the N200 implicates the ACC and other areas of the PFC (Liotti, Woldorff, Perez, & Mayberg, 2000; Nieuwenhuis et al., 2003).

Behavioral results replicated past research using this task; participants were faster to shoot armed Blacks than Whites, but faster to not shoot unarmed Whites than Blacks. For the N200, responses were larger to Whites than to Blacks. This can be seen in Figure 6–4. As can also be seen, this race effect was moderated by the object being held, with the largest N200s to unarmed Whites. By contrast, N200 amplitude did not differ for Black targets as a function of the object held. This pattern suggests more sensitivity to conflicting representations when viewing Whites than Blacks in this task, and sensitivity to the presence or absence of a weapon only for Whites. We think this result is particularly interesting given a bias toward making the shoot response in this task. Participants are around 100 milliseconds faster to make the shoot than not-shoot response. In addition, a running score is displayed to participants throughout the game, with more points awarded for correct shoot than not-shoot responses. To the extent there is a bias to shoot, Whites in general and unarmed Whites in particular would be the most stereotypically incongruent with that response tendency. Thus, conflict should

[5] Although they share a similar latency and anterior scalp distribution, the N200 discussed here in the context of conflict detection (sometimes also referred to as an N2) has been treated as conceptually distinct from the N200 discussed in the context of social categorization. It is possible that they reflect a common psychological and/or neural source, but that is not yet known.

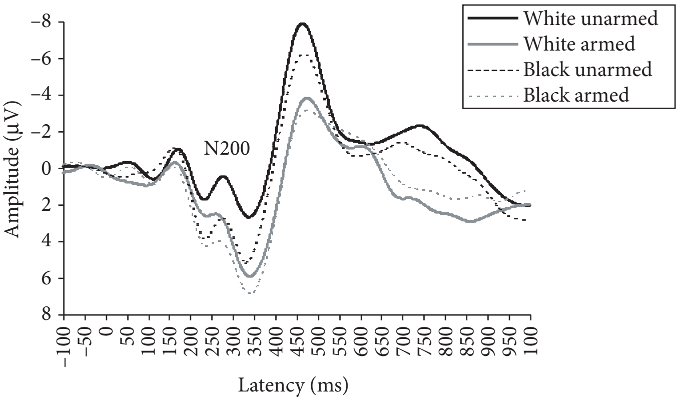


Fig. 6–4 N200 responses as a function of target race and object held, recorded during decisions to shoot. Waveforms as from the frontal scalp area (electrode Fz). Reprinted with permission from Correll, J., Urland, G.R., & Ito, T.A. (2006). Event-related potentials and the decision to shoot: The role of threat perception and cognitive control. *Journal of Experimental Social Psychology, 42*, 120–128. Copyright Elsevier.

be greatest in those conditions. Consistent with this, N200s were bigger to Whites in general and to unarmed Whites in particular.

It is also noteworthy that N200 differences correlated with behavior. The degree to which N200s to Whites exceeded those to Blacks predicted the degree of racial bias in response latencies (i.e., the degree to which participants were faster in making shoot responses to Blacks but in making not shoot responses to Whites). Said differently, neural responses indicative of a bigger race difference in conflict detection were associated with more racially biased behavior. We also found that individual differences in racial stereotypes predicted behavior; more strongly associating Blacks than Whites with violence and danger was associated with greater bias in behavior. Interestingly, we also found that racial differences in conflict monitoring mediated the relation between racial stereotypes and behavior—that is, participants who more strongly associated Blacks than Whites with violence were more biased in their behavior, and this was accounted for by stronger neural signals associated with conflict monitoring to Whites than Blacks. This mediational model can be seen in Figure 6–5. Together, these results support the application of behavior regulation models in understanding how stereotypes and prejudice affect behavior. They also suggest that perceptions of congruency between activated behavioral tendencies and beliefs differ as a function of target race and that these differences are predictive of behavior.

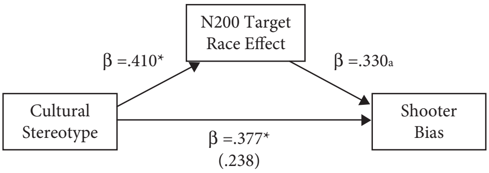


Fig. 6–5 Mediation of the relation between racial cultural stereotypes and racial bias in decisions to shoot by N200 race differences. The direct effect of the cultural stereotype on bias is significantly weaker after partialing out the ERP race effect (partial correlation is shown in parentheses). **p* < 0.05.

Conclusion

Perhaps no single stimulus has the power to convey as much about an individual as his or her face. Subtle variations in changeable features such as mouth position, eye gaze, and brow arch and more stable features like skin tone, eye brow shape, and nose width can be used to infer a wealth of social information. Among the inferences possible is categorization into social groups meaningful to the perceiver. Models of person perception have assumed that such information is processed in a relatively effortless manner (Bodenhausen & Macrae, 1998; Brewer, 1988; Fiske & Neuberg, 1990). The access to quickly occurring aspects of processing afforded by ERPs provides direct support for this. Studies assessing social categorization consistently demonstrate the ease and speed with which race and gender are processed. Race effects occur as quickly as 120 milliseconds after stimulus onset, and gender effects occur as quickly as 180 milliseconds. Moreover, this sensitivity is observed across a range of tasks, including when perceivers are attending to *(1)* another social dimension (e.g., race effects occur even when perceivers explicitly attend to gender) (Ito & Urland, 2003, Kubota & Ito, 2007), *(2)* a nonsocial cue (searching for dots on faces) (Ito & Urland, 2005), or *(3)* individual characteristics that foster person-based as opposed to category-based impressions (Ito & Urland, 2005). This pattern suggests that early perceptual aspects of social categorization are driven more by the properties of the individual being perceived than by the goals and intentions of the perceiver. They also suggest that processing manipulations that succeed in attenuating stereotyping and prejudice are likely to operate after the completion of rudimentary analyses that provide information about social category membership.

The studies we have reviewed also suggest that social category perception may be governed by different psychological processes at different points in time. As discussed earlier, the pattern of results in which White participants show larger N100s and P200s to Blacks, Asians, and males may be indicative of initial covert orienting to more novel targets and/or targets heuristically associated with a greater potential for threat. Moreover, the earliest stages of social category perception appear to reflect processing that is more gross than fine-grained, as suggested by the initial similarity in responses to in-group faces and multiracial faces that share features with the in-group (Willadsen-Jensen & Ito, 2006). Slightly later in processing, the larger N200s to Whites and females (*see also* Ito, Thompson, & Cacioppo, 2004; James, Johnstone, & Hayward, 2001) may represent greater, more individuated processing of racial in-group members and/or members of the more culturally dominant racial group, consistent with the large body of research showing that Whites and racial in-group members are spontaneously processed more deeply than other racial groups (Anthony, Cooper, & Mullen, 1992; Levin, 2000). The meaning of the larger N200s to females is less clear. Processing differences in favor of female over male targets have been obtained (Lewin & Herlitz, 2002; McKelvie, 1981; McKelvie et al., 1993; O'Toole et al., 1998), but these effects are more variable than differences in processing as a function of race.

Although important in their own right for understanding the earliest aspects of social category encoding, the attentional differences observed in response to different social groups are even more intriguing when considered in relation to the activation of stereotypes and prejudices. Studies using sequential priming paradigms to measure the implicit activation of beliefs and evaluations associated with different racial groups demonstrate that variations in ERP responses to brief presentations of faces predict the degree to which Blacks are more strongly associated with negative stereotypic content as compared to Whites (Ito & Urland, 2006; Willadsen-Jensen, Ito, & Park, 2006). We think it intriguing that differences in the way individuals attend to the faces of Blacks and Whites emerging within several hundred milliseconds after stimulus onset predict the types of beliefs and feelings that are activated by those faces. These may not be the only processing differences that contribute to the activation of stereotypes and prejudice, but these results do provide an indication of how quickly the processes that play a role in implicit bias manifest during perception.

Finally, ERP studies of stereotype regulation hold the promise of stimulating effective behavioral interventions by increasing our understanding of the neural mechanisms behind successful and unsuccessful behavior regulation. Consistent with the likelihood that cultural stereotypes make perceiving a Black than White target more congruent with responses associated with threat, we found greater evidence of conflict when participants were making simulated decisions to shoot on trials in which White targets were seen, especially when the Whites were unarmed. By contrast, conflict was weaker when perceiving Black targets, and there was no difference in conflict when the Blacks were armed or unarmed. If the detection of conflict is used to signal the need for higher-order cognitive control, the low degree of conflict on trials in which Black individuals are seen suggests an explanation for the racial bias seen in this task; the need for behavior regulation is not signaled to the same degree as on trials in which Whites are seen.

Factors such as perceivers' lack of awareness about or unwillingness to report on their reactions have challenged researchers in their attempts to understand how social category information influences behavior. The integration of tools and ideas from other levels of analysis holds promise in addressing some of these issues in new ways. We have tried to highlight, in particular, the benefits of research using ERPs in this chapter. We hope the continued development of social neuroscience will increase our understanding of the dynamic interplay between the mind and the neural processes that underlie it.

References

Amodio, D.M., Harmon-Jones, E., & Devine, P.G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink response and self-report. *Journal of Personality and Social Psychology, 84*, 738–753.

Anthony, T., Copper, C., & Mullen, B. (1992). Cross-racial facial identification: A social cognitive integration. *Personality and Social Psychology Bulletin, 18*, 296–301.

Bentin, S., & Deouell, L.Y. (2000). Structural encoding and identification in face processing: ERP evidence for separate mechanisms. *Cognitive Neuropsychology, 17*, 35–54.

Bodenhausen, G.V., & Macrae, C.N. (1998). Stereotype activation and inhibition. In R.S. Wyer, Jr. (Ed.), *Stereotype activation and inhibition* (pp. 1–52). Mahwah, NJ: Lawrence Erlbaum.

Botvinick, M.M., Carter, C.S., Braver, T.S., Barch, D.M., & Cohen, J.D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*, 624–652.

Brewer, M.C. (1988). A dual process model of impression formation. In R. Wyer & T. Scrull (Eds.), *Advances in social cognition* (Vol. 1, pp. 1–36). Hillsdale, NJ: Erlbaum.

Carter, C.S., Braver, T.S., Barch, D.M., Botvinick, M.M., Noll, D., & Cohen, J.D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science, 280*, 747–749.

Correll, J., Park, B., Judd, C.M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83*, 1314–1329.

Correll, J., Urland, G.R., & Ito, T.A. (2006). Event-related potentials and the decision to shoot: The role of threat and cognitive control. *Journal of Experimental Social Psychology, 42*, 120–128.

Czigler, I., & Geczy, I. (1996). Event-related potential correlates of color selection and lexical decision: Hierarchical processing or late selection? *International Journal of Psychophysiology, 22*, 67–84.

Donchin, E. (1981). Surprise!…Surprise? *Psychophysiology, 18*, 493–513.

Dovidio, J.F., Kawakami, K., Johnson, K., Johnson, C., & Hayward, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental and Social Psychology, 33*, 510–540.

Eimer, M. (1997). An event-related potential (ERP) study of transient and sustained visual attention to color and form. *Biological Psychology, 44*, 143–160.

Fabiani, M., Gratton, G., & Coles, M.G.H. (2000). Event-related brain potentials. In J.T. Cacioppo, L.G. Tassinary, & G.G Berntson (Eds.), *Handbook of psychophysiology* (2nd ed., pp 53–84). Cambridge, UK: Cambridge University Press.

Fazio, R.H., Jackson, J.R., Dunton, B.C., & Williams, C.J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69*, 1013–1027.

Fiske, S.T., & Neuberg, S.L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology, 23*, 1–73.

Gehring, W.J., Gratton, G., Coles, M.G.H., & Donchin, E. (1992). Probability effects on stimulus evaluation and response processes. *Journal of Experimental Psychology: Human Perception & Performance, 18*, 198–216.

Greenwald, A.G., McGhee, D.E., & Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

Hausmann, L.R. & Ryan, C.S. (2004). Effects of external and internal motivation to control prejudice on implicit prejudice: The mediating role of efforts to control prejudiced responses. *Basic and Applied Social Psychology, 26*, 215–225.

Ito, T.A., & Bartholow, B.D. (2009). The neural correlates of race. *Trends in Cognitive Sciences, 13*, 524–531.

Ito, T.A., Thompson, E., & Cacioppo, J.T. (2004). Tracking the timecourse of social perception: The effects of racial cues on event-related brain potentials. *Personality and Social Psychology Bulletin, 30*, 1267–1280.

Ito, T.A., & Urland, G.R. (2003). Race and gender on the brain: Electrocortical measures of attention to race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology, 85*, 616–626.

Ito, T.A., & Urland, G.R. (2005). The influence of processing objectives on the perception of faces: An ERP study of race and gender perception. *Cognitive, Affective, and Behavioral Neuroscience, 5*, 21–36.

Ito, T.A., & Urland, G.R. (2006). *The functional significance of differences in attention to Blacks and Whites*. Unpublished data.

James, M.S., Johnstone, S.J., Hayward, W.G. (2001). Event-related potentials, configural encoding, and feature-based encoding in face recognition. *Journal of Psychophysiology, 15*, 275–285.

Judd, C.M., Blair, I.V., & Chapleau, K.M. (2004). Automatic stereotypes versus automatic prejudice: Sorting out the possibilities in the Payne (2001) weapon paradigm. *Journal of Experimental Social Psychology, 40*, 75–81.

Kenemans, J.L., Kok, A., & Smulders, F.T.Y. (1993). Event-related potentials to conjunctions of spatial frequency and orientation as a function of stimulus parameters and response requirements. *Electroencephalography and Clinical Neurophysiology, 88*, 51–63.

Kubota, J.T., & Ito, T.A. (2007). Multiple cues in social perception: The time course of processing race and facial expression. *Journal of Experimental Social Psychology, 43*, 738–752.

Levin, D.T. (2000). Race as a visual feature: Using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology: General, 129*, 559–574.

Lewin, C., & Herlitz, A. (2002). Sex differences in face recognition-women's faces make the difference. *Brain and Cognition, 50*, 121–128.

Liotti, M., Woldorff, M.G., Perez, R., & Mayberg, H.S. (2000). An ERP study of the temporal course of Stroop color-word interference effect. *Neuropsychologia, 38*, 701–711.

MacDonald III, A.W., Cohen, J.D., Stenger, V.A., & Carter, C.S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science, 288*, 1835–1838.

Macrae, C.N., Bodenhausen, G.V., Milne, A.B., Thorn, T.M.J., & Castelli, L. (1997). On the activation of social stereotypes: The moderating role of processing objectives. *Journal of Experimental Social Psychology, 33*, 471–489.

McKelvie, S.J. (1981). Sex differences in memory for faces. *The Journal of Psychology, 107*, 109–125.

McKelvie, S.J., Standing, L., St. Jean, D., & Law, J. (1993). Gender differences in recognition memory for faces and cars: Evidence for the interest hypothesis. *Bulletin of the Psychonomic Society, 31*, 447–448.

Naatanen, R., & Gaillard, A.W.K. (1983). The orientating reflex and the N2 deflection of the event-related potential (ERP). In A.W.K. Gaillard & W. Ritter (Eds.), *Tutorials in ERP Research: Endogenous Components* (pp. 119–141). New York: North-Holland Publishing Company.

Nieuwenhuis, S., Yeung, N., Van Den Wildenberg, W., & Ridderinkhof, K.R. (2003). Electrophysiological correlates of anterior cingulate function in a go/no-go task: Effects of response conflict and trial type frequency. *Cognitive, Affective & Behavioral Neuroscience, 3*, 17–26.

O'Toole, A.J., Deffenbacher, K.A., Valentin, D., McKee, K., & Abdi, H. (1998). The perception of face gender: The role of stimulus structure in recognition and classification. *Memory and Cognition, 26*, 146–160.

Payne, B.K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81*, 181–192.

Pizzagalli, D., Regard, M., & Lehmann, D. (1999). Rapid emotional face processing in the human right and left brain hemispheres: An ERP study. *NeuroReport, 10*, 2691–2698.

Sinclair, L., & Kunda, Z. (1999). Reactions to a black professional: Motivated inhibition and activation of conflicting stereotypes. *Journal of Personality and Social Psychology, 77*, 885–904.

Smith, N.K., Cacioppo, J.T., Larsen, J.T., & Chartrand, T.L. (2003). May I have your attention, please: Electrocortical responses to positive and negative stimuli. *Neuropsychologia, 41*, 171–183.

Stroessner, S. J. (1996), Social categorization by race or sex: Effects of perceived non-normalcy on response times. *Social Cognition, 14*, 247–276.

Tanaka, J.W, Curran, T., Porterfield, A.L., & Collins, D. (2006). Activation of preexisting and acquired face representations: The N250 event-related potential as an index of face familiarity. *Journal of Cognitive Neuroscience, 18*, 1488–1497.

Wheeler, M.E., & Fiske, S.T. (2006). Controlling racial prejudice: Social cognitive goals affect amygdala and stereotype activation. *Psychological Science, 16*, 56–63.

Wijers, A., Mulder, G., Okita, T., Mulder, L.J.M., & Scheffers, M. (1989). Attention to color: An analysis of selection, controlled search, and motor activation, using event-related potentials. *Psychophysiology, 26*, 89–109.

Willadsen-Jensen, E.C., & Ito, T.A. (2006). Ambiguity and the timecourse of racial perception. *Social Cognition, 24*, 580–606.

Zarate, M.A., Bonilla, S., & Luevano, M. (1995). Ethnic influences on exemplar retrieval and stereotyping. *Social Cognition, 13*, 145–162.

Zarate, M.A., & Smith, E.R. (1990). Person categorization and stereotyping. *Social Cognition, 8*, 161–185.

CHAPTER 7
Self-Regulation in Intergroup Relations: A Social Neuroscience Framework

David M. Amodio

For many White Americans, an interracial interaction constitutes a regulatory challenge. Despite steady declines in self-reported prejudices over the past 60 years (Schuman, Steeh, Bobo, & Krysan, 1997), subtle forms of bias persist that may work their way unintentionally into intergroup behaviors (Devine, 1989; McConahay & Hough, 1976; Sears & Henry, 2005; Word, Zanna, & Cooper, 1974). Given the notion that America was founded on the principles of freedom and equality, the persisting racial biases held by most White Americans create a fundamental ideological conflict. Myrdall (1944) referred to this conflict as the "American Dilemma" more than 60 years ago, and Allport (1954) wrote about it at length in his seminal book *The Nature of Prejudice*.

A large body of social psychological research reveals that Myrdall's American Dilemma is a contemporary concern (Fiske, 1998; Gaertner & Dovidio, 1986; Katz & Hass, 1988). This work demonstrates that most Americans possess *implicit* forms of racial bias that can influence behavior without one's intention or awareness (Devine, 1989). Implicit processes are believed to operate automatically in thoughts, emotions, and behaviors, whereas explicit processes involve deliberation and awareness and are associated with intentional (e.g., egalitarian) responses (Greenwald & Banaji, 1995). Although implicit and explicit processes typically work in concert to orchestrate adaptive behavior, racial prejudice is a domain in which many people hold explicit intentions to respond without prejudice that contradict their implicit biases. Hence, to respond in line with intentions, regulatory processes are often required to override implicit biases. In this chapter, I describe a social neuroscience framework of the processes through which racial biases are activated and controlled. In what follows, I begin by reviewing the prevalent dual-process model of prejudice and stereotyping and then describe some ways in which recent social neuroscience research has extended our understanding of the activation and regulation of intergroup bias.

The dual-process approach to racial bias

The topic of intergroup relations has been a central theme of social psychology for decades, yet the theoretical approach guiding investigations of this topic has varied considerably. The focus on automatic versus controlled components of race bias is a relatively recent development, borne out of the application of cognitive psychological theories to social psychological issues (Fiske & Taylor, 1984). In an early demonstration of dissociable automatic and controlled components of prejudice and stereotyping, Devine (1989) applied a dual-process model that conceived of automatic biases as learned associations stored in a parallel-distributed semantic network. This model, which borrowed from theories in cognitive psychology (Shiffrin & Schneider, 1977; McClelland & Rumelhardt, 1985), suggested that implicit biases were learned through repeated exposure to associations between Black Americans and stereotypic and/or negatively valenced concepts. On the other hand, the controlled component represents individuals' consciously held beliefs and intentions. Devine demonstrated that racial biases could be activated automatically and affect one's interpersonal judgments regardless of one's explicit racial attitudes. That is, automatic stereotyping effects were shown to be dissociable from controlled, belief-based intentions. Subsequent theoretical treatments of racial bias, as with attitudes, have been generally consistent with this dissociation model (e.g., Bodenhausen & Macrae, 1998; Fazio, 1990; Fiske & Neuberg, 1990; Greenwald & Banaji, 1995; Smith & DeCoster, 2000; Wilson, Lindsay, & Schooler, 2000), offering refinements, clarifications, and applications to specialized domains.

Importantly, for the present concerns, social cognitive models of automatic and controlled processing posit two modes of processing. In the context of race bias, they assume that implicit stereotypes (i.e., traits ascribed to Black Americans) and implicit evaluations of Black people are processed through a single automatic mode of processing. In the dual-process framework, implicit stereotypes refer to linked representations of stigmatized group members (e.g., Blacks) and semantic concepts (e.g., traits, such as *hostile*). Implicit evaluative bias is thought to represent the net valence of these semantic links (e.g., Taylor & Falcone, 1982; for reviews, *see* Park & Judd, 2005; Dovidio, Brigham, Johnson, & Gaertner, 1996). Accordingly, this model assumes that implicit stereotyping and prejudice arise from the same mechanism, such that they are learned, activated, and unlearned in the same way. Similarly, dual-process models in social cognition generally assume that control reflects a single process. Although some theorists have suggested that control involves several deliberative steps, such as detecting the presence of bias, determining its magnitude and direction, and then adjusting one's response in kind (Wegener & Petty, 1997; Wilson & Brekke, 1994), this set of operations is thought to be supported by a single underlying resource.

Dual-process models in social cognition have been enormously useful for explaining a wide range of behaviors, and the dual-process framework continues to be the driving theoretical force in contemporary social psychology (Chaiken & Trope, 1999). Nevertheless, accumulating evidence from behavioral and cognitive neuroscience research suggests that the dual-process model may be due for some additional fine tuning and expansion (cf. Conrey et al., 2006). First, there are several phenomena that the traditional dual-process model does not explain well. For example, why do some egalitarian individuals fail to control expressions of bias more than others, despite similar motivations and effort to control? Why is control usually more successful when it is motivated for internal (personal) reasons rather than external (normative) reasons? In addition, the dual-process model assumes that implicit processes can have profound effects on behavior, yet evidence for such effects in the race bias literature is rather scant. Secondly, as our understanding of the brain's role in regulating social behavior evolves, it is becoming progressively clearer that the brain is not organized according to two simple processes. General systems for self-regulation reflect the coordinated activity of multiple underlying systems, ranging from more automatic to more controlled. As such, much of the recent social neuroscience research on racial bias may be seen as elucidating important subcomponents of automaticity and control in an effort to refine—and ultimately revise—the dual-process model of social cognition.

The social neuroscience approach

Social neuroscience refers broadly to the integrated study of the brain and social processes. Although the term *social neuroscience* is currently used to describe a wide range of research in humans and animals, I use it here to refer more specifically to the merger of social psychology and neuroscience (Cacioppo & Bernston, 1992; Ochsner & Lieberman, 2001). From this perspective, neuroscientific models of brain

Table 7–1 Processes Involved in Intergroup Bias and Their Associated Cognitive Functions and Neural Correlates

Role in intergroup bias	Cognitive process	Candidate structure(s)
Implicit prejudice	Classical fear conditioning; arousal; vigilance	Amygdala
Implicit stereotyping	Conceptual priming	Temporal cortex & left lPFC
Detecting bias/internal cues for regulation	Conflict monitoring	Dorsal ACC
Detecting external cues for engaging control	Mentalizing; regulating behavior to external social cues	mPFC, rostral ACC
Inhibition of implicit bias	Response inhibition	Ventral lPFC
Implementation of an intended response	Regulative control	Dorsal lPFC

Note: lPFC = lateral prefrontal cortex; mPFC = medial prefrontal cortex; ACC = anterior cingulate cortex. BA = Brodmann's Area

organization and function may be used to clarify and refine models of social cognition and behavior, just as an appreciation of social structures, goals, motivations, and relationships may be used to clarify functions of the brain. A major goal of this chapter is to highlight the ways in which the social neuroscience approach has helped to unpack the processes involved in the activation and regulation of racial bias. Table 7–1 lists a set of cognitive processes that have been identified in the social neuroscience literature on race bias to date. Considered together, they suggest important theoretical advances beyond the basic dual-process framework.

Social neuroscience contributions to theories of implicit race bias

The first questions about intergroup processes that were addressed from a social neuroscience approach focused primarily on mechanisms of implicit bias (*see* Eberhardt, 2005, for a review). Initial inquiries concerned the role of the amygdala in implicit prejudice. The amygdala refers to a small collection of nuclei located bilaterally in the medial temporal lobes that has been implicated in classical fear conditioning and emotional learning and, more generally, in the processes of arousal and vigilance (Fig. 7–1; Anderson et al., 2003; Cunningham, Raye, & Johnson, 2004; Fendt & Fanselow, 1999; Whalen, 1998). This structure is part of the brain's "rapid response" system that is activated and expressed within milliseconds of a potentially threatening event (LeDoux, 1992). This mode of rapid response is made possible by the short distance incoming sensory information must travel to the amygdala (Davis, 1992; but *see* Pessoa, McKenna, Gutierrez, & Ungerleider, 2002). For example,

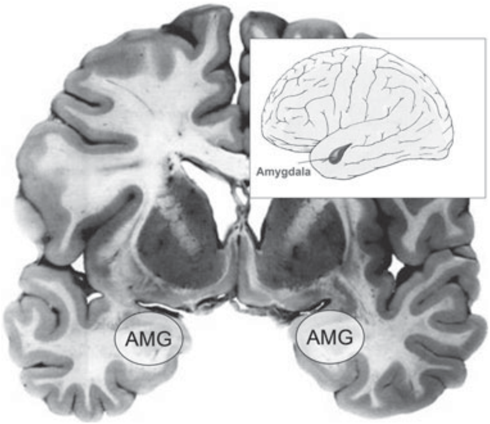


Fig. 7–1 The amygdala (AMG) comprises a set of small nuclei and is located bilaterally in the medial temporal lobe, as shown in the coronal brain slice. The inset shows the position of the left amygdala as it would appear within the temporal lobe when viewed from the side.

visual and auditory information is relayed by the thalamus via a single synapse to the amygdala for initial processing, whereas slower, more elaborative processing continues throughout the cortex (LeDoux, Cicchetti, Xagoraris, & Romanski, 1990). This "quick and dirty" detection quality provides an important mechanism for survival, but at the same time, the amygdala's response to particular stimuli is relatively resistant to change and prone to generalization (Bouton, 1994). The amygdala and its associated subcortical structures orchestrate adaptive behavioral responses, such as inhibition and approach/withdrawal, through multiple connections to brain stem structures, the thalamus, hypothalamus, basal ganglia, and medial prefrontal cortex (mPFC; Davis & Whalen, 1998).

Initial fMRI studies provided preliminary evidence for the link between implicit prejudice and the amygdala (Hart et al., 2000; Phelps et al., 2000). Phelps et al. (2000) found that the difference in amygdala activity in response to viewing Black vs. White faces was related to two other "implicit" measures of bias: *(1)* the Implicit Associations Test (IAT; Greenwald, McGhee, & Schwartz, 1998), a behavioral task assessing evaluative associations with Black versus White faces, and *(2)* the startle–eyeblink response to Black versus White faces, which provides a physiological index of amygdala activity (Lang, Bradley, & Cuthbert, 1990). However, Phelps et al. (2000) did not find a main-effect difference in amygdala to Black versus White faces. Similarly, Hart et al. did not observe a race effect on amygdala activity during the first block of trials but found that amygdala responses to Black faces habituated more slowly than responses to White faces in later trials. Nevertheless, these exciting discoveries threw the door open for researchers interested in the neural underpinnings of implicit and explicit racial biases.

Around the same time, social psychologists were already beginning to apply models of animal and human neuroscience to address some elusive theoretical questions about implicit race bias: Is implicit prejudice an emotional process that is fundamentally different from other types of associative learning (e.g., semantic associations)? Can individual differences in amygdala responses to race explain why some self-avowed egalitarians show more implicit evaluative bias than others on behavioral measures? My colleagues and I (Amodio, Harmon-Jones, & Devine, 2003) began to address these questions by comparing participants' startle–eyeblink responses to Black versus White faces. In our study, we identified participants who responded without prejudice for either personal reasons or for normative reasons (i.e., to avoid social disapproval) or for a combination of these two reasons. Our past work had shown that although all people motivated to respond without prejudice for personal reasons report positive *explicit* attitudes toward Black people, those who also tend to worry about social pressures showed higher levels of implicit evaluative bias than those who do not (Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002). That is, individuals with a combination of motivations to respond without prejudice appeared to be "conflicted"—they are explicitly egalitarian yet implicitly biased.

We wanted to know whether the "conflict" pattern could be explained by amygdala-based processes of race bias. To test this hypothesis we examined amygdala responses to Black versus White faces among participants who were motivated to respond without prejudice for personal or normative reasons, or for both reasons. We chose to use startle–eyeblink as an index of amygdala activity because it is capable of making temporally precise measurements, compared with the relatively slow hemodynamic response assessed by fMRI. Because the automatic activation of race bias is known to occur within a few hundred milliseconds following exposure to a target of bias, precise timing was critical. The startle–eyeblink measure works on the principle that people are more easily startled to a loud noise when they are in an aversive state and less easily startled when they are in an appetitive state, compared with baseline (Lang et al., 1990). For example, if you were sitting in a dark theater watching a horror movie and someone snapped their fingers behind your head, you'd likely jump in your seat. But if you were watching a love scene, which would presumably elicit an appetitive state, you would probably barely notice the finger snap.

The startle response is reflexive—that is, hard-wired and very difficult to suppress—and animal research has determined that this effect is modulated by the amygdala (Davis, 2006). One component of the whole-body startle reflex is the defensive eyeblink, and in humans, the magnitude of one's startle–eyeblink response can be assessed by measuring the contraction of the muscle surrounding the eye (orbicularis oculi) using surface electrodes. An important advantage of the startle–eyeblink index of amygdala activity is that it specifically indexes activity in the central nucleus of the amygdala—the region involved in fear processing. fMRI measures cannot clearly differentiate activity in the central nucleus from activity in other regions, such as the basal nucleus, which is involved in instrumental approach-related responses to positive as well as negative stimuli (Holland & Gallagher, 1999). Thus, the startle–eyeblink method is best suited for assessing amygdala activity associated with fear-related processing.

In our study, people viewed faces of Black, White, and Asian males and occasionally heard a *startle probe*, which was a very loud (96 dB) and short (50 ms) blast of white noise, delivered through headphones. The idea was that if seeing a picture of a Black person's face elicited negative affect, the probe should elicit a stronger blink when it occurs during the viewing of a Black person's versus a White person's face. We found that, as suspected, the "conflicted" participants showed larger startle–eyeblink responses to Black (vs. White) faces than those who responded without prejudice for purely personal reasons. This pattern emerged in eyeblink responses occurring as early as 400 milliseconds following the onset of a face picture and was very strongly pronounced in responses occurring 4000 milliseconds into picture-viewing. Hence, we concluded that an important source of the "conflict" in these individuals was the automatic activation of amygdala-based affective associations with Black people. More broadly, this study was the first to demonstrate a significant increase in an index of amygdala activity in response to Black faces compared with White faces across participant groups.

Subsequent research has shown similar effects using fMRI. Cunningham et al. (2004) circumvented the issue of slow timing in fMRI by presenting pictures of Black and White faces to participants for only 30 milliseconds, immediately followed by a colored shape, so that participants were not aware of having seen a face. Participants' task was to classify the shape as appearing on the left or right side of the monitor. The authors observed greater amygdala activity associated with the presentation of Black faces than White faces. Whereas Amodio et al. (2003) established the automaticity of their effect by measuring amygdala activity within milliseconds of its activation, before deliberative control could be engaged, Cunningham et al. (2004) established automaticity by using very fast presentations of faces that were intended to preclude control. Additional studies have further corroborated these findings (e.g., Lieberman, Hariri, Jarcho, Eisenberger, & Bookheimer, 2005; Wheeler & Fiske, 2005), which taken together suggest that our understanding of implicit prejudice may be enhanced by considering its relation to other functions ascribed to the amygdala, such as classical fear conditioning, arousal, and vigilance.

Different mechanisms for implicit stereotyping vs. implicit prejudice?

Much research has examined the neural basis of implicit evaluative bias, but few studies have examined implicit stereotypes. In the social psychological literature, stereotypes refer to the "cognitive" component of racial bias and typically correspond to sets of trait attributes ascribed to a social group (Fiske, 1998). The notion of independent affective and semantic components of person perception has a long history in social psychology (Abelson, Kinder, Peters, & Fiske, 1982; Allport, 1954; Park & Judd, 2004; Zajonc, 1980), and previous work has noted that this distinction may be represented in implicit processes (Dovidio, Evans, & Tyler, 1986; Greenwald & Banaji, 1995; Kawakami, Dion, & Dovidio, 1998; Rudman, Ashmore, & Gary, 2001; Wittenbrink, Judd, & Park, 1997, 2001). However, research has not yet advanced a theoretical framework to describe the distinct mechanisms of implicit prejudice versus implicit stereotyping or the nature of their relationship. In this section, I describe how a social neuroscience approach may be useful for elucidating such a framework.

Several social neuroscience studies have suggested that implicit prejudice involves basic neural systems for detecting threat and initiating rapid behavioral responses (e.g., Amodio et al., 2003; Cunningham et al., 2004; Lieberman et al, 2005; Wheeler & Fiske, 2005). On the other hand, implicit stereotypes involve relations between symbolic representations of abstract concepts and may function to bias judgments and to organize behavior (Allport, 1954; Dovidio et al., 1996; Dovidio, Esses, Beach, & Gaertner, 2004; Fiske, 1992). The ability to form conceptual representations is a higher-order cognitive capacity, and neuroscience research on conceptual priming suggests this type of processing is associated with regions of neocortex in the temporal lobe and posterior left prefrontal cortex (PFC, Fig. 7–2; Gabrieli, 1998). Importantly, implicit conceptual associations primarily rely on neocortex, whereas implicit affective associations rely on subcortical structures such as the amygdala (Squire & Zola, 1996). On the basis of the neuroscience literature, I have argued that implicit prejudice and implicit stereotypes reflect distinct underlying memory systems that are expressed through different sets of response channels (Amodio, 2008). For example, Amodio and Devine (2006) demonstrated that IAT measures of evaluative racial associations (in the absence of stereotype content) and stereotype associations (in the absence of evaluative content) concerning Black versus White faces were uncorrelated. We also showed that levels of implicit prejudice uniquely predicted participants' affective judgments of Black people, as well as the distance that they decided to sit from a Black student's belongings in a row of chairs while waiting to interact with him. By contrast, participants' levels of implicit stereotyping predicted the extent to which they formed stereotype-consistent impressions of a Black student, as well as their expectations that a Black activity partner would perform in a stereotypic way on tests of academic skills and sports trivia. These behavioral findings are consistent with the idea that at implicit levels of processing, prejudice, and stereotyping reflect different mechanisms.

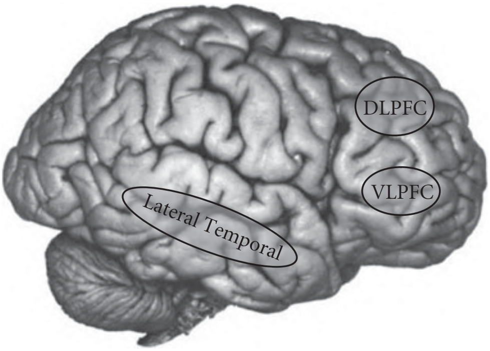


Fig. 7–2 Lateral view indicating temporal lobe and prefrontal cortex (PFC). Regions of dorsal and ventral lateral prefrontal cortex (dlPFC and vlPFC) have been associated with the controlled processing, and left PFC has been linked to semantic processes that play a role in stereotyping.

Despite the attention given to identifying the neural mechanisms of implicit prejudice, there is not currently any published evidence suggesting a neural substrate for racial stereotypes. Some research using event-related potentials (ERPs) has examined the time-course of brain activity associated with stereotype processing (Bartholow et al., 2001) and with conceptual categorizations made on the basis of race (Amodio, 2010; Correll, Urland, & Ito, 2006; Ito & Cacioppo, 2001; Ito & Urland, 2003). In these studies, ERP measures of attentional processes suggest that stereotyping and categorical processes are evident within 200 milliseconds of the presentation of a face. However, additional research is needed to examine neural substrates specific to implicit stereotyping *per se*, such as neural systems involved in implicit semantic memory as suggested by Amodio and Devine (2006). By elucidating these potentially dissociable mechanisms of implicit race bias, dual-process accounts may be refined to account for how implicit prejudice and stereotyping are acquired via different modes of operation, how they are expressed through different response channels, and how they may be extinguished through different procedures.

One important area of implicit social cognition that has not yet been explored concerns motor skill learning procedural memory (e.g., habit or skill learning)—the process through which repeated motor associations become automatized independently of explicit knowledge (Knowlton, Mangels, & Squire, 1996). Neural mechanisms for procedural memory have been dissociated from those of explicit knowledge, such that procedural memory is associated with activity in the basal ganglia, whereas explicit knowledge is associated with activation of the hippocampus (Foerde, Poldrack, & Knowlton, 2006). As prejudice researchers begin to focus more on the behavioral sequelae of implicit racial biases, it will be important to consider how implicit systems for motor learning interact with systems for implicit affective and semantic associations. For example, mechanisms of skill learning are especially important given that most social psychological assessments of implicit bias are made using behavioral tasks that involve repeated motor associations. Although the mechanism of motor skill learning may not be very informative to the constructs of implicit evaluation and stereotyping *per se*, it will likely inform interpretations of participants' performance on behavioral measures of implicit race bias.

Neural mechanisms for control in the context of race

Once implicit racial biases are activated, how are they controlled? To override any unwanted influences of bias on one's behavior, one must engage self-regulation processes that involve cognitive control (Devine, 1989; Macrae et al., 1994; Monteith, 1993; Gilbert & Hixon, 1991; Fazio, 1990). Past social psychological models of control have focused on deliberative aspects of self-regulation (Ajzen & Fishbein, 2000; Wegener & Petty; 1997; Wilson & Brekke, 1994). These models assume that control is initiated intentionally—that is, a person must consciously notice the presence of a biasing influence and then decide to take compensatory measures. In general, such models describe self-regulatory processes that operate when responses are be made deliberatively and without time constraints, such as in verbal or written self-reports (Fazio, 1999). However, many behaviors are relatively nondeliberative and are enacted under time constraint, such as during the rapid exchange in an animated social interaction or when making snap judgments about a person (e.g, Willis & Todorov, 2006). Social psychological models generally assume that such nondeliberative behaviors are driven by automatic processes, yet the complexity of such behaviors suggests that a considerable degree of regulation may be at play. The social neuroscience approach to understanding control in the context of race bias has been useful for unpacking aspects of self-regulation associated with more deliberative versus less deliberative processes.

Detecting bias and engaging control

In traditional models of self-regulation, control begins when an individual detects that bias is present. But what draws one's attention to the presence of bias in the first place? Although not explicitly stated, most dual-process theories in social and cognitive psychology assume a homuncular initiator of control—the idea being that a "little man" inside our head tells us when control is needed. As a solution to the "homunculus" problem of control, Botvinick, Braver, Barch, Carter, and Cohen (2001) proposed that there are independent cognitive systems for *(1)* determining when control is needed and *(2)* implementing intended behavior. In this model, it is assumed that several different response tendencies are often simultaneously activated in the brain in response to both internal and external cues. When two or more activated tendencies imply different behavioral responses, there is conflict in the system. The first component of the Botvinick et al.'s (2001) model of control monitors the degree of conflict in this system. As the degree of conflict rises, a second, regulatory system is engaged to orchestrate deliberative forms of control. Across several fMRI and ERP studies, conflict monitoring has been associated with activity of the dorsal anterior cingulate cortex (dACC) and the regulatory

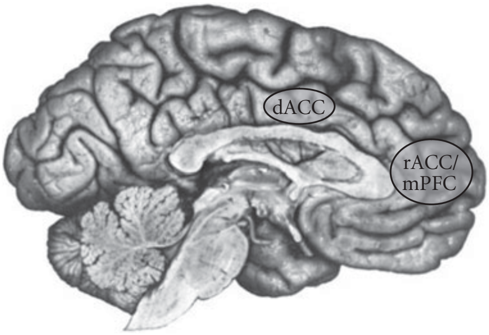


Fig. 7–3 Medial view of the brain illustrating the dorsal anterior cingulate cortex (dACC), medial prefrontal cortex (mPFC). The shaded areas of these regions are those typically activated in studies of prejudice control and person perception described in the text.

system has been linked to activity in the dorsolateral (dl)PFC (Fig. 7–3; Botvinick et al., 1999; Carter et al., 1998; van Veen & Carter, 2002).

Social neuroscientists interested in prejudice have applied the conflict monitoring model to address mechanisms of prejudice control (Amodio et al., 2004; Richeson et al., 2003). Previous models of prejudice control posit that failures to respond without prejudice result from a person's failure to override a prejudiced response, because of a lack of motivation and/or cognitive resources. By contrast, the conflict monitoring model suggests that failures to control bias might result because conflict was not detected in the first place. To test this hypothesis directly, my colleagues and I (Amodio et al., 2004) measured ERPs as participants completed the weapons identification task (Payne, 2001). In each trial in this task, a Black or White face prime is presented briefly (200 ms), followed by a target picture of either a handgun or handtool. Participants are instructed to categorize the target as a gun or tool irrespective of the prime. Previous research has shown that the presentation of a Black face facilitates the identification of guns and interferes with the identification of tools (Payne, 2001). That is, Black faces activate a prepotent stereotypic association with guns, and participants often fail to inhibit this automatic tendency, such that they erroneously identify tools as "guns" after seeing a Black face. Our research focused on the role of the ACC in response control on this task. Past work has shown that a specific component of the ERP called the error-related negativity (ERN) indexes activity of the ACC related to conflict monitoring (Gerhing et al., 1993; van Veen & Carter, 2002; Yeung, Botvinick, & Cohen, 2004). Therefore, we measured the amplitude of the ERN wave when participants failed versus succeeded in controlling their automatic tendency to classify tools as guns following a Black face. By using an ERP measure of ACC activity, we could examine changes in neural activity on the order of milliseconds and thus study the timing of the conflict monitoring process as it unfolded during the course of a response.

Despite their motivation to respond without bias, participants in our study made a disproportionate number of errors on trials that required stereotype inhibition. In other words, when a Black face prime was followed by a tool, participants had trouble overriding their automatic tendency to stereotype and often pressed "gun" erroneously. Nevertheless, when participants made this type of error, their ERN responses were larger than when they made errors on other types of trials, suggesting that at some level of processing, their conflict-monitoring systems were detecting a heightened degree of conflict caused by the unwanted stereotypic response tendency (Fig. 7–4). When the intended (i.e., correct) response was congruent with the automatic tendency, such as when Black face primes were followed by pictures of guns, ERN amplitudes were relatively low. These findings demonstrated a dissociation between conflict-monitoring and regulatory aspects of control in the context of race bias, providing evidence that prejudice control is a multicomponent process and that the detection of bias does not require deliberative processing (as suggested by previous social psychological models of control). The pattern of ERN responses was replicated by another ERP component linked to the ACC, the N2, that occurs approximately 100 to 200 milliseconds before a successfully controlled response. These N2 results revealed that conflict-monitoring levels were also higher just prior to the successful control of automatic stereotype effects.

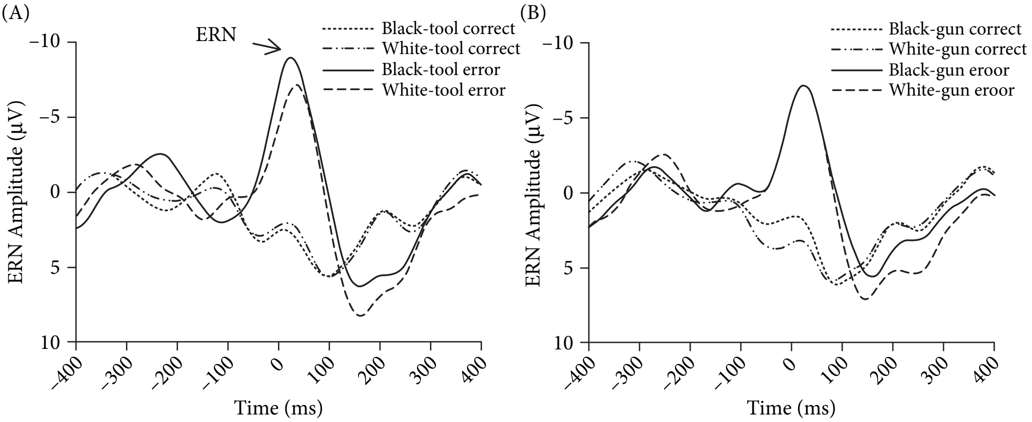


Fig. 7–4 Response-locked event-related potential waveforms for correct and incorrect tool (A) and gun (B) trials as a function of race of face. The larger error-related negativity (ERN) elicited on Black-tool trials reflects the heightened activity of the conflict-monitoring system when an automatic stereotyping tendency conflicts with subjects' intention to make correctly categorize the target as "tool." Zero indicates the time of response.

Finally, we found that the magnitude of participants' ERN response on trials that required stereotype inhibition was strongly correlated with behavioral estimates of controlled processing (derived using the process-dissociation procedure; Payne, 2001; Jacoby, 1991), as well as behavioral accuracy on trials requiring stereotype inhibition. That is, participants with more sensitive conflict-monitoring systems were generally better at inhibiting stereotypes throughout the task.

The role of conflict-related ACC activity in prejudice control and its relation to lower levels of race-biased behavior has since been replicated in subsequent ERP research (Amodio, Devine, & Harmon-Jones, 2008; Amodio, Kubota, Harmon-Jones, & Devine, 2006). Although fMRI studies have not yet shown a relationship between conflict-related ACC activity and behavioral control of race bias, some research has shown that simply viewing faces of Black individuals elicits greater ACC (and PFC) activity compared with viewing faces of White individuals (Cunningham et al., 2004; Richeson et al., 2003). Future research is needed to determine whether activations elicited by viewing faces might be related to controlled processing and the regulation of responses to race.

Explaining individual differences in the ability to effectively regulate racial responses

For the social psychologist, the primary appeal of the social neuroscience approach is that it promises to illuminate difficult social psychological questions from new angles. Having identified conflict monitoring as an important component in the regulation of prejudice, the next step was to apply the conflict monitoring framework to address phenomena that have been difficult to explain with more traditional models of control. One such phenomenon is the oft-observed finding that some egalitarian individuals have difficulty regulating their behavioral expressions of bias, whereas others who report equally egalitarian attitudes are more effective (Amodio et al., 2003; Devine et al., 2002; Devine et al., 1991; Monteith et al., 1993). My colleagues and I have hypothesized that variability in egalitarians' ability to inhibit expressions of automatic race bias may relate to the sensitivity of their conflict monitoring systems. As described above, previous research suggests that these "good" and "poor" regulators of bias can be identified by their motivations to respond without prejudice. Among individuals motivated to respond without prejudice for

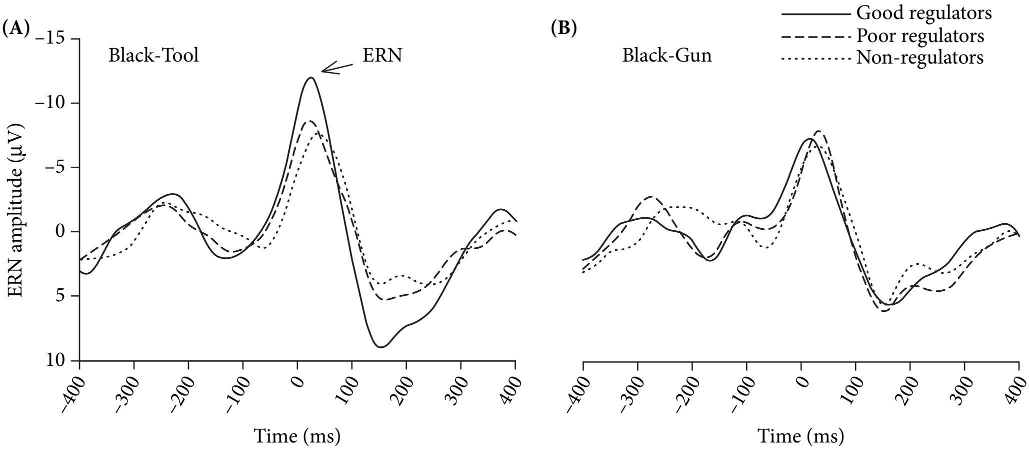


Fig. 7–5 Mean error-related negativity (ERN) amplitudes associated with Black-tool (A) and Black-gun trials (B) as a function of regulation group. Good regulators showed larger ERNs only on trials requiring the inhibition of a stereotype-based response. ERNs did not vary by group when responses did not require stereotype inhibition. Zero represents the time of response.

primarily personal (internal) reasons, those who are also concerned about external social pressures tend to express greater bias on behavioral and physiological measures than those who are not concerned with external pressures (Amodio et al., 2003; Devine et al., 2002).

To test the hypothesis that the ability to effectively inhibit race bias among egalitarian individuals is related to conflict-monitoring, we recruited participants matching the "good regulator" and "poor regulator" profiles on the basis of their internal and external motivations to respond without prejudice (Plant & Devine, 1998) and recorded ERPs as they completed the weapons identification task. Both groups showed equivalent (and significant) levels of automatic stereotyping in their behavior on the task (although these groups are known to differ in levels of implicit evaluation; Amodio et al., 2003; Devine et al., 2002). Both groups also reported positive explicit attitudes toward Black people (Brigham, 1993), and thus both needed to inhibit automatic stereotypes to respond in line with explicit beliefs. But as suggested by past findings (e.g., Devine et al., 2002), good regulators exhibited greater controlled processing on the task, as indicated by process—dissociation estimates, and responded more accurately on trials requiring the inhibition of stereotypes (i.e., Black-tool trials) than poor regulators. Was this effect to the result of differences in conflict monitoring? Indeed, good regulators showed significantly larger ERN amplitudes than poor regulators on trials requiring the inhibition of automatic stereotypes but did not differ on trials that did not require stereotype inhibition (Fig. 7–5; Amodio et al., 2008). Additional analyses showed that ERN amplitudes mediated the effect of regulation group on controlled processing and response accuracy. Thus, we found that the conflict monitoring mechanism for initiating controlled processes accounted for the puzzling finding that some egalitarians were more effective in responding without bias than others.

Mechanisms for regulating bias according to internal versus external cues

A hallmark of the social psychological approach is an emphasis on the power of the situation. For example, normative influences, such as pressure from peers or authority figures, can have profound effects on the ways people think and behave (Asch, 1956; Cialdini & Trost, 1998), and modern normative standards proscribe expressions of racial bias (Crosby et al., 1980; Plant et al., 2003). Although traditional social psychological models do not distinguish between mechanisms underlying internal versus external forces on behavior (cf. Carver & Sheier, 1978), several different lines of research suggest that internal and external impetuses for control may involve different processes. In the intergroup literature, individual differences in the strength of personal and normative motivations to respond without prejudice tend to be independent (Dunton & Fazio, 1997; Plant & Devine, 1998). Research on motivation has identified different qualities of behavior motivated by personal versus normative reasons, such that personally motivated behaviors tend to be more stable and consistent than normatively motivated behaviors (Deci & Ryan, 2000; Ryan & Connell, 1989). Deci and Ryan's theory concerning different motivations for behavior (Self-Determination Theory; Deci & Ryan, 2000) has been applied to explain individual differences in implicit and explicit expressions of race bias (Amodio et al., 2003; 2008; Devine et al., 2002). In light of these findings, my colleagues and I have considered the possibility that internally and externally driven forms of control may involve different underlying mechanisms related to distinct neural processes (Amodio et al., 2006).

Interestingly, the notion that behaviors may be regulated by either internal or external impetuses for control has not been addressed by the neuroscience literature. A survey of literature on conflict monitoring reveals that most, if not all, studies have focused exclusively on internally driven forms of control in the absence of external social pressures (i.e., in most cognitive neuroscience studies, tasks are completed in a private room with minimal social interaction). However, recent neuroscience studies on empathy and mentalizing are relevant to this issue because they concern the way an individual processes information about others (Frith & Frith, 1999). In studies of empathy and mentalizing, these externally oriented processes are typically associated with activity in regions of the mPFC and rostral (r)ACC (Fig. 7–3; Amodio & Frith, 2006; Harris, Todorov, & Fiske, 2005; Mitchell, Banaji, & Macrae, 2005; Singer et al., 2004; *see also* Greene et al., 2001, for activations in orbital frontal cortex). Although this body of research has not emphasized a regulatory role for these medial frontal activations, my colleagues and I have hypothesized that these more anterior regions of mPFC may be important for externally driven forms of self-regulation, in contrast to dACC regions linked to conflict among internal cues for self-regulation (Amodio et al., 2006). We tested this hypothesis by measuring ERPs while participants completed the weapons identification task either *(1)* in private or *(2)* while being observed (via video monitor) by an experimenter for signs of prejudice. As in past work, the ERN component was taken as an index of conflict-monitoring processes. To assess activation of the rACC/mPFC, we examined the error-positivity (P_e) wave—a positive-polarity ERP component that immediately follows the ERN and is strongest at fronto-central scalp sites (Fig. 7–6a). Past work has localized this wave to the rACC and neighboring regions of mPFC (Hermann et al., 2004; Nieuwenhuis et al., 2001; van Veen & Carter, 2002). The P_e has a slower time-course than the ERN—it peaks approximately 200 milliseconds following an error response—and has been associated with the conscious perception of an unintended response. Furthermore, its putative neural generator in the rACC/mPFC suggests stronger connections to areas of the brain linked to theory of mind, social cognition, and reward processing, whereas the dACC is more richly connected to regions of brain linked to attention and motor control (Amodio & Frith, 2006).

Given the distinctions in connectivity between the dorsal and rostral regions of the ACC and mPFC, we expected that behavioral control driven by one's internal (personal) motivations would relate to conflict-monitoring and thus dACC activity, as indicated by the ERN. We expected that behavioral control motivated by social pressures would also be associated with more complex social cognitive processing and thus rACC/mPFC activity, as indicated by the P_e. All participants reported being personally motivated to respond without bias, such that they would make a strong effort to inhibit the influence of stereotypes

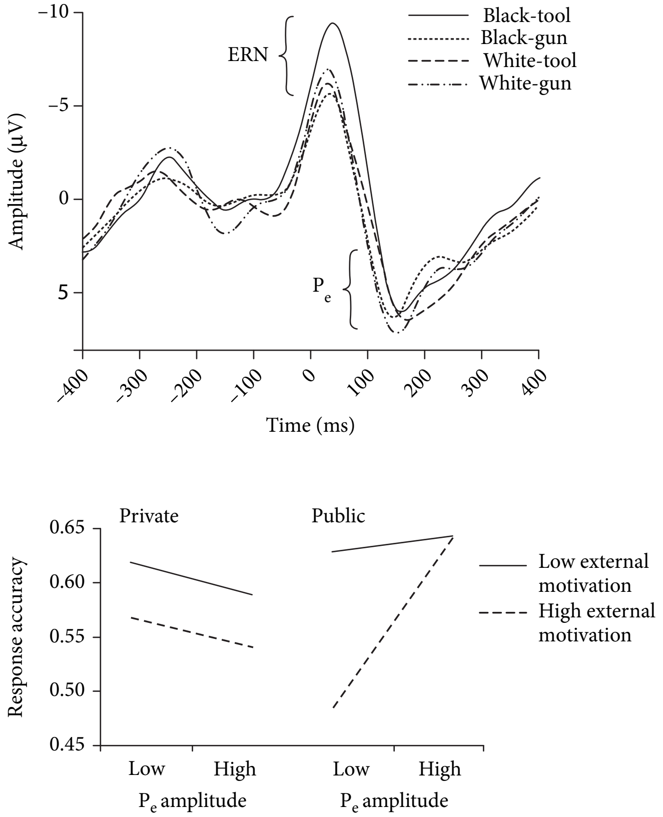


Fig. 7–6 The top panel shows error-related negativity (ERN) and error-positivity (P_e) waveforms (see labels), elicited during weapons identification task (zero indicates the time of response). The bottom panel shows predicted values for behavioral control predicted by error-positivity (P_e) amplitudes as a function of private vs. public response condition for participants with low sensitivity and high sensitivity to external social pressures. Behavioral control represents a probability estimate derived using the process-dissociation procedure. Predicted values show that P_e amplitudes predicted control only among externally sensitive participants who responded in the public condition.

on their responses. In addition, we preselected participants who reported being either high or low in sensitivity to external (normative) pressures to respond without prejudice, using Plant and Devine's (1998) scale. This way, we could test the strong hypothesis that the rACC/mPFC, as indicated by the P_e wave, is important for regulating behavior on the basis of external social cues, only for people sensitive to such cues.

Results showed that ERN and P_e amplitudes were uncorrelated across participants. As in past research, larger ERNs predicted a pattern of less-biased responding across conditions and for all participants, consistent with the idea that internal cues are always present. However, as hypothesized, the P_e wave emerged as a strong predictor of control among participants in the public response condition who reported being highly sensitive to external pressures (Fig. 7–6). An additional set of analyses confirmed that the ERN and P_e waves influenced behavior by affecting controlled, but not automatic, forms of processing. Overall, this pattern of findings provided initial evidence that internally versus externally driven forms of prejudice control arise from independent neural mechanisms associated with the dACC and rACC/mPFC, respectively. More broadly, this work is unique in that it suggests a regulatory role of the mPFC, beyond the information processing function typically ascribed to this region. Nevertheless, because our ERP measures were not well-suited for identifying the specific neural structures underlying the effects, additional fMRI research will be needed to confirm our theorized regulatory role of the dACC and rACC/mPFC regions.

Mechanisms for implementing intentional responses

Once the need for control is detected by monitoring processes, additional mechanisms of executive function are activated to override unwanted impulses with intentional responses. Generally speaking, executive function refers to a set of processes for implementing intentional behavior, most of which have been linked to regions of lateral PFC (Baddeley, 1986; Botvinick et al., 2001; Miller & Cohen, 2001; Shallice, 1982). However, it is notable that the same regions of PFC associated with cognitive control have been associated with a long list of functions, including working memory, episodic retrieval, rehearsal, semantic monitoring, motivational orientation, and attentional gating, to name a few (S. Gilbert et al., 2006). Although the specific regions of PFC activated by tasks that engage these different processes are sometimes distinguishable, they are often highly overlapping. Therefore, the observation of lateral PFC activity is not in itself diagnostic of a specific process (Cacioppo et al., 2003; Poldrack, 2006), and researchers must be careful to validate their interpretations of PFC activity as reflecting control by showing that it predicts actual behavioral control or is correlated with individual differences in motivations to control. In all likelihood, the range of processes linked to lateral PFC corresponds to some aspect of organizing and implementing deliberative and intentional responses. With this issue in mind, I now turn to the roles of prefrontal cortical regions involved in the regulation of race bias.

Inhibiting unwanted racial biases

The concept of inhibition has a long history in psychology and philosophy. Descartes famously believed that humans should strive to inhibit the base urges of the body, and Freud's psychodynamic approach to therapy centered on an individual's ability to inhibit unwanted drives and fantasies. The emphasis on inhibition as a component of control continues to be a major theme in modern psychology, including in popular social psychological models of control (e.g., Bodenhausen & Macrae, 1998; Monteith, Sherman, & Devine, 1998; for a discussion of alternative views, *see* Botvinick et al., 2001; D. Gilbert, 1998). Research on inhibitory mechanisms in the brain has focused on the right ventrolateral PFC (vlPFC; also, inferior frontal cortex). Initial findings from studies of patients with lesions in these areas suggested that the right vlPFC is uniquely associated with performance on response inhibition tasks such as the stop-signal task and Wisconsin card-sorting task (Aron, Robbins, & Poldrack, 2004). Applying this model to the regulation of race bias, Lieberman et al. (2005) examined changes in vlPFC activity in the context of race-biased responding using fMRI. Participants in this study viewed faces of White or Black individuals presented in the center of the computer screen. At the bottom of the screen, participants viewed either two faces (one Black and one White face) or two group labels ("African American" or "Caucasian"), positioned on the left and right sides. Participants matched the centered face to one of the faces or group labels presented below. Lieberman et al. (2005) reasoned that matching a face with a symbolical lexical representation of the group requires more complex cognitive processing than matching pictorial representations of faces. Moreover, the authors argued that the process of labeling involves the active inhibition of automatically activated affective responses to faces. Consistent with this reasoning, participants in this study showed greater amygdala activity when matching faces of Black versus White people but showed no amygdala effects when matching labels. By comparison, matching labels of Black faces elicited significantly greater vlPFC activity compared with matching labels of White faces, and the degree of vlPFC activation during labeling was negatively correlated with amygdala activation in response to Black faces, suggesting that vlPFC may actively inhibit implicit prejudice responses (*see also* Cunningham et al., 2004).

Implementing intended egalitarian responses

The most important component of prejudice control is the implementation of an intended, nonbiased response, because in the end, it is one's behavior that leads to discrimination. It is worth noting that that phrase "prejudice control" may be misleading, as it connotes the promotion of more favorable responses toward Black individuals rather than a lack of bias. However, a truly egalitarian response is one that is unaffected by bias (e.g., Fiske & Neuberg, 1990), and thus *prejudice control* represents the ability to respond in an intentional (e.g., accurate) manner irrespective of the potentially biasing effects of automatic prejudices and stereotypes (Payne, 2005; Amodio et al., 2008). Given research suggesting that frontal cortical regions are particularly well-designed for regulating behavior, but not so effective at regulating thoughts or emotions, I define prejudice control more specifically as the implementation of intentional behavioral responses.

As in the cognitive and affective neuroscience literatures, social neuroscience research on controlled processes in the context of race has focused on the dlPFC (Fig. 7–2; Amodio, in press; Amodio et al., 2003, 2004; Bartholow, Dickter, & Sestir, 2006; Cunningham, 2004; Richeson et al., 2003). Amodio et al. (2003) suggested that lateral and orbital PFC are likely to support more intentional aspects of racial responses, on the basis of theorizing in cognitive and affective neuroscience (Miller & Cohen, 2001; Rolls, 2000), and researchers applying conflict-monitoring theory to issues of prejudice control have posited that the dlPFC is important for the implementation of a nonprejudiced response (Amodio et al., 2004; Richeson et al., 2003). For example, Richeson et al. (2003) found that exposure to Black (vs. White) faces elicited activations of the ACC and dlPFC (but not the amygdala) in their White participants, which the authors interpreted as spontaneous efforts to exert control when viewing a Black face. Following the fMRI scanner task, these participants interacted with a Black experimenter and then, under instruction of a White experimenter, completed a standard Stroop color-naming task. Richeson et al. (2003) hypothesized that any self-regulatory effort expended during the interaction, presumably to inhibit signs of race bias, would deplete participants' cognitive resources, leaving them to perform more poorly on the Stroop task. Indeed, participants who exhibited greater PFC activity while viewing Black faces performed worse on the Stroop task following the interracial interaction.

Similarly, participants in a study by Cunningham et al. (2004) viewed faces of White and Black individuals and indicated whether the image appeared on the right or left side of the screen. When faces were clearly visible (i.e., presented for 525 ms), Black faces activated the ACC and some dlPFC areas more than White faces. Given that these regions have been implicated in aspects of control in past research, it is possible that these activations related to the control of prejudice. However, because these activations were not correlated with participants' self-reported prejudice attitudes or with a behavioral index of control, the role of the PFC in the control of a prejudiced response was ambiguous.

The findings of Richeson et al. (2003) and Cunningham et al. (2004) are consistent with the idea that the PFC is related to controlled forms of processing, but additional research is needed to show that these activations are directly related to the control of prejudice. Indeed, few studies to date have reported a direct link between PFC activity and the regulation of bias and controlled patterns of behavior. With this issue in mind, my colleagues and I recently used an EEG measure of dlPFC activity (cf. Pizzagalli, Sherwood, Henriques, & Davidson, 2005) to examine the role of this brain region in the behavioral regulation of prejudice (Amodio, Devine, & Harmon-Jones, 2007). A large body of literature has suggested that left versus right asymmetries in frontal cortical activity are associated with approach versus withdrawal motivation (Harmon-Jones, 2003), and we were interested in the roles of motivation and PFC activity in the regulation of race bias. Using EEG, we measured changes in PFC activity just after participants realized they had responded in a prejudiced manner on a task, and then again when they were given an opportunity to engage in an activity designed to reduce their level of prejudice. In line with predictions, we observed a reduction in left frontal activity when participants believed they had acted in a prejudicial way, suggesting a drop in approach motivation. This change in frontal EEG was correlated with an increase in guilt but was unrelated to changes in other emotions (anxiety, sadness, other-directed negative emotion, or positive emotion). However, when participants were given a chance to redress their prejudiced behavior by reading magazine articles on how to reduce race bias, left frontal activity was increased. Importantly, a stronger interest in prejudice-reduction activities was associated with greater left frontal activity, whereas their desire to engage in other activities unrelated to prejudice was not related to brain activity.

Although Amodio et al. (2007) demonstrated a relatively direct link between changes in dlPFC activity and a behavioral measure of prejudice control, this study did not assess the role of the dlPFC in controlling responses as they unfolded "in the moment." A more recent study used EEG to monitor dlPFC activity while participants completed a task requiring the control of racial stereotypes (Amodio, 2010). As expected, greater left dlPFC activity during the task was associated with stronger attentional orienting to Black versus White faces and greater behavioral control over stereotype-based response tendencies. This study provided the first clear evidence for the role of the PFC in the control of prejudice.

It is worth noting that a larger issue in the study of prejudice control concerns theoretical ambiguity regarding the target of control—that is, what exactly is being controlled? Thoughts? Emotions? Behavior? All of the above? Philosophers of the mind and early psychologists were particularly focused on controlling emotions and, to a lesser extent, thoughts (e.g., Plato, 1993). The focus on regulating thoughts and emotions, as opposed to behavior, has extended to modern social psychology (*see* Wegner & Bargh, 1998, for a review). Ironically, however, research consistently shows that humans are generally unable to regulate their emotions (e.g., Gross & Levenson, 1993; Ochsner & Gross, 2003) and thoughts (e.g., Wegner, 1994; Wegner et al., 1987). Similar effects have been shown in the context of prejudice and stereotyping (Macrae et al., 1994; Monteith, Devine, & Sherman, 1998; Monteith, Spicer, & Tooman, 1998; Wyer, Sherman, & Stroessner, 1998). By contrast, research has shown that humans are much more effective at regulating aspects of their behavior, irrespective of their thoughts and emotions (e.g., DePaulo, 1992; Ekman & O'Sullivan, 1991). The idea that behaviors, but not thoughts and emotions, can be effectively controlled is generally consistent with anatomical connectivity of the frontal cortex. Although largely derived from research on monkey and rat brains, anatomical studies suggest that the lateral regions of frontal cortex associated with control are primarily interconnected to neural regions linked to motor activity (e.g., motor cortex, basal ganglia; Lehéricy et al., 2004) but have relatively sparse connections to the amygdala (Gabbot et al., 2005; Ghashghaei & Barbas, 2002). Furthermore, the lateral PFC has particularly dense receptor fields for dopamine, a neurotransmitter important for orchestrating goal-driven behavior, suggesting that the lateral PFC may be more strongly involved in orchestrating behavior as opposed to inhibiting thoughts or emotions. This issue is particularly important for prejudice reduction efforts. If it turns out that controlled processes pertain primarily to behavior, then strategies that focus on the regulation of behavior rather than unwanted thoughts or emotions may be most effective (and vice versa). Certainly, the question of "what's being controlled" is complex, and prejudice researchers will benefit from input on

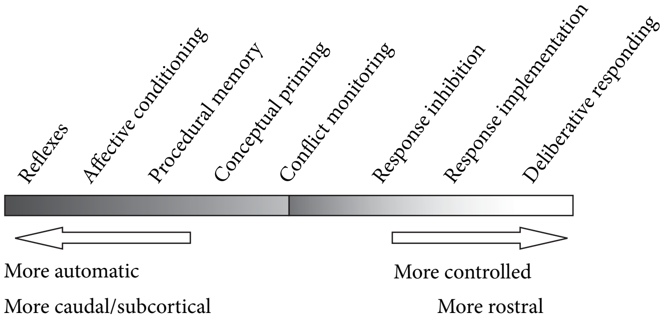


Fig. 7–7 Different processes involved in intergroup responses vary along a continuum from more automatic to more deliberative. Progression along this continuum corresponds roughly to neuroanatomy, such that more automatic processes tend to have more caudal/subcortical substrates that are evolutionarily older, whereas more deliberative processes tend to have more rostral cortical substrates that are evolutionarily newer. The process of conflict monitoring may represent the mechanism by which implicit processes become explicit, although this hypothesis remains speculative.

this question offered by the behavioral neuroscience literature.

A SOCIAL NEUROSCIENCE MODEL OF RACE BIAS ACTIVATION AND CONTROL

The findings reviewed in this chapter suggest an expanded view of the basic dual-process model of prejudice that has dominated the field for nearly 20 years. Although the body of social neuroscience research on prejudice and stereotyping is still quite small, researchers have made efficient use of models developed in the larger cognitive and behavioral neuroscience literatures to inform social psychological questions. As illustrated in Figure 7–7, there are several important subcomponents contained within the classic dual-process framework that fall along a continuum from more automatic to more controlled processes. This multicomponent framework corroborates and expands on the nuances of automaticity and control identified by previous theorists (e.g., Bargh, 1989; Devine & Monteith, 1999; Fiske & Neuberg, 1990). One may begin to wonder: How many processes are there? I have described six processes that appear to have roots in distinct neural systems, but this number is likely to change as our understanding of neural function and social behavior evolves. The point of a multiprocess approach is not so much to determine the correct number but to acknowledge that multiple interacting processes are at play that may not be easily captured by a classic dichotomous view of automaticity versus control (Conrey et al., 2005; Sherman et al., 2008). By considering how different neurocognitive mechanisms function, interact with each other, and influence behavior, prejudice researchers will continue to refine their models of intergroup bias in the service of social issues as well as science.

REFERENCES

Abelson, R. P., Kindre, D. R., Peters, M. D., & Fiske, S. T. (1982). Affective and semantic components in political person perception. *Journal of Personality and Social Psychology, 42*, 619–630.

Ajzen, I., & Fishbein, M. (2000). Attitudes and the attitude-behavior relation: Reasoned and automatic processes. In W. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology* (pp. 1–33). New York: John Wiley & Sons.

Allport, G. W. (1954). *The Nature of Prejudice*. Reading, MA: Addison-Wesley.

Amodio, D. M. (2008). The social neuroscience of intergroup relations. In W. Stroebe and M. Hewstone (Eds.), *European Review of Social Psychology* (Vol. 19, pp. 1–54). Hove, UK: Psychology Press.

Amodio, D. M. (2010). Coordinated roles of motivation and perception in the regulation of intergroup responses: Frontal cortical asymmetry effects on the P2 event-related potential and behavior. *Journal of Cognitive Neuroscience, 22*, 2609–2617.

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91*, 652–661.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2007). A dynamic model of guilt: Implications for motivation and self-regulation in the context of prejudice. *Psychological Science, 18*, 524–530.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2008). Individual differences in the regulation of intergroup bias: The role of conflict monitoring and neural signals for control. *Journal of Personality and Social Psychology, 94*, 60–74.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*, 268–277.

Amodio, D. M., Harmon-Jones, E., & Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink responses and self-report. *Journal of Personality and Social Psychology, 84*, 738–753.

Amodio, D. M., Harmon-Jones, E., & Devine, P. G. (2007). Mechanisms for the regulation of intergroup responses: Insights from a social neuroscience approach. In E. Harmon-Jones & P. Winkielman (Eds.), *Fundamentals of Social Neuroscience* (pp. 353–375). New York: Guilford.

Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., & Covert, A. E. (2004). Neural signals for the detection of unintentional race bias. *Psychological Science, 15*, 88–93.

Amodio, D. M., Kubota, J. T., Harmon-Jones, E., & Devine, P. G. (2006). Alternative mechanisms for regulating racial responses according to internal vs. external cues. *Social Cognitive and Affective Neuroscience, 1*, 26–36.

Anderson, A. K., Christoff, K., Stappen, I., et al. (2003). Dissociated neural representations of intensity and valence in human olfaction. *Nature Neuroscience, 6*, 196–202.

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences, 8*, 170–177.

Asch, S. E. (1956). Studies of independence and submission to group pressure: I. A minority of one against a unanimous majority. *Psychological Monographs, 70*(9, Whole No. 417).

Baddeley, A. D. (1986). *Working Memory*. New York: Oxford University Press.

Bartholow, B. D., Dickter, C. L., & Sestir, M. A. (2006). Stereotype activation and control of race bias: Cognitive control of inhibition and its impairment by alcohol. *Journal of Personality and Social Psychology, 90*, 272–287.

Bartholow, B. D., Fabiani, M., Gratton, G., & Bettencourt, B. A. (2001). A psychophysiological analysis of cognitive processing of and affective responses to social expectancy violations. *Psychological Science, 12*, 197–204.

Bodenhausen, G. V., & Macrae, C. N. (1998). Stereotype activation and inhibition. In R. S. Wyer, Jr. (Ed.), *Advances in Social Cognition, Vol. 11*. (pp. 1–52). Mahwah, NJ: Erlbaum.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*, 624–652.

Botvinick, M. M., Nystrom, L. E., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature, 402*, 179–181.

Bouton, M. E. (1994). Conditioning, remembering, and forgetting. *Journal of Experimental Psychology: Animal Behavior Processes, 20*, 219–231.

Brigham, J. C. (1993). College Students' Racial Attitudes. *Journal of Applied and Social Psychology, 23*, 1933–1967.

Cacioppo, J. T., & Berntson, G. G. (1992). Social psychological contributions to the decade of the brain: Doctrine of multilevel analysis. *American Psychologist, 47*, 1019–1028.

Cacioppo, J. T., Berntson, G. G., Lorig, T. S., Norris, C. J., Rickett, E., & Nusbaum, H. (2003). Just because you're imaging the brain doesn't mean you can stop using your head: A primer and set of first principles. *Journal of Personality and Social Psychology, 85*, 650–661.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science, 280*, 747–749.

Carver, C. S., & Scheier, M. F. (1978). Self-focusing effects of dispositional self-consciousness, mirror presence, and audience presence. *Journal of Personality and Social Psychology, 36*, 324–332.

Chaiken, S., & Trope, Y. (1999). *Dual-Process Theories in Social Psychology*. New York: Guilford Press.

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology, Vol. 2*. (pp. 151–192). New York: McGraw-Hill.

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. (2005). Separating multiple processes in implicit social cognition: The Quad-Model of implicit task performance. *Journal of Personality and Social Psychology, 89*, 469–487.

Correll, J., Urland, G. R., & Ito, T. A. (2006). Shooting straight from the brain: Early attention to race promotes bias in the decision to shoot. *Journal of Experimental Social Psychology, 42*, 120–128.

Crosby, F., Bromley, S., & Saxe, L. (1980). Recent unobtrusive studies of Black and White discrimination and prejudice: A literature review. *Psychological Bulletin, 87*, 546–563.

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of Black and White Faces. *Psychological Science, 15*, 806–813.

Cunningham, W. A., Raye, C. L., & Johnson, M. K. (2004). Implicit and explicit evaluation: fMRI correlates of valence, emotional intensity, and control in the processing of attitudes. *Journal of Cognitive Neuroscience, 16*, 1717–1729.

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

Davis, M. (2006). Neural systems involved in fear and anxiety measured with fear-potentiated startle. *American Psychologist, 61*, 741–756.

Davis, M. & Whalen, P. J. (2001). The amygdala: Vigilance and emotion. *Molecular Psychiatry, 6*, 13–34.

Deci, E. L. & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self–determination of behavior. *Psychological Inquiry, 11*, 227–268.

DePaulo, B. M. (1992). Nonverbal behavior and self-presentation. *Psychological Bulletin, 111*, 203–243.

Devine, P. G. (1989). Prejudice and stereotypes: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5–18.

Devine, P. G., & Monteith, M. M. (1999). Automaticity and control in stereotyping. In S. Chaiken & Y. Trope (Eds.), *Dual Process Theories in Social Psychology* (pp. 339–360). New York: Guilford Press.

Devine, P. G., Monteith, M. M., Zuwerink, J. R., & Elliot, A. J. (1991). Prejudice with and without compunction. *Journal of Personality and Social Psychology, 60*, 817–830.

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E, & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology, 82*, 835–848.

Dovidio, J. F., Brigham, J. C., Johnson, B. T., & Gaertner, S. L. (1996). Stereotyping, prejudice and discrimination: Another look. In C. N. McCrae, C. Stangor, & M. Hewstone (Eds.), *Stereotypes and Stereotyping* (pp. 276–319). New York: Guilford.

Dovidio, J. F., Esses, V. M., Beach, K. R., & Gaertner, S. L. (2004). The role of affect in determining intergroup behavior: The case of willingness to engage in intergroup affect. In D. M. Mackie & E. R. Smith (Eds.), *From Prejudice to Intergroup Emotions: Differentiated Reactions to Social Groups* (pp. 153–171). Philadelphia: Psychology Press.

Dovidio, J. F., Evans, N., & Tyler, R. B. (1986). Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology, 22*, 22–37.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 23*, 316–326.

Eberhardt, J. L. (2005). Imaging race. *American Psychologist, 60*, 181–190.

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist, 46*, 913–920.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 23, pp. 75–109). New York: Academic Press.

Fazio, R. H., & Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes.

In S. Chaiken, & Y. Trope (Eds.), *Dual Process Theories in Social Psychology* (pp. 97–116). New York: Guilford.

Fendt, M., & Fanselow, M. S. (1999). The neuroanatomical and neurochemical basis of conditioned fear. *Neuroscience and Biobehavioral Review, 23*, 743–760.

Fiske, S. T. (1992). Thinking is for doing: Portraits of social cognition from Daguerreotype to laserphoto. *Journal of Personality and Social Psychology, 63*, 877–889.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum model of impression formation: Form category-based to individuating process as a function of information, motivation, and attention. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 23, pp. 1–108). San Diego, CA: Academic Press.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology, Vol. 2* (pp. 357–411). New York: McGraw-Hill.

Fiske, S. T., & Taylor, S. E. (1984). *Social Cognition*. New York: Random House.

Foerde, K., Knowlton, B. J., & Poldrack, R. A. (2006). Modulation of competing memory systems by distraction. *Proceedings of the National Academies of Sciences, 103*, 11,778–11,783.

Frith, C. D., & Frith, U. (1999). Interacting minds––a biological basis. *Science, 286*, 1692–1695.

Gabbott, P. L., Warner, T. A., Jays, P. R., Salway, P., & Busby, S. J. (2005). Prefrontal cortex in the rat: projections to subcortical autonomic, motor, and limbic centers. *Journal of Comparative Neurology, 492*, 145–177.

Gabrieli, J. D. (1998). Cognitive neuroscience of human memory. *Annual Review of Psychology, 49*, 87–115.

Gaertner, S. L., & Dovidio, J. F. (1986). The aversive form of racism. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, Discrimination, and Racism* (pp. 61–89). San Diego, CA: Academic Press.

Gerhing, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science, 4*, 385–390.

Ghashghaei, H. T., & Barbas, H. (2002). Pathways for emotion: interactions of prefrontal and anterior temporal pathways in the amygdala of the rhesus monkey. *Neuroscience, 115*, 1261–1279.

Gilbert, D. T. (1999). What the mind's not. In S. Chaiken & Y. Trope (Eds.), *Dual-process Theories in Social Psychology* (pp. 3–11). New York: Guilford Press.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology, 60*, 509–517.

Gilbert, S. J., Spengler, S., Simons, J. S., et al. (2006). Functional specialization within rostral prefrontal cortex (area 10): A meta-analysis. *Journal of Cognitive Neuroscience, 18*, 932–948.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105–2108.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition. *Psychological Review, 102*, 4–27.

Greenwald, A., McGhee, D., & Schwartz, J. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

Gross, J. J., & Levenson, R. W. (1993). Emotional suppression: Physiology, self-report, and expressive behavior. *Journal of Personality and Social Psychology, 64*, 970–986.

Harmon-Jones, E. (2003). Clarifying the emotive functions of asymmetrical frontal cortical activity. *Psychophysiology, 40*, 838–848.

Harris, L. T., Todorov, A., & Fiske, S. T. (2005). Attributions on the brain: Neuro-imaging dispositional inferences, beyond theory of mind. *NeuroImage, 28*, 763–769.

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.

Herrmann, M. J., Rommler, J., Ehlis, A. C., Heidrich, A., & Fallgatter, A. J. (2004). Source localization (LORETA) of the error-related-negativity (ERN/N_e) and positivity (P_e). *Cognitive Brain Research, 20*, 294–299.

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

Ito, T. A., & Cacioppo, J. T. (2000). Electrophysiological evidence of implicit and explicit categorization processes. *Journal of Experimental Social Psychology, 36*, 660–676.

Ito, T. A., & Urland, G. R. (2003). Race and gender on the brain: Electrocortical measures of attention to race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology, 85*, 616–626.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513–541.

Katz, I., & Hass, R. G. (1988). Racial ambivalence and American value conflict: Correlational and priming studies of dual cognitive structures. *Journal of Personality and Social Psychology, 55*, 893–905.

Kawakami, K., Dion, K. L., & Dovidio, J. F. (1998). Racial prejudice and stereotype activation. *Personality and Social Psychology Bulletin, 24*, 407–416.

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science, 273*, 1399–1402.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review, 97*, 377–395.

LeDoux, J. E., Cicchetti, P., Xagoraris, A., & Romanski, L. M. (1990). The lateral amygdaloid nucleus: Sensory interface of the amygdala in fear conditioning. *Journal of Neuroscience, 10*(4), 1062–1069.

LeDoux, J. E. (1992). Emotion and the amygdala. In J. P. Aggleton (Ed.), *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction* (pp. 339–351). New York: Wiley-Liss.

Lehéricy, S., Ducros, M., Van de Moortele, P. F., et al. (2004). Diffusion tensor fiber tracking shows distinct corticostriatal circuits in humans. *Annals of Neurology, 55*, 522–529.

Lieberman, M. D., Hariri, A., Jarcho, J. M., Eisenberger, N. I., & Bookheimer, S. Y. (2005). An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. *Nature Neuroscience, 8*, 720–722.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., Thorn, T. M. J., & Castelli, L. (1997). On the activation of social stereotypes: The moderating role of processing objectives. *Journal of Experimental Social Psychology, 33*, 471–489.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*, 159–188.

McConahay, J. B., & Hough, J. C. (1976). Symbolic racism. *Journal of Social Issues, 32*, 23–45.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*, 167–202.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005).The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience, 17*, 1306–1315.

Monteith, M. J. (1993). Self-regulation of stereotypical responses: Implications for progress in prejudice reduction. *Journal of Personality and Social Psychology, 65*, 469–485.

Monteith, M., Devine, P. G., & Zuwerink, J. (1993). Self-directed vs. other-directed affect as a consequence of prejudice-related discrepancies. *Journal of Personality and Social Psychology, 64*, 198–210.

Monteith, M. J., Sherman, J. W., & Devine, P. G. (1998). Suppression as a stereotype control strategy. *Personality and Social Psychology Review, 2*, 63–82.

Monteith, M. J., & Spicer, C. V., & Tooman, G. (1998). Consequences of stereotype suppression: Stereotypes on AND not on the rebound. *Journal of Experimental Social Psychology, 34*, 355–377.

Myrdal, G. (1944). *An American Dilemma*. New York: Harper & Row.

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P. H., & Kok, A. (2001). Error-related brain potentials are differently related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology, 38*, 752–760.

Ochsner, K., & Gross, J. J. (2005). The cognitive control of emotion. *Trends in Cognitive Sciences, 9*, 242–249.

Ochsner, K. N., & Lieberman, M. D. (2001). The emergence of social cognitive neuroscience. *American Psychologist, 56*, 717–734.

Park, B., & Judd, C. M. (2005). Rethinking the link between categorization and prejudice within the social cognition perspective. *Personality and Social Psychology Review, 9*, 108–130.

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81*, 181–192.

Payne, B. K. (2005). Conceptualizing control in social cognition: How executive functioning modulates the expression of automatic stereotyping. *Journal of Personality and Social Psychology, 89*, 488–503.

Plato (1993). *The Republic* (Translated by A.D. Lindsay). New York: Knopf.

Pessoa, L., McKenna, M., Gutierrez, E., & Ungerleider, L. G. (2005). Neural processing of emotional faces requires attention. *Proceedings of the National Academy of Sciences, 99*, 11,458–11,463.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., et al. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience, 12*, 729–738.

Pizzagalli, D. A., Sherwood, R., Henriques, J. B., & Davidson, R. J. (2005). Frontal brain asymmetry and reward responsiveness: A source localization study. *Psychological Science, 16*, 805–813.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75*, 811–832.

Plant, E. A., Devine, P. G., & Brazy, P. C. (2003). The bogus pipeline and motivations to reduce prejudice: Revisiting the fading and faking of racial prejudice. *Group Processes and Intergroup Relations, 6*, 187–200.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10*, 59–63.

Richeson, J. A., Baird, A. A., Gordon, H. L., et al. (2004). An fMRI examination of the impact of interracial contact on executive function. *Nature Neuroscience, 6*, 1323–1328.

Rolls, E. T. (2000). The orbitofrontal cortex and reward. *Cerebral Cortex, 10*, 284–294.

Rudman, L. A., Greenwald, A. G., & McGhee, D. E. (2001). Implicit self-concept and evaluative implicit gender stereotypes: Self and ingroup share desirable traits. *Personality and Social Psychology Bulletin, 27*, 1164–1178.

Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology, 57*, 749–761.

Sears, D. O., & Henry, P. J. (2005). Over thirty years later: A contemporary look at symbolic racism. In M. P. Zanna (Ed), *Advances in Experimental Social Psychology, Vol. 37* (pp. 95–150). San Diego, CA: Elsevier Academic Press.

Schuman, H., Steeh, C., Bobo, L., & Krysan, M. (1997). *Racial Attitudes in America: Trends and Interpretations*. Cambridge, MA: Harvard University Press.

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London, B298*, 199–209.

Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review, 115*, 314–335.

Shiffrin, R., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127–190.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science, 303*, 1157–1162.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*, 108–131.

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences, 93*, 13,515–13,522.

Taylor, S. E., & Falcone, H. (1982). Cognitive bases of stereotyping: The relationship between categorization and prejudice. *Personality and Social Psychology Bulletin, 8*, 426–432.

van Veen, V., & Carter, C. S. (2002). The timing of action-monitoring processes in the anterior cingulate cortex. *Journal of Cognitive Neuroscience, 14*, 593–602.

Wegener, D. T., & Petty, R. E. (1997). The flexible correction model: The role of naïve theories of bias in bias correction. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 29, pp. 141–208). Mahwah, NJ: Erlbaum.

Wegner, D. M., Schneider, D. J., Carter, S., & White, T. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology, 53*, 5–13.

Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review, 101*, 34–52.

Whalen, P. J. (1998). Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala. *Current Directions in Psychological Science, 7*, 177–188.

Wheeler, M. E., & Fiske, S. T. (2005) Controlling racial prejudice and stereotyping: Social cognitive goals affect amygdala and stereotype activation. *Psychological Science*, 16, 56–63.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*, 592–598.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, 116, 117–142.

Wilson, T., Lindsey, S., & Schooler, T. (2000). A model of dual attitudes. *Psychological Review, 107*, 101–126.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology, 72*, 262–274.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Evaluative versus conceptual judgments in automatic stereotyping and prejudice. *Journal of Experimental Social Psychology, 37*, 244–252.

Word, C. O., Zanna, M. P., & Cooper, J. (1976). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology, 10*, 109–120.

Wyer, N. A., Sherman, J. W., & Stroessner, S. J. (1998). The spontaneous suppression of racial stereotypes. *Social Cognition, 16*, 340–352.

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review, 111*, 931–959.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35*, 151–175.

CHAPTER 8
Perceiving Humanity or Not: A Social Neuroscience Approach to Dehumanized Perception

Lasana T. Harris & Susan T. Fiske

Defining what is human is complicated. Scientific inquiry on this abstract topic requires an interdisciplinary approach. To present an overarching view involves understanding various perspectives and some philosophy as well. In fact, the study of philosophy often begins by asking the question "What is the mind?" (Appiah, 2003). This question is related to the more important question "What is human?" because the lay psychological answer often is that human beings have a mind, or an inner life, and having this inner life is what separates human beings from other entities. From the initial question of what it means to have a mind follows a further inquiry: How do I know that I have a mind? Descartes' *cogito ergo sum* provides a partial answer but begs the question, "How do I know that others also think?" All of these philosophical questions have "the mind" as a subject, and we will argue that a failure to consider the mind of others is associated with dehumanizing them. The following philosophical thought experiment provides the premise for our argument and sheds some light on this idea.

Consider the following scenario borrowed from Appiah (2003). Imagine there was a machine that looked, felt, sounded like, and did everything your mother did. How would you be able to tell that this machine was not your mother? The lay psychological answer is that this machine would not have the subjective experiences that your mother can have daily: The robot would not feel as your mother feels. This suggests that you think about the mind or "inner life" of your mother and use that as a determinant of her really being human. This may also suggest that you distinguish your mother from everything that is not your mother by imagining the contents of her mind. Essentially, this question and this lay psychological answer help reveal *perceived humanity*—the psychological process of perceiving an entity as human—which is the opposite of *dehumanized perception*—a failure to consider the inner life or mind of another.

We can use social neuroscience to understand *dehumanized perception*, this failure to think about another person's mind (mentalizing). This extreme form of prejudice entails perceiving a person as less than, not quite, or not at all human. Social neuroscience as an interdisciplinary approach combines the literature on two phenomena central to dehumanized perception—dehumanization and mentalizing—using cognitive and developmental neuroscience, as well as social psychology to explore this phenomenon. Social neuroscience also has the advantage of being reciprocal in nature; social psychological theory makes predictions about neuroscience data, and the subsequent results can then inform social psychological theory and predict behavioral data. This allows insight into the functional significance of both neural systems and abstract psychological concepts.

This chapter argues that *dehumanized perception* may be a psychological response to social targets who elicit the negative basic emotion disgust (*see* Harris & Fiske, 2006). We review social neuroscience data showing that the medial prefrontal cortex (mPFC)—an area implicated in mentalizing and social cognition—is not as active for these dehumanized targets as for other social targets. We then review subsequent social psychological predictions generated by the neural data; these data show that participants fail to think about the minds of these dehumanized targets to the same extent as other social targets. Participants also describe these dehumanized targets as ill-intentioned, inept, unfamiliar, dissimilar, strange, and not uniquely human or quite typically human. We conclude with a discussion of some factors that may moderate *dehumanized perception*, perhaps relevant to the hope of reducing some of the thought processes and emotions that underlie human atrocities.

Mentalizing

A specific type of social cognition involves the inference of another's mind. Commonly referred to in the cognitive and developmental neuroscience literature as mentalizing or Theory of Mind (ToM), this cognitive process can be automatic or controlled (Frith & Frith, 2001). Imagining a social target's mind is influenced by information inferred from the person, the environment, or a third source, such as prior experience. This research generally has separated people's understanding of other people's individual dispositions (hypothesized to be an early-developing process) and their beliefs (hypothesized to develop later; *see* Saxe et al., 2004). ToM includes goals (inferred from bodily actions), attention (inferred from gaze direction), and emotion (inferred from others' expressions). People generally believe that goals, like intentions, can predict behavior. Intent can also be attributed to animate objects (Heider & Simmel, 1944), suggesting that the process is not reserved for people. Other ToM research further separates perceived intent, biological motion, episodic memory retrieval, and decouples mental states from reality (e.g., Frith & Frith, 2001; Gallagher & Frith, 2003).

People also specifically infer other personalities, a process that understands a social target's mind as containing both goals and intent. In fact, because of people's ability to calculate dispositional attributions, Heider (1958) labeled people naïve scientists. Inferring other personalities may be viewed as an attempt to gain an insight into global warmth or intentionality (addressing the friend-foe dimension) and competence (addressing the other's ability to enact those intentions; *see* Fiske, Cuddy, & Glick, 2006). The first personality inference suggests the target's general good or ill intent; the second suggests the target's degree of agency. Inferred warmth and competence predict the behavior that guides social interaction.

Sometimes people would like to judge the behavioral tendencies of nonhuman agents as well, although they know such agents do not have internal mental processes. Consequently, people also make dispositional attributions to objects (Harris & Fiske, 2008), suggesting that this mentalizing process is not reserved for people. Anthropomorphism and personification of animals and objects possibly reflect this tendency to assume that agents possess internal states (Kwan, Gosling, & John, 2003). Inferring another's internal states allows prediction of that other's actions. Nevertheless, most objects are not viewed as having a mind. Therefore a failure to infer a person's disposition involves a failure to perform a basic cognitive process and hints at *dehumanized perception*.

As the social psychological literature on dispositional attributions to social targets shows, people automatically infer dispositions from even thin slices of behavior (Ambady & Rosenthal, 1993; Dunning et al., 1989). In addition, these spontaneous first impressions often prove to have strong traces in memory and influence later judgments of the individual (Todorov et al., 2007). This rapid mental inference ability first appeared in Asch's (1946) original impression formation study. Attribution theory within social psychology has focused almost exclusively on how people make spontaneous dispositional attributions about others and the dimensions that moderate this effect.

Perspective-taking, a more conscious mentalizing, describes the process of intentionally inferring another's internal mental states. This controlled process moderates conscious and unconscious prejudice (Galinsky & Moscowitz, 2000). In particular, inferring others' internal states deactivates stereotypes through the greater overlap between self and other, allowing the other to seem similar to the self (Galinsky, Ku, & Wang, 2005). These effects occur in both real social groups and minimalist groups (Galinsky & Moscowitz, 2000). Prior to this work, social psychology had a long history of implicating perspective-taking in moral reasoning (Kohlberg, 1976), altruism (Batson, 1991, 1998), and aggression (Richardson et al., 1994). When people fail to take the perspective of dehumanized targets, they feel disgust, a strictly negative emotion often linked to perceived moral violations and subsequent aggressive responses (Haidt et al., 1997).

The chapter approaches perceived humanity through a series of experiments in which participants simply view social targets, report their own experienced emotion, and comment on the targets' perceived traits. The remainder of the chapter is organized by type of data, either neural or self-report. We first examine the role of *neural* activity in person perception, using the *Stereotype Content Model (SCM)* to generate predictions. We then explore both the functional significance of *neural* regions and social psychological *theories of dehumanization* to generate hypothesis about the *self-report* data. Finally, if *dehumanized perception* is a type of prejudice, then moderators of prejudice such as *intergroup contact* should relate to *dehumanized perception*. We test this final hypothesis with *self-report* data. This social neuroscience approach therefore utilizes *dehumanized perception* to understand people's perception of human beings.

Neural evidence for dehumanized perception

Stereotype Content Model

The SCM (Fiske, Cuddy, Glick, & Xu, 2002) predicts differentiated prejudices. It incorporates a fundamental friend–foe plus capability judgment; the SCM proposes that we appraise societal groups as intending either help or harm (*warmth*) and as capable or not to enact those intentions (*competence*; Fiske, Cuddy, & Glick, 2006; Fiske et al., 2002). These dimensions are rooted in classic person perception (Rosenberg, Nelson, & Vivekananthan, 1968) and differentiate out-groups into four low-high warmth x competence clusters. The four combinations of the competence and warmth dimensions produce four distinct emotions toward social groups: pride, envy, pity, and disgust (*see* Fig. 8–1).

Thus, not all out-groups provoke unambivalent animosity. Out-groups stereotyped as either competent or warm (but not both) elicit ambivalent emotions, whereas in-groups and allies perceived as high on both dimensions receive a positive response. These latter responses (pride, admiration) assume self-relevant, positive outcomes and are reserved for the cultural defaults (e.g., the middle-class). Out-group prejudices occur in the remaining three quadrants, and some are worse than others. Moderate prejudices are ambivalent, mixing positive and negative reactions. In one mixed case, envy and jealousy (which resent another's positive outcomes) are elicited by groups stereotyped as competent but not warm (e.g., rich people); envy admits respect but harbors dislike. In the other mixed combination, groups stereotyped as warm, but not competent (e.g., elderly people), elicit pity and sympathy (emotions reserved for people with uncontrollable negative outcomes). Pity admits benign reactions but also disrespect. Only the most extreme out-groups, the

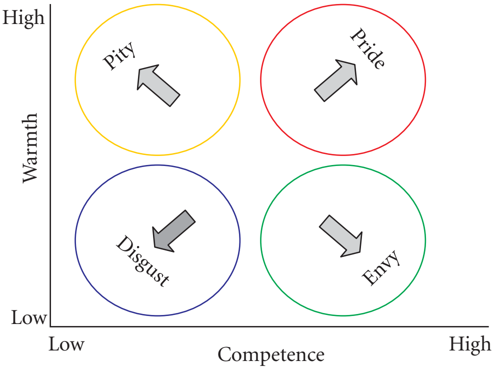


Fig. 8–1 The Stereotype Content Model

low-low, receive unabashed disliking and disrespect: Groups stereotyped as neither warm nor competent elicit the worst kind of prejudice—disgust and contempt—based on perceived moral violations and negative outcomes that they allegedly caused themselves. Both people and nonhuman agents can elicit disgust, making it unique among the SCM emotions, as a basic, not-just-social emotion.

The Behavior from Intergroup Affect and Stereotypes (BIAS) map predicts behavioral orientations to groups based on their perceived warmth and competence (Cuddy, Fiske, & Glick, 2007). Groups appearing high on both SCM dimensions receive active and passive facilitation. The mixed quadrants receive both positive and negative behavior: passive facilitation and active harm go to those who elicit envy, whereas active facilitation and passive harm go to those who elicit pity. Groups perceived as low in warmth and competence and who elicit disgust are subjected to both active and passive harm—behaviors consistent with historical and present day examples of dehumanization.

Neural Data

Targets from SCM social groups all elicit activity in a brain region reliably implicated in mentalizing and social cognition—the medial prefrontal cortex (mPFC; *see* Amodio & Frith, 2006)—with the notable exception of social targets who elicit the negative basic emotion disgust (Harris & Fiske, 2006). These apparently dehumanized targets instead elicit activity in two separate neural regions, the insula (Harris & Fiske, 2006; Harris & Fiske, 2009) and amygdala (Harris & Fiske, 2009), consistent with a disgust response. In these studies, participants saw pictures of a social target for either 2 seconds (Harris & Fiske, 2006) or 500 milliseconds (Harris & Fiske, 2007) after a 12-second fixation cross in a slow event-related design. The task was simply to indicate via button press which of the four SCM emotions participants felt toward the person: pride, envy, pity, or disgust. In both studies, participants assigned each emotion to the corresponding pictured social target from that respective SCM space at a rate well above chance, suggesting that the in-scanner ratings agreed with the pretest ratings of these social group members representative in the SCM space.

The neural activity for the each social target categorized by emotion was compared to the neural activity during the fixation baseline and to each other category in 3:1 contrasts. Only the social targets that elicited disgust failed to generate activity in the mPFC above the fixation baseline and compared to the other social targets. This is an indication of *dehumanized perception*—a result that suggests participants do not think about the minds of these dehumanized targets to the same extent as other social targets. There was also a small but contiguous overlap of mPFC voxels in response to the other social targets who did activate the area above baseline. A region-of-interest analysis of this overlap indicated that the effect size for activity related to dehumanized targets in this area was significantly smaller than activity for the other social targets. This evidence suggests that dehumanized targets are processed at a different end of the humanized-dehumanized continuum compared to all other social targets.

Nevertheless, to elicit mPFC activity to dehumanized targets, we asked some participants to engage in a mentalizing process about the targets. Participants asked to infer either the vegetable preference (like/dislike) or age category (over middle-aged/under middle-aged) show greater mPFC activity when judging these same dehumanized targets' preference than when categorizing them (Harris & Fiske, 2007). The other social targets from the SCM space all elicit mPFC activity, regardless of judgment task. All this evidence suggests that the functional significance of the mPFC may relate to the question of perceived humanity. We next explore the mPFC more closely and test the hypotheses generated by its function.

Experimental Social Psychological Evidence for Dehumanized Perception

Medial Prefrontal Cortex

The mPFC is a large strip of frontal cortex anterior to the cingulate. It functions as a socially tuned area of a reward network that indexes mentalizing processes and social cognition (Harris, McClure, van den Bos, Cohen, & Fiske, 2007; van den Bos et al., 2007). Social neuroscience reliably finds mPFC activity when participants think about social stimuli (Amodio & Frith, 2006). This is not surprising, if people are intrinsically rewarding (Harris et al., 2007; *see also* Fiske, 2004, pp. 23–24; Kwan et al., 2004; Sears, 1983; Taylor & Brown, 1988; Taylor & Gollwitzer, 1995). The mPFC is divided into three functionally and connectively distinct regions—posterior medial frontal cortex (pMFC), anterior medial frontal cortex (aMFC), and orbital medial frontal cortex (oMFC). The pMFC and aMFC share the Talairach boundary y = 10, whereas the aMFC and oMFC share the Talairach boundary z = 2. Furthermore, the pregenual cingulate is a unique region within mPFC, subsumed as part of the aMFC but distinguished from more superior and anterior parts of that region of cortex. The majority of mentalizing and social cognition tasks activate the aMFC, including the pregenual cingulate (Amodio & Frith, 2006). These labels will be used for the remainder of the chapter to refer to and distinguish these functionally distinct areas of cortex.

Reliably, the mPFC, along with the superior temporal sulcus (STS) activate in mentalizing tasks (Abu-Akel, 2003, Blakemore et al., 2003, Brunet et al., 2000, Calarge et al., 2003, Castelli et al., 2000, Frith & Frith, 2001, Gallagher & Frith, 2002, Sabbagh, 2004; Saxe, Carey, & Kanwisher, 2004; Saxe & Wexler, 2005), including dispositional attribution (Harris et al., 2005). Social neuroscience work also includes these areas in face perception, person perception, and impression formation (Harris & Fiske, 2006; Haxby, Gobbini, & Montgomery, 2004; Todorov, Gobbini, Evans, & Haxby, 2007).

Greater mPFC activation appears in social compared to nonsocial cognition. For example: *(1)* social cognition tasks in which participants form an impression of a person versus an object (e.g. Mason & Macrae, 2004; Macrae, Heatherton, & Kelley, 2004; Mitchell, Banaji, & Macrae, 2005); *(2)* reactions involving familiarity and interpersonal affect (Gobinni & Haxby, 2007; Gobinni, Leibenluft, Santiago, & Haxby, 2004; Haxby, Gobbini, & Montgomery, 2004; Leibenluft, Gobbini, Harrison, & Haxby, 2004; Ochsner et al., 2004); *(3)* thinking about the self (Macrae et al., 2004); *(4)* personal (vs. impersonal) moral judgments (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001); *(5)* thinking about other players in games involving trust and second-guessing of their decisions (Gallagher, Jack, Roepstorff, & Frith, 2002; McCabe, Houser, Ryan, Smith, & Trouard, 2001; Sanfey, Rilling, Nystrom, & Cohen, 2003); and *(6)* mentally navigating the social versus physical world (Kumaran & Maguire, 2005) all greater activate regions of the mPFC. Although they span a variety of areas, these studies all require thinking about the minds of people. Evidence therefore seems to converge on the mPFC as necessary for thinking about people and associated mentalizing processes.

Self-Report Data

Experimental social psychological evidence also shows that people do not think about the mental state of dehumanized targets (Harris & Fiske, 2009). Participants in a between-subjects design saw a picture of a social target from one of the four quadrants of the SCM space. Participants first described a day in the life of the social target. Participants next rated these social targets on a number of dimensions, including ease of mentalizing them and ease of inferring their disposition. We performed 3:1 contrasts comparing the dehumanized targets with the average response to the three other social targets. Participants used fewer verbs that required mental state inference (e.g., quench vs. drink) to describe the dehumanized targets compared to other social agents. Participants also rated dehumanized targets as more difficult to mentalize and to infer dispositions (Harris & Fiske, 2009). This evidence helps demonstrate dehumanized perception.

Theories of Dehumanization

Allport (1954) described dehumanization as the worst type of prejudice, excluding out-groups from full humanity. Implicit in this and explicit in more modern accounts is the idea that extreme prejudice reduces the target to less than human, sometimes as an animal and sometimes an automaton (Haslam, 2006). Social psychological theory underscores the idea of perceiving some out-groups as less than people: Bar-Tal (1989) theorizes that groups acting outside societal norms would be excluded from other human groups. Struch and Schwartz (1989) argue that all out-groups allegedly possess a lesser degree of humanity than the in-group. Staub (1989), in discussing evil, often speaks of moral exclusion—the belief that some social groups operate beyond moral rules and values (cf. Opotow, 1990). All these theories share the point that people may think of out-groups in a significantly different way than they think of in-groups and themselves.

Most relevant to *dehumanized perception* is the work of Jacques-Philippe Leyens and colleagues on out-group infra-humanization[1] (Demoulin et al., 2005; Leyens et al., 2001, 2003). The theory posits that out-groups are believed not to experience, thus are not attributed, complex human emotions. Participants in these studies are willing to attribute negative and even positive basic emotions to these out-groups but not complex secondary emotions[2] (Leyens et al., 2001, 2003). Therefore, an enemy may feel sadness but not regret.

Dehumanization has been divided along the dimensions of typical humanity and unique humanity (Haslam, 2006). Typical humanity describes the aspects of being human fundamental to lay definitions of humanity, such as complex emotions. Unique humanity describes aspects of humanity not shared by other species, such as intelligence and language (Haslam, 2006). Denial of either of the characteristics leads to a perception of the social target as not human—they are reduced to the level of automata and animals respectively.

Rating Data on Dehumanization Dimensions

The various theories of dehumanization suggest a number of rating dimensions that should dissociate such targets. In the same study that measured mentalizing (Harris & Fiske, 2009), participants also rated the social targets on various dimensions of dehumanization—complex emotions, typical humanity, and uniquely human characteristics (intelligent and articulate). Participants rated dehumanized targets who elicit the basic emotion disgust as lower on uniquely human characteristics (articulate, intelligent), and less typically human. This all suggests that dehumanized targets are denied some aspects of unique and typical humanity.

Social targets who elicit envy are sometimes perceived as not fully human in other ways. These out-group social targets serve as a comparison group for the dehumanized targets. Interestingly, participants rated social targets who elicited the ambivalent emotion envy as lower on complex emotions and lower on typical humanity, but higher on the uniquely human dimensions—articulate and intelligence—than other social targets from the SCM space. This suggests that these social targets are not perceived as typically human to the same extent as other social targets—they may be seen as automata. However, these social targets do elicit envy, an ambivalent emotion that entails both disliking and respect.

This stands in contrast to dehumanized targets, who are lower on both unique and typical humanity. This more dramatic denial of humanity seems to be an extreme form of prejudice. Note that all these perceptions occur on a continuum from most human to least human. Participants realize rationally that homeless people are literally human, but they respond to them as if they are not.

[1] The concept of dehumanized perception discussed in this chapter is similar to, but not the same as, the concept of infrahumanization. Infrahumanization surrounds the *attribution* of secondary emotions to outgroups whereas dehumanized perception involves extreme outgroups not *eliciting* complex social emotions in the perceiver. This latter concept does not involve a denial of affect; in fact the affect elicited by the extreme group may drive the entire processes. Therefore, dehumanized perception is, as the name implies, a perceptual cognitive process possibly mediated by an affective reaction to the perceived social group of the target.

[2] Secondary emotions require higher cognitive ability and are only available later in cognitive development. Complex social emotions refer to emotions that can only be elicited in the actual, imagined, or implied presence of other human beings. Although there is a lot of overlap between these two definitions of higher-order emotions, the distinction is made to emphasize the social nature of some higher-order emotions central to the concept of dehumanized perception.

Contact moderates dehumanized perception

Our social neuroscience approach demonstrates thus far that dehumanized perception involves both a failure of mentalizing and less attribution of humanity. Other tasks beyond mentalizing activate the mPFC (*see* Amodio & Frith, 2006). Similarity and familiarity—the psychological dimensions examined in some of these tasks—correlate with intergroup contact (Allport, 1958; Islam & Hewstone, 2001; Sherif & Sherif, 1938). Additionally, because dehumanized perception appears to be a form of prejudice, contact as a moderator of prejudice may also moderate dehumanized perception. In particular, this section focuses on contact, a variable at the overlap of the neural and social psychological literatures.

Prior Evidence Related to Contact

If dehumanized perception is indeed a form of prejudice, and social psychological theory suggests that the prejudice can be reduced by intergroup contact (Allport, 1958; Islam & Hewstone, 2001; Sherif & Sherif, 1938), then perhaps contact with dehumanized targets may reduce it. This contact hypothesis suggests that low familiarity and similarity may distinguish dehumanized targets. Additionally, the neural literature suggests that similarity and familiarity are related to mPFC function.

Familiarity

Familiarity generates positive affect. For example, mere exposure demonstrates that simply repeated conscious or unconscious exposure to a neutral stimulus enhances liking for it (Murphy, Monahan, & Zajonc, 1995). The most familiar person is the self, and people prefer letters associated with their initials (Pelham, Mirenberg, & Jones, 2002). Friends have an easier time inferring each other's thoughts and feelings than strangers do (Stinson & Ickes, 1992). Therefore, positive affect related to the self is generated when an attitude object is familiar.

Familiarity breeds liking in neural data as well. A few imaging studies of familiarity in particular, and positive social affect more generally, nicely illustrate mPFC activity associated with familiarity. Greater mPFC activity is observed when mothers look at faces of their own child rather than other children, and mothers' activation are also greater when they look at familiar rather than unfamiliar children (Gobbini & Haxby, 2007; Gobbini, Leibenluft, Santiago, & Haxby, 2004; Leinbenluft et al., 2004). In addition, participants given an immediate reward for performance also exhibit increased mPFC activity (McClure et al., 2004)—a source of positive affect. Thus, familiarity and the positive affect it generates may also be generated in a social contact experience.

Similarity

Research within experimental social psychology on similarity (e.g., self-other biases, self-referential effect, self-esteem, in-group bias; *see* Fiske & Taylor, 2008, for a review) has illustrated that the self serves as a positive attitude-object. Things similar to the self become associated with positive affect, and even objects inherit additional value when owned (the endowment effect; Thaler, 1980).

The work on similarity within social neuroscience links the mPFC and thinking about the self. A number of studies show mPFC activation in tasks where participants reflect on themselves, access self-knowledge, or compare the self to another (Johnson et al., 2002; Kelley et al., 2002; Lieberman, Jarcho, & Satpute, 2004). Self-regulation of affect also activates the mPFC (Ochsner et al., 2004). Finally, self-reflection may be a tool that allows one to infer the mental states of others (Heal, 1986), and it does activate the mPFC, along with the posterior cingulate and precuneus (Mitchell, Banaji, & Macrae, 2005).

Rating Data on Familiarity, Similarity, and Contact

Dehumanized targets are rated significantly less familiar. A lack of familiarity with the dehumanized targets may help account for the lack of mentalizing. Relatedly, the in-group social targets who elicit pride and activate the mPFC are rated as more familiar, along with the envied social targets. Both these social targets elicit the largest mPFC effect (Harris & Fiske, 2006). Thus, familiarity may moderate dehumanized perception.

Dehumanized targets are also rated as less similar. Social targets who elicit the in-group emotion pride are rated as more similar, along with social targets who elicit envy. As just noted, both these social targets elicit the largest mPFC effect (Harris & Fiske, 2006). Thus, similarity also may moderate dehumanized perception.

Recent data have also revealed that participants are less likely to interact with dehumanized targets (Harris & Fiske, 2009). This evidence all suggests that contact with dehumanized targets may moderate dehumanized perception. Certainly, a lack of contact deprives people of any opportunity to know another person—that person's preferences, habits, thoughts, aspirations, flaws, and so forth—the mental states that moderate perceived humanity.

Conclusion

We proposed that people do not much consider the minds of some social targets. We combined evidence from neuroscience and social psychology in an account of dehumanized perception. We described neuroscience studies whose data were predicted by social psychology, showing less mPFC activity to the dehumanized targets. We also described experimental social psychological studies whose data were predicted by neuroscience, demonstrating fewer mental state inferences and lower ratings on subjective mentalizing. Furthermore, we have outlined possible dimensions along which dehumanized perception may occur, in the hope of finding moderating mechanisms that may ameliorate this extreme form of prejudice.

Social targets who elicit the negative basic emotion disgust elicit reduced mPFC activity in participants, but individuating processes re-activate this area of cortex. Similarly, participants do not think about the minds of dehumanized targets—a function associated with the mPFC—and they describe dehumanized targets as more difficult subjects for mental inference processes. Finally, dehumanized targets, consistent with both the social psychological and neuroscience literatures, are seen as dissimilar, unfamiliar, and not uniquely human. Social targets who do activate the mPFC are perceived as more similar, familiar, and uniquely human.

The neural data possibly reveal something important about both mPFC function and social psychological theory. Because dehumanized targets elicit the basic negative emotion disgust, rather than a more complex uniquely social emotions such as pride, envy, or pity, the mPFC—already described as an affective area—may be involved in some forms of affect. Perhaps these more complex uniquely social (and partially positive) emotions may also require mental inference. The rating data have dual implications as well, suggesting that because the dehumanized targets are rated as not typically or uniquely human, but as strange and dissimilar, then people who engage in frequent meaningful contact with dehumanized targets may not show this effect. These data altogether suggest that perceived humanity could be defined as at least partially positive feelings resulting from meaningful contact with people. This feeling is accompanied by spontaneous thinking about their minds. Thus, perceiving another's humanity entails mPFC activity that may signal that a target is person. This does not argue that the mPFC has a purely social function but that it may discriminate relatively more human from relatively less human targets.

We are continuing to explore the function of the mPFC with neuro-imaging studies aimed at differentiating its role in person perception from its many other functions (e.g., positive affect). In the process, we have created a scale of dehumanized perception and a database of social targets that reliably elicit the SCM emotions. These tools can be used in future research on this topic. Thus, our social neuroscience beginnings promise a line of research aimed at understanding a basic human process that overlaps various disciplines.

Tragically, history has shown how possible it is for people to dehumanize others. Historical anecdotes suggest that people have the capacity to perform extreme acts of violence against others following a dehumanized perception of the victims. The metaphor of Jews and Tutsis as cockroaches in Nazi Germany and war-torn

Rwanda, respectively, Blacks as three fifths of a person in the U.S. Constitution, and Iraqi prisoners portrayed as dogs and beasts of burden by their American and British torturers are but a few instances.[3] Additional empirical scientific evidence is necessary before claiming that dehumanized perception, a failure to think about the mind of another, is a necessary component for these extreme acts of violence and prejudice. But the historical evidence certainly suggests the phenomenon.

Imagine a utopia without violence, racism, and all the associated social ills—a place where each person is treated as a unique individual yet is liked and respected. Utopian fantasies aside, social neuroscience does hold the promise of helping to make such a place just a bit more of a reality. Although it will not "solve all of the world's ills," as one enthusiastic undergrad once exclaimed in an introductory class, social neuroscience thus far has allowed a more comprehensive understanding of how people think about people. In building our case of dehumanized perception, we have relied on research concerning person- and face-perception, prejudice, moral reasoning, self-perception, attribution, mentalizing, and a host of other topics.

To function in an interdisciplinary manner, philosophy of mind and philosophy of science—the bedrock fields for psychological and neuroscience inquiry—must communicate between two distinct ways of understanding human behavior (*see* Adolphs, 2004). Philosophy of mind in this instance provided a theoretical framework, a thought-experiment provided a premise, and empirical experiments tested hypotheses generated from both neuroscience and social psychological experimental data. In the process, this chapter addressed perceived humanity. Although an answer to the more abstract question of defining humanity may certainly not emerge, scientific knowledge generated along the way that can serve as a point of departure for other eager travelers along this road.

[3] Also, perpetrators of extremely immoral acts are often reported as not human (although for an argument as well that they are soulless, *see* the Chapter 18 in this volume).

References

Abu-Akel, A. (2003). A neurobiological mapping of theory of mind. *Brain Research Reviews, 43*, 29–40.

Adolphs, R. (2004). Emotion, social cognition, and the human brain. In J. Cacioppo, & G. Berntson (Eds.), *Essays in Social Neuroscience*. Cambridge, MA: MIT Press.

Allport, G. W. (1954). *The Nature of Prejudice*. Reading, MA: Addison-Wesley.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64*, 431–441.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews, Neuroscience, 7*, 268–277.

Appiah, K. A. (2003). Thinking it through: An introduction to contemporary philosophy. Bar-Tal, D. (1989). Delegitimization: The extreme case of stereotyping and prejudice. In D. Bar-Tal, C. Graumann, A. Kruglanski, & W. Stroebe (Eds.), *Stereotyping and Prejudice: Changing Conceptions*. New York: Springer-Verlag.

Batson, D. (1997). Is empathy-induced helping due to self-other merging? *Journal of Personality and Social Psychology, 73*, 495–509.

Blakemore, S. J., et al. (2003). The detection of contingency and animacy from simple animations in the human brain. *Cerebral Cortex, 13*, 837–844.

Brewer, M. B. (1979). Ingroup bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin, 86*, 307–324.

Brunet E., Sarfati, Y., Hardy-Bayle, M. C., & Decety, J. (2000). A PET investigation of the attribution of intentions with a nonverbal task. *NeuroImage, 11*, 157–166.

Calarge, C., Andreasen, N. C., & O'Leary, D. S. (2003). Visualizing how one brain understands another: A PET study of theory of mind. *American Journal of Psychiatry, 160*, 1954–1964.

Castelli, F., Happe, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage, 12*, 314–325.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology, 92*, 631–648.

Demoulin, S., Torres, R. R., Perez, A. R., et al. (2005). Emotional prejudice can lead to infra-humanisation. *European Review of Social Psychology, 15*, 259–296.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57*, 1082–1090.

Ferguson, T. J., & Wells, G. L. (1980). Priming of mediators in causal attribution. *Journal of Personality and Social Psychology, 38*, 461–470.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2006). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Science, 11*, 77–83.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*, 878–902.

Fiske, S. T., Harris, L. T., & Cuddy, A. J. (2004). Why ordinary people torture enemy prisoners. *Science, 306*, 1421–1632.

Fiske, S. T., Lin, M., & Neuberg, S. (1999). The continuum model: Ten years later. In Y. Trope & S. Chaiken (Eds.), *Dual-process Theories in Social Psychology* (pp. 231–254). New York: Guilford Press.

Fiske, S. T., & Taylor, S. E. (2008). *Social Cognition: From Brains to Culture*. New York: McGraw-Hill.

Frith, U., & Frith, C. (2001). The biological basis of social interaction. *Current Directions in Psychological Science, 10*, 151–155.

Gallagher, H. L., & Frith, C. D. (2002). Functional imaging of 'theory of mind.' *Trends in Cognitive Sciences, 7*, 77–83.

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia, 45*, 32–41.

Gobbini, M. I., Leibenluft, E., Santiago, N., & Haxby, J. V. (2004). Social and emotional attachment in the neural representation of faces. *NeuroImage, 22*, 1628–1635.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105–2108.

Harris, L. T., & Fiske, S. T. (2003). BOLD activations to Black and White faces under different social goal conditions. Unpublished data.

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the lowest of the low: Neuro-imaging responses to extreme outgroups. *Psychological Science,17*, 847–853.

Harris, L. T., & Fiske, S. T. (2007). Social groups that elicit disgust are differentially processed in mPFC. *Social Cognitive Affective Neuroscience, 2*, 45–51.

Harris, L. T., & Fiske, S. T. (2008). The brooms in Fantasia: Neural correlates of anthropomorphizing objects. *Social Cognition, 26*, 210–223.

Harris, L. T., & Fiske, S. T. (2009). Dehumanized perception: The social neuroscience of thinking (or not thinking) about disgusting people. In M. Hewstone & W. Stroebe (Eds.), *European Review of Social Psychology* (Vol. 20, pp. 192–231). London: Wiley.

Harris, L. T., McClure, S. M., van den Bos, W., Cohen, J. D., & Fiske, S. T. (2007). Regions of MPFC differentially tuned to social and non-social affective evaluation. *Cognitive and Behavioral Neuroscience, 7*, 309–316.

Harris, L. T., Todorov, A., & Fiske, S. T. (2005). Attributions on the brain: Neuro-imaging dispositional inferences beyond theory of mind. *Neuroimage, 28*, 763–769.

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. *Brain Imaging, 11*, 2351–2355.

Haslam, N. (2006). *Dehumanization: an integrative review. Personality and Social Psychology Review, 10*, 252–264.

Haxby, J. V., Gobbini, M. I., & Montgomery, K. (2004). Spatial and temporal distribution of face and object representations in the human brain. In M. Gazzaniga (Ed.), *The Cognitive Neurosciences*. Cambridge, MA: MIT Press.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York: Wiley.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57*, 243–259.

Jones, E. E. (1979). The rocky road from acts to dispositions. *American Psychologist, 34*, 107–117.

Kelley, H. H. (1972). Attribution in social interaction. In E. E. Jones, D. E. Kanouse, H.H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the Cause of Behavior* (pp. 1–26). Hillsdale, NJ: Lawrence Elbaum & Associates.

Kumaran, D., & Maguire, E. A. (2005). The human hippocampus: Cognitive maps or relational memory? *Journal of Neuroscience, 25*(31), 7254–7259.

Leibenluft, E., Gobbini, M. I., Harrison, T., & Haxby, J. V. (2004). Mothers' neural activation in response to pictures of their children and other children. *Biological Psychiatry, 56*(4), 225–232.

Leyens, J. Ph., Cortes, B. P., Demoulin, S., et al. (2003). Emotional prejudice, essentialism, and nationalism. *European Journal of Social Psychology, 33*, 703–718.

Leyens, J. P., Rodriguez-Perez, A., Rodriguez-Torres, R., et al. (2001). Psychological essentialism and the differential attribution of uniquely human emotions to ingroups and outgroups. *European Journal of Social Psychology, 31*, 395–411.

Lieberman, M. D., Hariri, A., Jarcho, J. M., Eisenberger, N. I., & Bookheimer, S. Y. (2005). An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. *Nature Neuroscience, 8*, 720–722.

McArthur, L. Z. (1972). The how and what of why: Some determinants and consequences of causal attribution. *Journal of Personality and Social Psychology, 22*, 171–193.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences USA, 98*(20), 11,832–11,835.

Macrae, C. N., Heatherton, T. F., & Kelley, W. M. (2005). A self less than ordinary: The medial prefrontal cortex and you. In M. Gazzaniga (Ed.), *The Cognitive Neurosciences.* Cambridge, MA: MIT Press.

Mason, M. F., & Macrae, C. N. (2004). Categorizing and individuating others: The neural substrates of person perception. *Journal of Cognitive Neuroscience, 16*(10), 1785–1795.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience, 17*(8), 1306–1315.

Murphy, S. T., Monahan, J. L.., & Zajonc, R. B. (1995). Additivity of nonconscious affect: Combined effects of priming and exposure. *Journal of Personality and Social Psychology, 69*, 589–602.

Ochsner, K. N., Knierim, K., Ludlow, D. H., et al. (2004). Reflecting upon feelings: An fMRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience, 16*, 1–27.

Opotow, S. (1990). Moral exclusion and injustice: An introduction. *Journal of Social Issues, 46*, 1–20.

Pelham, B. W., Mirenberg, M. C., & Jones, J. T. (2002). Why Susie sells seashells by the seashore: Implicit egotism and major life decisions. *Journal of Personality and Social Psychology, 82*, 469–487.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., et al. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience, 12*, 729–738.

Rosenberg, S., Nelson, C., & Vivekananthan, P. S., (1968). A multidimensionsal approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9*, 283–294.

Sabbagh, M. A. (2004). Understanding orbitofrontal contributions to theory of mind reasoning: Implications for autism. *Brain and Cognition, 55*, 209–219.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science, 300*, 5626, 1755–1758.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology, 55*, 87–124.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporoparietal junction. *Neuropsychologia, 43*(10), 1391–1399.

Semin, G. R., & Fiedler, K. (1988). The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology, 54*, 558–568.

Staub, E. (1989). *The Roots of Evil: The Origins of Genocide and Other Group Violence.* New York: Cambridge University Press.

Stinson, L., & Ickes, W. (1992). Empathetic accuracy in the interactions of male friends versus male strangers. *Journal of Personality and Social Psychology, 62*, 787–797.

Struch, N. & Schwartz, S. H. (1989). Intergroup aggression: Its predictors and distinctness from in-group bias. *Journal of Personality and Social Psychology, 56*, 364–373.

Thaler, R. (1980). Towards a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, 1, 39–60.

Todorov, A., Gobinni, M. I., Evans, K. K., & Haxby, J. V. (2007). Spontaneous retrieval of affective person knowledge in face perception. *Neuropsychologia, 45*, 163–173.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*, 1623–1626.

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actor's faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology, 83*, 1051–1065.

van den Bos, W., McClure, S. M., Harris, L. T., Fiske, S. T., & Cohen, J. D. (2007). Dissociating affective evaluation and social cognitive processes in ventral medial prefrontal cortex. *Cognitive and Behavioral Neuroscience, 7*, 337–346.

Wheeler, M. E., & Fiske, S. T. (2005). Controlling racial prejudice: Social-cognitive goals affect amygdala and stereotype activation. *Psychological Science, 16*, 1, 56–63.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context Variability in automatically activated attitudes. *Journal of Personality and Social Psychology, 83*, 815–827.

Zebrowitz, L. A. (1999). *Reading Faces: Window to the Soul?* Boulder, CO: Westview Press.

CHAPTER 9
Us versus Them: The Social Neuroscience of Perceiving Out-groups

Nalini Ambady & Reginald B. Adams, Jr.

What are the neural processes involved in perceiving out-groups? The pace of inquiry into this topic has picked up considerably since the pioneering neuro-imaging studies conducted in 2000 (Hart et al., 2000; Phelps et al., 2000), yielding several insights into the neural processes involved in perceiving, responding to, and regulating responses to out-groups. The three chapters in the volume elegantly synthesize the social neuroscience work on perceiving out-groups and suggest several areas for further inquiry. We first summarize the insights provided by these chapters and then go on to outlining areas for further inquiry.

Harris and Fiske consider the most extreme out-groups—the dehumanized, who are often regarded as less than human. Their previous pioneering work has shown that both the mental and emotional capacities of the dehumanized are considered to be diminished compared to normal human beings, and they elicit negative emotional reactions from others. Based on this reasoning, Harris and Fiske argue that the perception of dehumanized targets should be associated with activation in the insula, which is the neural response associated with the emotion of disgust and should also be associated with less activity in the neural areas associated with mentalizing and attributing minds and thoughts to others—the medial prefrontal cortex. Indeed, that is what they found, suggesting that dehumanization is associated with distinct neural responses that parallel the behavioral reactions of being disgusted by and attributing less humanity to the most negatively stigmatized out-groups.

Amodio disentangles the influence of automatic and controlled processing in the regulatory processes influencing racial bias against out-groups. He persuasively argues that dual-process models regarding stereotyping and prejudice are too simplistic. His work with his colleagues, using psychophysiological methods such as the startle eye blink and event-related potentials, provides evidence that motivational and regulatory abilities nuance these traditional models. For example, one experiment examined the responses of people who were all egalitarian in their beliefs and values but differed in whether their motivation to respond without prejudice was externally or internally motivated. Those who were internally motivated to respond without prejudice showed a greater regulation of prejudice, as reflected in the ERN, an event-related potential associated with conflict monitoring, than did people who were externally motivated to respond without prejudice. This work suggests that we need to refine and enrich our models of the processes underlying stereotyping and prejudice. Motivational and self-regulatory factors influence both our behavioral as well as neural responses to others.

Ito summarizes her work examining the time-course of processing in-groups and out-groups in the brain using event-related potentials and reports several interesting findings.

Her work shows that neural activity in White subjects reflects early attention being paid to Blacks and males compared to Whites and females. She generally finds similar results even when processing goals and attention being paid to the particular categories are manipulated. She finds similar results with other out-groups as well. Thus, white participants show larger N100 and P200 potentials to Asians and Blacks than to Whites and a larger N200 potential to in-group members. These intriguing results suggest that neural processing and categorization into in-groups and out-groups occurs relatively early in the time-course. Interestingly, similar results were not obtained for racially ambiguous group members, who were initially responded to in the same way as the in-group and were differentiated only later in neural processing.

Thus, the chapters in this section indicate quite clearly that distinct neural responses are associated with the processing of both extreme (those who are dehumanized) as well as less extreme (those who are biracial) out-groups. Moreover, the neural processing of in-group and out-group members occurs quite early.

Compound social cues

As the preceding chapters indicate, most work examining neural processing in relation to group membership has examined responses to single cues such as ethnicity or race rather than the combined influence of multiple social cues. But human social perception is based on a number of different cues. Multiple social messages are simultaneously conveyed and encoded even from a single channel of communication such as the face. The face, for example, can simultaneously convey information about a person's gender, race, age, and attention. The physiognomy of the face, such as the facial maturity of the face (*see* Zebrowitz, 1997), is in itself a cue that influences social perception. Emotional expressions and eye gaze conveyed by the face also affect social perception. In the next section we consider the separate and combined effects of these three different facial cues: *(1)* facial appearance; *(2)* eyes gaze, and *(3)* emotion on the perception of group membership.

Facial appearance

One common feature of each chapter included in this section is the use of the human facial appearance as a primary vehicle for investigating intergroup perception. Ito's chapter focuses on early categorization based on facial appearance cues, Harris and Fiske's examines dehumanized perception that can result once someone is categorized as belonging to an out-group, and Amodio discusses how we regulate negative bias arising from such stereotype activation. As the use of facial stimuli readily lends itself to neuroscientific inquiry, discussion of face processing in relation to intergroup perception warrants further commentary.

Because face processing is a form of visual perception, it is necessarily influenced and constrained by the mechanics governing vision. That said, it is undeniable that social cognition exerts a powerful moderating influence on how we see the world, gating attention, determining whether individuation or categorization occurs, and influencing how we interpret visual cues we perceive. It is clear from the research conducted to date that race is readily detected from facial appearance cues, even as early as 100 milliseconds. However, as discussed below, it is quite clear that social experience impacts the extent to which these early processes influence the individual responses to race.

Race Perception, Face Processing, and the Fusiform

Decades of research has revealed an out-group homogeneity effect in the form of an own-race bias when remembering faces (*see* Sporer, 2001). It has recently been found that this effect is modulated by the speed with which individuals racially categorize out-group relative to in-group faces. The faster individuals categorized out-group faces, the less they individuate them, and thus the better their relative memory performance for in-group faces (Levin, 1996). Golby et al. (2001) further investigated this out-group homogeneity effect by examining fusiform responsivity—purported to be specialized for encoding structural aspects of facial identity—in Black and White participants when viewing photographs of same- versus other-race faces. After viewing the faces during fMRI scanning, participants were given a face recognition task. The ubiquitous own-race memory bias was found. More importantly, the magnitude of this memory effect was directly related to differential activation in the fusiform for in-group versus out-group face processing. One might conclude from this study that the greater fusiform activation is likely related to greater perceptual expertise associated with processing in-group versus out-group faces. However, intergroup contact has not generally been found to reliably explain such out-group homogeneity effects (*see* Eberhardt, 2005). Thus, these effects are arguably more likely driven by an affective, motivated process than an incidental cognitive process.

Race Perception, Perceived Threat, and the Amygdala

Hart et al. (2000) found that in White participants, amygdala responses habituated more quickly to White than Black faces. They offered a couple of explanations for this effect: *(1)* greater familiarity with in-group faces, consistent with the perceptual expertise account, or *(2)* racial bias toward out-group members, consistent with a more affective, motivated account. In a related study conducted around the same time, Phelps et al. (2000) discovered that White participants, although not revealing a difference in amygdala response to Black versus White faces when subjected to a direct comparison, did show a correlation between amygdala response to Black faces and their implicit attitudes toward Blacks, measured using the Implicit Association Test (IAT). This work therefore suggests responses that result less from perceptual expertise than from learned emotional associations. Importantly, other work has demonstrated that variation in neural responsivity to in-group versus out-group faces also appears to be influenced by conscious attempts to control such racial bias (Cunningham et al., 2004; Richeson et al., 2003).

The studies reviewed above examined responses that occur when facial appearance gives rise to categorization according to race membership. However, the specific influence that race prototypical appearance plays in neural responsivity to faces remains to be examined. Several behavioral studies have demonstrated modulation in the affective responses of White participants to Black faces, high versus low in Afro-centric features (Blair, Judd, & Fallman, 2004; Eberhardt, Dasgupta, & Banaszynski, 2003; Livingston & Brewer, 2002; Maddox, 2004). Further these effects appear to operate separately from race categorization. For example, the differential presence of Afro-centric features has been found to profoundly influence responses to faces even when clearly categorized as White (Blair, Judd, & Fallman, 2004). Future neuroscientific inquiry into the role of such facial appearance cues in intergroup perception will help to better clarify our understanding of the mental operations underlying racial categorization and stereotype activation.

Eye gaze

The eyes are generally believed to be one of the most powerful channels of human social communication (Emery, 2000), with the perception of gaze often argued to play a critical role in the development of Theory of Mind (ToM; Baron-Cohen, 1995). Eye contact has been found to yield greater Galvanic Skin Response (Nichols & Champness, 1971), EEG arousal (Gale, Lucas, Nissin, & Harpham, 1972; Gale, Kingsley, Brookes, & Smith, 1978; Gale, Spratt, Chapman, & Smallbone, 1975), increased heart rate (Kleinke & Pohlen, 1971), and increased amygdalar responsivity (Kleinke & Pohlen, 1971; Kawashima et al., 1999) in observers. Amygdala damage has been found to undermine the ability to orient to the eyes of another during social communication (Adolphs et al., 2005) and consequently to follow another's direction of attention (Akiyamaet al., 2007). Furthermore, a number of recent studies have demonstrated differential cortical (Hori et al., 2005) and amygdala responsivity (Adams et al., 2003; Sato et al., 2004) to threat displays as a function of gaze direction (direct versus averted), suggesting that eye gaze combines with specific facial expressions to modulate threat responses.

A set of recent studies by Adams and his colleagues (Adams & Kleck, 2003, 2005; Adams, Gordon, Baird, Ambady, & Kleck, 2003, Hess, Adams & Kleck, 2007; *see also* Ganel & Goshen-Gottstein, Goodale, 2005; Graham & Labar, 2007; Sander, Granjean, & Kaiser, 2007) serves to further illustrate and clarify the important role that gaze can play in emotion processing. Adams and Kleck (2003; 2005), for example, demonstrated that anger faces coupled with direct gaze (approach signals) and fear faces coupled with averted gaze (avoidance signals) are perceived as more intense and recognized more quickly and accurately than anger faces coupled with averted or fear faces coupled with direct gaze (cf. Hess, Adams, & Kleck, 2007; Sanders, Grandjean, Kaiser, Wehrle, & Scherer, 2007). A recent study by Fox et al. (2007) revealed that emotional expressions can also influence gaze processing *even* at the level of attention allocation. Specifically, they found that for participants high in trait anxiety, fear coupled with averted gaze yielded greater reflexive attentional shifts in observers compared to that found for either anger or neutral expressions, whereas anger coupled with direct gaze yielded greater attention capture effects than either fear or neutral expressions.

A recent study by Richeson, Todd, Trawalter, and Baird (2008) directly examined the influence of direct versus averted eye gaze in intergroup perception. In this paper, they articulated a similar rationale for predicting the influence of gaze in amygdala responsivity during race perception as that described above for the role of eye gaze in emotion perception. Because direct eye gaze signals approach, as does the hostile intent stereotypically associated with Black males, they argued that eye gaze and race should be expected to combine to communicate a heightened threat response. Their results support this contention, demonstrating greater amygdala responsivity in White participants when viewing Black relative to White faces but only when combined with direct eye gaze. In a related study, they also found that direct gaze coupled with Black versus White faces selectively captured attention to a greater degree (Trawalter, Todd, Baird, & Richeson, 2008). More recently eye gaze was found significantly impact the otherwise ubiquitous cross-race memory effect in the effect was only apparent for direct-gaze faces (Adams, Pauker, & Weisbuch, 2010). These findings therefore further highlight the importance of considering the combined influence of compound social cues during intergroup perception.

Emotional expressions

Another cue that has received considerable attention in social and affective neuroscience is that of emotion or affect. Numerous studies have investigated neural activation in response to emotional displays and have yielded some fascinating findings (for a meta-analytic review, *see* Phan et al., 2002).

Recently, social psychologists have turned their attention to the effects of emotional displays on the perception of in-group and out-group members. For example, Hugenberg and Bodenhausen (2003) investigated whether target race would moderate the recognition advantage for happy faces. White participants took part in an emotion recognition task which used computer-generated Black and White faces displaying anger, joy, and sadness as stimuli. Participants were quicker at evaluating happiness in White targets compared to Black target faces but were faster at categorizing the angry and sad expressions of Black target faces. The results of this study suggest the race of the target individual displaying the emotion and the type of emotion being displayed both affect emotion recognition. Ackerman and colleagues (2006) have also investigated the role of emotion—particularly anger—in eliminating out-group homogeneity effects. White participants were shown White and Black faces displaying angry or neutral expressions and later completed a memory task for the previously presented faces. Results showed that the out-group homogeneity bias was apparent for neutral faces but not for angry Black faces. In addition, when participants were placed under cognitive constraints, angry Black faces were more accurately recognized than angry White faces, suggesting out-group heterogeneity. Taken together, these studies suggest that race can play and important role in emotion recognition.

But very few studies so far have investigated how emotional displays affect the perception of out-groups at the neural level. In one study, Chiu, Deldin, and Ambady (2004) examined responses of high- and low-prejudiced individuals to combinations of group membership as well as emotion. The index of neural processing was the contingent negative variation (CNV) component of the event-related brain potential. The CNV is a slow negative ERP elicited by a warning stimulus that requires anticipation of a target stimulus (Walter, Cooper, Aldridge, McCallum, & Winter, 1964; Picton & Hillyard, 1988). The component is quantifiable into two distinct subcomponents: an "early" CNV and a "late" CNV (Rohrbaugh, Syndulko, & Lindsley, 1976). The early CNV is thought to index initial attention to the information carried by the warning stimulus, the expected degree of expenditure of cognitive effort to respond to the target stimulus, and the degree of motivation to respond to the target stimulus (Low & McSherry, 1968; Forth & Hare, 1989; Hamon & Seri, 1987). Moreover, the presence of the early CNV is generally thought to be a cortical reflection of controlled, rather than automatic, psychological processes in response to an S1 that requires anticipation of a subsequent S2 (Picton & Hillyard, 1988; Shiffrin & Schneider, 1977). The late CNV is measured just prior to the onset of the target stimulus and reflects the additional contribution of cortical resources required for motor response preparation (Brunia & Damen, 1988; Damen & Brunia, 1994).

High- and low- prejudiced individuals selected on the basis of their responses to the Modern Racism Scale (McConahay, Hardee, & Batts, 1981) were asked to make evaluative judgments of emotionally and racially salient facial stimuli. Specifically, participants were asked to make a socially relevant judgment (i.e., do I want to work with this person?) regarding in- and out-group members. Low-prejudiced participants demonstrated the greatest CNV in anticipation of making evaluative responses of angry black faces than to any other category of faces. This finding is consistent with work showing that individuals monitor automatic reactions to negative stereotypes elicited by out-group stimuli (Bodenhausen & Macrae, 1998; Monteith et al., 1993; Plant & Devine, 1998; Richeson et al., 2003). The high-prejudiced group, in contrast, showed the most decreased CNV in anticipation of angry black targets compared to all other targets, supporting theories suggesting that the individuals high in explicit prejudice may be characterized by a decreased tendency, or motivation, to monitor automatic prejudiced responses to negative stereotypes (e.g., Bodenhausen & Macrae, 1998; Monteith et al., 1993; Plant & Devine, 1998). This works suggests that multiple factors combine to influence neural responses in theoretically meaningful ways, including target race, the emotion expressed by the target, and individual differences among the participants.

Compound social cues and the shared signal hypothesis

Thus far, we have considered the effects of distinct facial cues such as physiognomic (i.e., invariant facial appearance cues) and emotional cues (i.e., variant facial expressions). Little is known about how such cues give rise to the unified perceptions that guide our impressions and interactions with others, but it stands to reason that various forms of social information meaningfully interact, even when from distinct sources such as facial expressions and facial appearance cues. Very few studies have examined the effects of these multiple or compound cues on neural processing. Indeed, as illustrated in Chiu, Deldin, and Ambady's (2004) work described above, interactions in cortical responsivity occurred not only for individual differences in level of prejudice but also as a function of the combined influence of race and the presence of stereotypically congruent emotional expressions (i.e., anger), thus demonstrating the complex interplay of social messages conveyed by the human face in a manner that is theoretically tractable.

Preliminary insight into the combinatorial effects of processing multiple social cues at once has been offered by the work Adams and colleagues, which examines what he has referred to as the *shared signal hypothesis* (e.g., Adams & Kleck, 2005; Adams, Ambady, Macrae, & Kleck, 2006). The shared signal hypothesis is based on an understanding that social visual cues, even ones from distinct sources, share basic low-level signal values, such as warmth or aggression or the likelihood to approach or avoid. The shared signal hypothesis predicts that these cues can combine in congruent or incongruent ways, which should have different consequences on perception. Several studies (e.g., Adams, Ambady, Macrae, & Kleck, 2006; Adams & Kleck, 2003, 2005; Adams, Hess, Kleck, & Wallbott, 2004; Hess, Adams, & Kleck, 2004; Marsh, Adams, & Kleck, 2005) support this contention, demonstrating that social cues such as gaze direction and gender of an expresser can influence the efficiency with which a given emotional display is processed and how it is interpreted when the combination represents congruent versus incongruent signal values.

Chiu, Deldin, and Ambady's (2004) findings are consistent with the shared signal hypothesis in that aggression associated with anger is also stereotypically ascribed to Blacks. As already noted, other recent work similarly demonstrates interactivity based on race and emotion cues (c.f., Hugenberg, 2005; Ackerman et al., 2006) as well as race and eye gaze cues (Richeson et al., 2008), again in a manner consistent with the shared signal hypothesis. Thus, in combination, these powerful social cues should be expected to mutually influence neural activation related to social perception, cognition, and behavior. Yet, very little is known about the cognitive and neural effects of perceiving such compound social cues. The possibility of such a functional correspondence among otherwise distinct social cues, however, offers exciting possibilities for future research in this area that can help illuminate our understanding of intergroup perception and contribute to the emerging literature on compound social cue processing.

Another important consideration related to these issues concerns the specific nature of the information driving social perception, whether exerting an upstream impact on categorical thinking or, rather, a downstream impact driven by categorical thinking. In other words, to what extent and under what conditions do facial cues give rise to category and stereotype activation, and similarly, do category and stereotype activation influence how facial cues are processed and interpreted? In this way, the influence of combined social cues on category and stereotype activation will likely be driven, at least in part, by the number of shared signals, particularly ones that are congruent and stereotypically consistent. Conversely, whether a person is categorized as in-group versus out-group has a powerful top-down influence on whether faces are individuated (Levin, 1996) or susceptible to stereotype activation and consequent overgeneralization to a group or category.

Conclusion: toward an examination of compound cues

The mutual role that certain cues, such as facial appearance, gender, and emotional expression, play in upstream group categorization and stereotype activation merit further investigation. Social psychological research over the past several decades has been consumed by investigating the influence of category memberships in personal construal (Allport, 1954; Bodenhausen & Macrae, 1998), but only recently has work begun to examine the perceptual determinants that give rise to categorization in the first place (e.g., Cloutier & Macrae, 2006). Such research begins to highlight factors that can moderate the extent to which certain social cues give rise to categorization and stereotype activation, whether they do so automatically, and whether such cues are even attended to in the first place. These factors can include differences motivated by individual goals, beliefs, and prejudices that exert top-down influences on what information is attended to, but they can also involve stimulus-based effects, such as the combined influence of expressive information (i.e., anger) and appearance-based cues (i.e., race or gender) that instead exert upstream influences on whether stereotype activation occurs.

Finally, it is important to consider that categorization and stereotype activation from facial information are not mutually inclusive processes. One illustration that makes this point is the impact that race-prototypical appearance can have on stereotype-based social responses above and beyond mere category-based judgments (*see* Blair, Judd, & Fallman, 2004, Livingston & Brewer, 2002).

In fact, categorization is not essential for stereotyping to occur. For example, stereotypical Black facial appearance cues yield negative biases in the context of criminal sentencing even for Whites (i.e., in the absence of overt categorization as Blacks). On the other hand, otherwise identical faces are seen and processed differently based solely on hairstyle—a cue that exerts a powerful top-down indicator of one's racial group membership (MacLin & Malpass, 2001). It is thus essential to consider category- and stereotype-based processing as separate but related.

In sum, what has the social neuroscience approach taught us about perceiving outgroups? Skeptics about social neuroscience should be somewhat appeased when they read the chapters in this section. Even 10 years ago we knew almost nothing about the neural processing of in-groups and out-groups. Since then, the rapid increase is knowledge has been remarkable. We now know that there are distinct neural responses to in-groups and outgroups both in the time-course as well as the areas of the brain. We know that a number of factors affect neural activation in response to in-groups and out-groups, including facial appearance, motivation, emotion, self-regulation, and individual differences. We also know that neural responses that relate to other social cues, such as eye gaze, are impacted by group membership. This work illustrates how social psychology can contribute to knowledge in cognitive neuroscience, in showing how basic cognitive and neural processes involving attention, cognition, emotion, and memory are affected by myriad factors such as social category membership and motivational states, to name a few. But much remains to be discovered, especially about how multiple cues combine to affect neural responses. It is difficult to even imagine the gains in knowledge that will be acquired 10 years from now. These are exciting and heady times for social neuroscience.

Author notes

Work on this chapter was funded, in part, by a NIH R01 MH070833–01A1 grant to N.A.

References

Ackerman, J. M., Shapiro, J. R., Neuberg, S. L., et al. (2006). They all look the same to me (unless they're angry): From out-group homogeneity to out-group heterogeneity. *Psychological Science 17*, 836–840.

Adams, R. B., Jr., Ambady, N., Macrae, C. N., & Kleck, R. E. (2006). Emotional expressions forecast approach-avoidance behavior. *Motivation & Emotion, 30*, 177–186.

Adams, R. B., Jr., Gordon, H. L., Baird, A. A., Ambady, N., & Kleck, R. E. (2003). Effects of gaze on amygdala sensitivity to anger and fear faces. *Science, 300*, 1536.

Adams, R. B., Jr., Hess, U., Kleck, R. E., & Wallbott, H. (2004). The influence of perceived gender on the perception of emotional dispositions. In A. Kappas (Ed.), *Proceedings of the XIth Conference of the International Society for Research on Emotions, 16–20 August 2000, Quebec City* (pp. 17–19). Amsterdam: ISRE publications/University of Amsterdam.

Adams, R. B., Jr., & Kleck, R. E. (2003). Perceived gaze direction and the processing of facial displays of emotion. *Psychological Science, 14*, 644–647.

Adams, R. B., Jr., & Kleck, R. E. (2005). The effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion, 5*, 3–11.

Adams, R. B., Jr., Pauker, K., & Weisbuch, M. (2010). Looking the other way: The role of gaze direction in the cross-race memory effect. *Journal of Experimental Social Psychology, 46*, 478–481.

Adolphs, R., Gosselin, F., Buchanan, T. W., et al (2005) A mechanism for impaired fear recognition after amygdala damage. *Nature, 433*, 68–72.

Akiyama, T., Kato, M., Muramatsu, T., Umeda, S., Saito, F., & Kashina, H. (2007). Unilateral amygdala lesions hamper ***attention****al orienting triggered by gaze direction.* Cerebral Cortex, 17, *2593–2600.*

Allport, G. W. (1954). *The Nature of Prejudice.* Reading, MA: Addison-Wesley.

Amodio, D. M. (in press). Self-regulation in intergroup relations: A social neuroscience framework. To appear in Todorov, A., Fiske, S. T., & Prentice, D. (Eds.), *Social Neuroscience: Toward Understanding the Underpinnings of the Social Mind*. London: Oxford University Press.

Baron-Cohen, S. (1995) *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: MIT press.

Blair, I. V., Judd, C. M., & Fallman, J. L. (2004). The automaticity of race and Afrocentric facial features in social judgments. *Journal of Personality and Social Psychology, 87*, 763–778.

Bodenhausen, G., & Macrae, N. (1988). Stereotype activation and inhibition. In R. Wyer (Ed.), *Advances in Social Cognition* (pp. 1–52). New Jersey: Erlbaum.

Brunia, C., & Damen, E. (1988). Distribution of slow-potentials related to motor preparation & stimulus anticipation in a time estimation task. *Electroencephalagraphy and Clinical Neurophysiology, 69*, 234–243.

Chiu, P., Ambady, N., & Deldin, P. (2004). CNV in response to emotional in- and out-group stimuli differentiates high- and low-prejudiced individuals. *Journal of Cognitive Neuroscience, 16*, 1830–1839.

Cloutier, J., Mason, M. F., & Macrae, C. N. (2005). The perceptual determinants of person construal: Reopening the social-cognitive toolbox. *Journal of Personality and Social Psychology, 88*, 885–894.

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of Black and White faces. *Psychological Science, 15*, 806–813.

Damen, E., & Brunia, C. (1994). Is a stimulus conveying task-relevant information a sufficient condition to elicit a stimulus-preceding negativity. *Psychophysiology, 31*, 129–139.

Eberhardt, J. L. (2005). Imaging race. *American Psychologist, 60*, 181–190.

Eberhardt, J. L., Dasgupta, N., & Banaszynski, T. L. (2003). Believing is seeing: The effects of racial labels and implicit beliefs on face perception. *Personality & Social Psychology Bulletin, 29*, 360–370.

Emery, N. J. (2000). The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews, 24*, 581–604.

Forth, A., & Hare, R. (1989). The contingent negative variation in psychopaths. *Psychophysiology, 26*, 676–682.

Fox, E., Mathews, A., Calder, A. J., & Yiend, J. (2007). Anxiety and sensitivity to gaze direction in emotionally expressive faces. *Emotion, 7*, 478–486.

Gale, A., Kingsley, E., Brookes, S., & Smith, D. (1978). Cortical arousal and social intimacy in the human female under different conditions of eye contact. *Behavioural Processes, 3*, 271–275.

Gale, A., Lucas, B., Nissim, R., & Harpham, B. (1972). Some EEG correlates of face-to-face contact. *British Journal of Social & Clinical Psychology, 11*, 326–332.

Gale, A., Spratt, G., Chapman, A. J., & Smallbone, A. (1975). EEG correlates of eye contact and interpersonal distance. *Biological Psychology, 3*, 237–245.

Ganel, T., Goshen-Gottstein, Y., & Goodale, M. A. (2005). Interactions between the processing of gaze direction and facial expression. *Vision Research, 45*, 1191–1200.

Golby, A. J., Gabrieli, J. D. E., Chiao, J. Y., & Eberhardt, J. L. (2001). Differential responses in the fusiform region to same-race and other-race faces. *Nature Neuroscience, 4*, 845–850.

Graham, R., & Labar, K. S. (2007). Garner interference reveals dependencies between emotional expression and gaze in face perception. *Emotion, 7*, 296–313.

Hamon, J., & Seri, B. (1987). Relation between warning stimuli and contingent negative variation in man. *Activitas Nervosa Superior, 29*, 249–256.

Harris, L. T., & Fiske, S. T. (in press). Perceiving humanity or not: A social neuroscience approach to dehumanized perception. To appear in Todorov, A., Fiske, S. T., & Prentice, D. (Eds.), *Social Neuroscience: Toward Understanding the Underpinnings of the Social Mind*. Oxford University 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.

Hess, U., Adams, R. B., Jr., & Kleck, R. E. (2004). Dominance, gender and emotion expression. *Emotion, 4*, 378–388.

Hess, U., Adams, R. B., Jr., & Kleck, R. E. (2007). Looking at you or looking elsewhere: The influence of head orientation on the signal value of emotional facial expressions. *Motivation & Emotion, 31*, 137–144.

Hori, E., Tazumi, T., Umeno, K., et al., (2005). Effects of facial expression on shared attention mechanisms. *Physiology and Behavior, 84*, 397–405.

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science, 14*, 640–643.

Ito, T. A. (*in press*). Perceiving social category information from faces: Using ERPs to study person perception. To appear in Todorov, A., Fiske, S. T., & Prentice, D. (Eds.), *Social Neuroscience: Toward Understanding the Underpinnings of the Social Mind*. Oxford University Press.

Kawashima, R., Sugiura, M., Kato, T., et al. (1999). The human amygdala plays an important role in gaze monitoring. A PET study. *Brain, 122*, 779–783.

Kleinke, C. L., & Pohlen, P. D. (1971). Affective and emotional responses as a function of other person's gaze and cooperativeness in a two-person game. *Journal of Personality & Social Psychology, 17*, 308–313.

Levin, D. T. (1996). Classifying faces by race: The structure of face categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1364–1382.

Livingston, R. W., & Brewer, M. B. (2002). What are we really priming? Cue-based versus category-based processing of facial stimuli. *Journal of Personality & Social Psychology, 82*, 5–18.

Low, M., & McSherry, J. (1968). Further observations of psychological factors involved in CNV genesis. *Electroencephalography and Clinical Neurophysiology, 25*, 203–207.

MacLin, O. H., & Malpass, R. S. (2001). Racial categorization of faces: The ambiguous-race face effect. *Psychology, Public Policy and Law, 7*, 98–118.

Maddox, K. (2004). Perspectives on racial phenotypicality bias. *Personality and Social Psychology Review, 8*, 383–401.

Marsh, A. A., Adams, R. B., Jr., & Kleck, R. E. (2005). Why do fear and anger look the way they do? Form and social function in facial expressions. *Personality and Social Psychological Bulletin, 31*, 73–86.

McConahay, J., Hardee, B., & Batts, V. (1981). Has racism declined in America? It depends on who is asking and what is asked. *Journal of Conflict Resolution, 25*, 563–579.

Monteith, M., Devine, P., & Zuwerink, J. (1993). Self-directed versus other-directed affect as a consequence of prejudice-related discrepancies. *Journal of Personality and Social Psychology, 64*, 198–210.

Phelps, E., O'Connor, K., Cunningham, W, et al. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience, 12*, 729–738.

Picton, T. & Hillyard, S. (1988). Endogenous components of the event-related brain potential. In T. Pitcotn (Ed.), *Human Event-related Potentials: EEG Handbook* (pp. 361–426). Amsterdam: Elsevier.

Plant, A., & Devine, P. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75*, 811–832.

Richeson, J. A., Todd, Trawalter, S., & Baird A. (2008). Eye-gaze direction modulates race-related amygdala activity. *Group Processes and Intergroup Relations, 11*, 233–246.

Richeson, J., Baird, A., Gordon, H., et al. (2003). An fMRI investigation of the impact of interracial contact on executive function. *Nature Neuroscience, 6*, 1323–1328.

Rohrbaugh, J., Syndulko, K., & Lindsley, D. (1976). Brain wave components of the contingent negative variation in humans. *Science, 191*, 1055–1057.

Sanders, D., Grandjean, D., Kaiser, S., Wehrle, T., & Scherer, K. R. (2007). Interaction effects of perceived gaze direction and dynamic facial expression: Evidence for appraisal theories of emotion. *European Journal of Cognitive Psychology, 19*, 470–480.

Sato, W., Yoshikawa, S., Kochiyama, T., & Matsumura, M. (2004). The amygdala processes the emotional significance of facial expressions: An fMRI investigation using the interaction between expression and face direction. *Neuroimage, 22*, 1006–1013.

Shriffin, R., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127–190.

Sporer, S. W. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law, 7*, 36–97.

Trawalter, S., Todd, A., Baird, A. A., & Richeson, J. A. (2008). Attending to threat: Race-based patterns of selective attention. *Journal of Experimental Social Psychology, 44*, 1322–1327.

Walter, W., Cooper, R., Aldridge, V., McCallum, W., & Winter, A. (1964). Contingent Negative Variation: an electric sign of sensori-motor association and expectancy in the human brain. *Nature, 203*, 380–384.

Zebrowitz, L. A. (1997). *Reading Faces : Window to the Soul?* Boulder, CO: Westview Press.

This page intentionally left blank

PART III

Regulation of Social Behavior

This page intentionally left blank

CHAPTER 10
Self-Regulation and Evaluative Processing

Dominic J. Packer, Amanda Kesek, & William A. Cunningham

The study of attitudes and the processes by which people determine whether objects in their environments are good or bad, trustworthy or untrustworthy, approachable or worth running away from has long been central to social psychology (Allport, 1935; Jung, 1921/1971). In their 1959 review, Katz and Stotland referred to the concept of attitude as "an orphan child, born in controversy and fostered in hostility," but noted that it was a construct that few "have been able to abandon" (p. 427). Indeed unable to abandon it, the subsequent 50 years of attitude research showed great conceptual and empirical advances. In this chapter, we focus on recent neuroscientific contributions to this literature, with particular attention to what this research reveals about the processes than underlie complex evaluations. We suggest that evaluations can be construed as falling on a continuum from those that are relatively simple (e.g., a strong negativity to spiders) to those that are relatively complex (e.g., a conflicting and ambivalent reaction to the death penalty). We present evidence that prefrontal brain regions support increasingly complex evaluations by directing and modulating the reprocessing of information to allow for the integration of relatively simple evaluations with additional information about context, social norms, as well as the goals of the perceiver.

Attitude researchers have sought to identify separate components of evaluation and to distinguish between different types of attitudes. Katz and Stotland (1959), for example, outlined the now classic view that cognitive, affective, and behavioral components can be combined in varying combinations to create different sorts of attitudes (*see also* Eagly & Chaiken, 1993). More recently, dual process models have proposed a distinction between automatic/implicit evaluations and controlled/explicit evaluations (*see* Chaiken & Trope, 1999). Automatic evaluations are rapidly generated and unintentional, reflecting the associational history of an attitude object; attitude objects that in the past have been paired with negative connotations or consequences generate negative automatic evaluations when they are re-encountered (e.g., Bargh, 1989; Devine, 1989; Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Greenwald & Banaji, 1995). Controlled evaluations, on the other hand, result from slower, more deliberative processing that takes situational goals, social norms, and novel information into account to construct contextually flexible evaluations (e.g., Fazio, 1990).

The automatic versus controlled distinction has generated a great deal of recent interest among social psychologists, particularly with respect to the study of prejudice. This is, at least in part, because measures of automatic evaluations are thought to provide an index of individuals' attitudes in the absence of social desirability concerns. For example, changes in North American social norms have made the explicit expression of prejudicial attitudes largely unacceptable, and blatant expressions of dislike for other social groups are now quite rare (at least in typical university samples). In contrast, implicit attitude measures, such as the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), Bona Fide Pipeline (Fazio, Jackson, Dunton, & Williams, 1995), and Affect Misattribution procedure (Payne, Cheng, Govorun, & Stewart, 2005), routinely find evidence of negativity toward stigmatized social groups (e.g., Nosek, Banaji, & Greenwald, 2002). The racial IAT, for example, measures the extent to which Black faces are preferentially associated with negative stimuli, relative to the extent to which White faces are preferentially associated with positive stimuli. Scores on this and other implicit measures have been shown to predict subtle and nonverbal types of bias (Dovidio, Kawakami, & Gaertner, 2002; Dovidio, Kawakami, Johnson, Johnson, & Howard, 1997), suggesting one reason why discrimination remains a problem despite dramatic changes in social norms with regard to the explicit expression of prejudice.

Automatic evaluations have obvious survival value. In particular, organisms with the ability to respond rapidly and preconsciously to stimuli that have been associated with negative outcomes in the past are much more likely to maintain their physical integrity in dangerous environments. Neuropsychological research suggests that perceptual information about a stimulus follows a subcortical route proceeding from the thalamus to the amygdala, which on the basis of prior associations generates a motivational inclination to either approach or avoid the stimulus. In response, projections from the hypothalamus prepare the body to make a rapid physiological response by altering sympathetic and parasympathetic nervous system activity (*see* Le Doux, 2000; Panksepp, 1998).

Despite the advantages of an automatic system, an organism capable only of automatic responses would be entirely dependent on its prior associational history and the immediate environment. As a consequence, such an organism would be relatively ill-equipped to deal with complex environments and unable to plan for the future (e.g., to delay gratification; Mischel, Ebbesen, & Ziess, 1972). In humans at least, conscious reflective processes, supported by a well-developed prefrontal cortex (PFC), allow for the construction of more complex evaluations that are responsive to long-term goals and social norms (Amodio & Frith, 2006; Crone & Van der Molen, 2004; McClure, Laibson, Loewenstein, & Cohen, 2004; Zelazo, 2004; Zelazo & Cunningham, 2007).

Attitudes, evaluations and iterative reprocessing

Although the distinction between automatic and controlled aspects of thought and feeling is a useful heuristic, we suggest that automatic and controlled evaluations are not strictly dichotomous (*see* Cunningham & Johnson, 2007). Instead, as people have greater opportunity (e.g., time) to process a stimulus, reflective processes come online, which allows for more complex forms of evaluation. Rather than generating an entirely new evaluation (e.g., a separate "explicit" attitude), we suggest that these reflective processes interact with already active automatic responses. In their recent Iterative Reprocessing (IR) model, Cunningham and Zelazo (2007) postulate that reflective processes reseed the processing stream, highlighting some aspects of information, backgrounding others, and/or retrieving additional information. Newly active patterns of representation are then fed into the same automatic system that processed the information initially, generating a more nuanced evaluation. As such, evaluation is not the result of a single process occurring within a fixed time window. Some judgments may be reached rapidly and remain stable across the lifespan, whereas others may be continually altered and updated as new information and situations are encountered (*see also* Cunningham, Packer, Van Bavel, & Kesek, 2009).

Although the terms *attitude* and *evaluation* are generally employed synonymously, we have found it useful to draw a distinction between them (Cunningham et al., 2009; Cunningham & Zelazo, 2007). We use the term *attitude* to refer to all pre-existing valenced information that a person has about a stimulus, either from

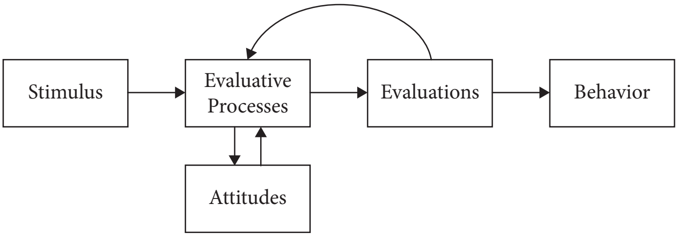


Fig. 10–1 **Simple Schematic of the Iterative Reprocessing Model.**

prior learning or because of innate preferences. In contrast, we use the term *evaluation* to refer to the current state of the evaluative system, which is influenced (although not exclusively) by activated aspects of the relevant attitude. Evaluations reflect the currently determined motivational significance and reward/punishment value of a stimulus. Importantly, the current evaluation of a stimulus is not reliant solely on stored attitude representations: *evaluative processes* construct evaluations by drawing upon pre-existing attitudes and integrating them with new information about the stimulus, along with contextual information and current goals (*see* Cunningham & Zelazo, 2007).

According to the IR model, and as illustrated in Figure 10–1, these evaluative processes are iterative, and information about a stimulus is continually fed back through the system (Cunningham et al., 2009; Cunningham & Zelazo, 2007). With each iteration, the current evaluation can be combined with additional contextual and motivational information to create a new, updated evaluation. As time passes, and more reflective processes (mediated by the PFC) come online, there is greater opportunity for the elaboration or modulation of earlier evaluations. Thus, the distinction between relatively automatic and controlled evaluations can be conceptualized as reflecting different points on an iterative continuum (Cunningham & Johnson, 2007). Automatic evaluations arise after relatively few iterations and are thus, on average, more dependent on highly accessible attitude representations (Fazio et al., 1986). Controlled evaluations arise after additional iterations and reflect the integration of attitudes with relevant contextual information and goal-states.

Automatic evaluative processing

The neural structures subserving relatively automatic versus relatively reflective processing can be differentiated by comparing brain activity between tasks in which people attend versus do not attend to their evaluations of stimuli. In attended, reflective tasks, participants are asked to think about and report their evaluations (e.g., the valence of faces, names, or concepts). In unattended, nonreflective tasks, participants may be asked to report on a non-evaluative aspect of the same stimuli (e.g., the gender of faces; Iidaka et al., 2001; *see also* Anderson, Christoff, Panitz, De Rosa, & Gabrieli, 2003a), or stimuli may be presented so rapidly as to prevent conscious detection (Cunningham et al., 2004a; Whalen et al., 1998). Because these nonreflective tasks reduce or eliminate conscious evaluative processing, patterns of neural activation that differ depending on the evaluative properties of stimuli (e.g., that differ between positive and negative stimuli) can be assumed to manifest automatic processes.

Automatic evaluative and emotional processing—particularly of negativity—has consistently been linked to activity in the amygdala (Le Doux, 1996; Whalen, 1998). This almond-shaped structure, buried deep in the medial temporal lobe, supports fear-conditioning (e.g., Armony & Dolan, 2002; Davis, 1995) and the perception of fear in others (e.g., Adolphs et al., 1999; Hadjikhani & de Gelder, 2003). Across modalities, the amygdala generally responds more strongly to negative than positive stimuli (Anderson et al., 2003a; Isenberg et al., 1999; Morris et al., 1996; Small et al., 2003), even when stimuli are presented subliminally (Cunningham et al., 2004a; Morris, Ohman,

& Dolan, 1998; Whalen et al., 1998). Recent research suggests, however, that the amygdala may be responsive to the intensity or arousal value of stimuli, rather than their valence (Anderson et al., 2003b; Small et al., 2003). In an fMRI study, Cunningham, Raye, and Johnson (2004b) asked participants to rate concept words (e.g., murder, happiness) in either an evaluative (i.e., good vs. bad) or non-evaluative (i.e., concrete vs. abstract) fashion. Across both conditions, and controlling for stimulus valence, they found that activity in the amygdala was predicted by participants' ratings of stimulus intensity. Valence, on the other hand, was associated with activity in the right anterior insula, which responded more strongly to negative than positive stimuli. Given large-scale projections from the insula to the hypothalamus, and the role that both of these structures play in the modulation of sympathetic and parasympathetic nervous system activity, this finding implies that valence may be represented in the brain as a physiological orientation toward a stimulus (i.e., a behavioral tendency/readiness to approach or avoid; *see* Cunningham & Zelazo, 2007; Damasio, 1994, 1996; Critchley, Weins, Rotshtein, Ohmen, & Dolan, 2004).

Conscious expression and elaboration

In the remainder of this chapter, we review findings regarding the role of the PFC in the construction of complex evaluations and in the modulation of early, relatively automatic evaluations. We suggest that the PFC is involved in the conscious expression of evaluations, as well as their elaboration. By drawing upon additional information, elaborative processing in the PFC may change the nature of an evaluation (e.g., its valence) or, alternately, may leave the value of the evaluation relatively unchanged but embed it in a more complex cognitive structure (e.g., employ a stereotype to justify a prejudice). At other times, the PFC may actively modulate evaluative processing, either suppressing or enhancing automatic responses from subcortical regions in further iterations of the evaluative cycle.

Consistent with the prediction that prefrontal regions are involved in the conscious consideration and expression of evaluations, comparison of evaluative and non-evaluative conditions in the concept–word study described above revealed heightened activation in prefrontal regions when participants made explicitly evaluative judgments (Cunningham et al., 2004b). Specifically, the evaluative task recruited greater activity in the anterior cingulate cortex (ACC), right anterior PFC, and bilateral regions of orbital frontal cortex (OFC). Similarly, a related study, which asked participants to rate the names of famous people (e.g., Adolf Hitler, Bill Cosby) either in terms of their valence (good vs. bad) or their historical status (past vs. present) found greater activity in medial and ventrolateral PFC for evaluative than non-evaluative judgments (Cunningham, Johnson, Gatenby, Gore, & Banaji., 2003).

We have suggested that prefrontal regions are likely to be particularly important for the construction of complex evaluations and that the PFC supports the integration of initial evaluations with additional attitudinal as well as contextual information about a stimulus. In line with this contention, Cunningham et al. (2003) observed greater activity in the ventrolateral PFC when participants evaluated famous people toward whom they were ambivalent (i.e., who they evaluated positively and negatively at the same time). Similarly, in Cunningham et al. (2004b), participants' ratings of how much they typically try to control their initial reactions to concepts (which correlated strongly with ambivalence) predicted prefrontal activity in response to those concepts. Importantly, in both studies, the correlations between stimulus ambivalence/control and prefrontal activity were greater in the evaluative than non-evaluative conditions, indicating that conscious, reflective processing may be required for the representation—and possibly the resolution—of complex evaluations.

Interactive processes

It is important to note that activity in the amygdala and insula was evident in both the non-evaluative and evaluative tasks in these studies. The fact that these regions were engaged when participants made non-evaluative ratings suggests that they are involved in relatively automatic processing and that detection of stimulus valence and stimulus intensity does not require conscious attention. However, the fact that these regions were also active when participants made evaluative ratings further suggests that they were not supplanted or replaced by more reflective processes. Indeed, the amygdala showed greater activation in the evaluative than non-evaluative tasks (*see* Cunningham et al., 2004b), implying that amgydalic activity reflects both initial associatively driven automatic evaluations, as well as subsequent reflective evaluative processing. These findings point to a hierarchical evaluative system: as higher-order processes (supported by cortical regions) come online, information is continually fed back through lower-order processes (supported by subcortical regions) to generate updated evaluations (Cunningham, Espinet, DeYoung, & Zelazo, 2005; Cunningham & Zelazo, 2007). As such, even as conscious deliberation starts to exert an influence on evaluative processing, evaluative states themselves are likely to go on being represented in subcortical brain structures.

In these studies, participants were directly asked to reflect on and report their evaluations (in the evaluative conditions). Although explicitly evaluative situations are fairly common in day-to-day life (e.g., choosing what to eat in a restaurant, selecting a job candidate), the issue of what triggers a shift from relatively automatic to relatively reflective processing in the absence of an explicit evaluative goal remains a question. What, in other words, besides instructions to deliberatively evaluate a stimulus causes continued iterations of evaluative processing? Stimulus ambivalence may be one such trigger; when automatic associative processes do not give rise to a simple binary (good or bad) evaluation, an individual may engage in further reflective processing to resolve the inconsistency. More reflective processing may also be triggered by incongruities between the current evaluation of a stimulus and feedback from the environment. For example, the need for further deliberation may be signaled if approaching a positively evaluated stimulus does not yield the rewards expected to accompany it (or, even worse, if it is punished). Research suggests that the OFC is involved in the comparison of expectations (i.e., current evaluations) with rewards (e.g., Blair, 2004; Beer, Heery, Keltner, Scabini, & Knight, 2003; Rolls, 2000; Rolls, Hornak, Wade, & McGrath, 1994). Detection of a disparity between expectations and outcomes, or the presence of uncertainty (e.g., ambivalence), triggers activity in the ACC, a region associated with conflict monitoring (e.g., Carter et al., 1998; Cohen, Botvinick, & Carter, 2000). The ACC, in turn, signals that the organism's current evaluative state requires some adjustment, triggering prefrontal regions to engage in evaluative reprocessing (Bunge & Zelazo, 2006; Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004). Continued evaluative processing updates the current evaluation with additional information recruited from prestored attitudes, as well as the environment, to achieve a better match to reality or a more valid evaluation (Cunningham & Zelazo, 2007).

In addition to elaborating and updating evaluations with more information, prefrontal activity may also serve to embed evaluations within more cognitively complex structures or schemas. An individual's evaluation of a single stimulus (e.g., a sports utility vehicle) is often subsumed within a larger value system or ideology (e.g., environmentalism). At times, reflective processing may alter a current evaluation to make it consistent with a set of personal or societal values (*see* discussion of modulation below). At other times, a current evaluation may be left unchanged, but reflective processing may be employed to justify and account for it. The intergroup relations literature suggests that stereotypes about social groups are often used as justifications for possessing negative attitudes toward them (e.g., Crandall & Eshleman, 2003; Jost & Banaji, 1994). Stereotypes are, in part, causal schemas that attribute certain outcomes (e.g., low societal status) to the dispositional characteristics of a group (e.g., incompetence), which carry with them evaluative connotations (Fiske, Cuddy, Glick, & Xu, 2002). Perceivers will evaluate a member of a stigmatized group negatively if they believe that the causally justifying stereotype can/should be applied to that person (e.g., Kunda & Spencer, 2003; Sinclair & Kunda, 1999).

It is likely that prefrontal regions are also involved in this type of elaborative processing, in which evaluations are integrated with pre-existing cognitive structures. Although there is little direct evidence in the evaluative domain, this contention is consistent with a recent study by Satpute et al. (2006), which examined the neural correlates of causal reasoning. In one condition, participants rated whether two concepts were semantically associated (e.g., ring–emerald); in another condition, they rated whether two concepts were causally related (i.e., whether one causes the other: moon–tide). Determining the nature of the causal relationship between concepts was associated with heightened activity in the dorsolateral PFC (an area associated with reasoning and working memory tasks) as well as the precuneus (a posterior midline region linked to the integration of episodic memories). To the extent that stereotypes and other cognitive schemas serve to justify evaluations by embedding them within a causal structure, we would expect to see similar patterns of activation when individuals elaborate on their evaluations in this way (Quadflieg et al., 2009).

In our model, reflective processes are thought to drive and direct automatic evaluative processes—that is, it is not that reflective processes necessarily generate a distinct evaluative state themselves, but rather they foreground/background information for additional iterations of evaluative processing within the automatic system. As such, with reflective processing, evaluations (processed automatically) can be shaped by current motivations and goals and modulated by reflective thought. Support for this idea comes from studies of emotional regulation that demonstrate a common set of prefrontal activations accompanying both the deliberate down-regulation (make yourself feel less emotional) and up-regulation (make yourself feel more emotional) of amygdala activations (Ochsner, Bunge, Gross, & Gabrieli, 2002; Ochsner et al., 2004).

One mechanism by which reflection can alter automatic processing is by directing attention to motivationally salient aspects of stimuli. This system can detect potentially relevant features and redirect attention, such that significant stimuli receive enhanced processing. Consistent with this, recent evidence suggests that chronic differences in orientation toward valenced information (e.g., positivity vs. negativity bias), as well as situational variables, may direct attention and thereby influence the perception of emotional intensity. In one study, participants were presented with positively and negatively valenced stimuli during fMRI scanning (Cunningham, Raye, & Johnson, 2005). After scanning, participants completed an individual differences measure of their prevention and promotion focus orientation (e.g., *see* Higgins, 1997). Individuals scoring high on promotion focus tend to be attentive to and motivated by the achievement of gains, whereas individuals scoring high on prevention focus tend to be oriented toward avoidance of losses. Results indicated that more promotion-focused participants had greater activation in the amygdala, anterior cingulate gyrus, and extrastriate cortex for positive stimuli. Conversely, more prevention -focused participants had greater activation in same these regions for negative stimuli. Thus, amygdala and attentional brain regions were not universally tuned toward a particular valence but, rather, toward stimuli that were motivationally important for the individual (as defined by chronic goal states).

More direct evidence for the motivated direction of automatic processing by reflective processes comes from work by Cunningham, Van Bavel, and Johnsen (2008). In this study, participants were presented with famous names and asked to focus on either the positive or negative aspects of the person (e.g., ignoring anything bad, how good is this person?). Activity in bilateral amygdala and insula was found to vary as a function of evaluative fit—that is, when focusing on negativity, greater amygdala and insula activity was found to bad rather than good. The opposite pattern was found for the positive focus condition, such that greater activity was observed in these regions to good rather than bad names. Taken together, these studies suggest that reflective thought engaged by task demands and motivational concerns can direct and modulate the processing of valenced information to generate situationally appropriate responses.

Similar effects have been found in studies of prejudice, where participants are typically motivated to control or inhibit negative evaluations that they may have about certain social groups. Specifically, greater amygdala activation to Black, rather than White, faces was found when participants used social (racial) categories to make evaluative judgments about them, but greater activation to White, rather than Black, faces when people tried to treat the faces as individuals (Wheeler & Fiske, 2005). The motivation to individuate modulated the automatic evaluative signal. More directly, in another study, when faces were presented subliminally, 12 of 13 White participants had greater amygdalic activation to Black faces compared to White (Cunningham et al., 2004a; *see also* Hart et al., 2001; Leiberman, Hariri, Jarcho, Eisenberger, & Bookheimer, 2005; Phelpset al., 2001). Yet, when stimuli were presented supraliminally and participants had opportunity to regulate initial reactions, there was an equivalent amygdalic response to Black and White faces. Consistent with modulation of evaluative processing by prefrontal regions, this decreased activation in amygdala to Black relative to White faces was accompanied by increased activation in areas of the ACC and lateral PFC (dorsolateral PFC and rostrolateral PFC).

Like the findings for race, greater activity in amygdala and insula has been found for images of members of other stigmatized groups (e.g., obese individuals and transexuals; Krendl, Macrae, Kelley, Fugelsang, & Heatherton, 2006). However, unlike in studies involving racial stimuli, in which personal goals and/or social norms may encourage participants to inhibit negative responses to Black faces, negative responses to these sorts of stigmatized others may be considered more normatively acceptable. Perhaps for this reason, in this study, heightened activity in amygdala and insula was accompanied by greater activity in the ACC and lateral PFC in response to the stigmatized images. Consistent with the idea that individuals foreground automatically activated negative information when it is congruent with motivational concerns, lateral PFC-mediated foregrounding may have led to the amplification of negative emotional evaluations.

Cognitive development and evaluative processing

Where on the continuum of reflective processing one is operating at any particular moment depends on a confluence of factors (e.g., time, opportunity, and motivation; Fazio 1990). An additional factor likely to influence reflective processing is PFC functionality, whether defined in terms of brain damage (e.g., Bechara, 2004) or in terms of the neural development that continues through adolescence (e.g., Zelazo, 2004). As discussed, self-regulation depends critically on the PFC, and development in these regions is thought to underlie the emergence of various cognitive processes. With cognitive and neural development, children become more sophisticated in their ability to consciously reflect on thoughts, actions, people, and situations. Although even very young children are capable of quickly evaluating whether a stimulus is good or bad, the capacity for conscious reflection allows children to make increasingly complex evaluations.

PFC development underlies children's transformation from stimulus-bound infants to goal-oriented individuals capable of regulation and reflection. For example, one mark of PFC maturation is a shift in the ratio of gray to white matter. Gray matter volume reaches adult levels earlier in the OFC than in more lateral areas of the PFC, which achieve maturity only at the end of adolescence (Giedd et al., 1999). The protracted developmental course of lateral areas of PFC has also been documented using measures of cortical thickness (Nagy, Westerberg, & Klingberg, 2004). Importantly, the regions of the brain involved in relatively automatic, affective responses to stimuli develop somewhat earlier than regions associated with more controlled processing (Zelazo & Cunningham, 2007). With PFC development, children are able to override the relatively automatic, affective responses mediated by the amygdala and OFC. Rather than simply responding to the salient aspects of a stimulus, conscious reflection allows children to integrate positive and negative information and generate a more considered evaluation of a stimulus.

Evidence for the development of more controlled processing in the face of stimuli that elicit automatic, appetitive responses comes from studies assessing the ability of young children to delay gratification. For example, Prencipe and Zelazo (2005) found that when given the option between a small reward now or a larger reward later, there were increases in the tendency to delay gratification among children between the ages of 3 and 5 years. The development of the ability to control affective reactions to stimuli has also been studied using a modified version of the Iowa Gambling Task (Bechara, Damasio, Damasio, & Anderson, 1994). Kerr and Zelazo (2004) administered a version of this task that used only two decks of cards (one advantageous and one disadvantageous) and presented information about rewards and losses in the form of happy faces. Over the course of the task, 4- and 5-year-olds developed a preference for the advantageous decks, whereas 3-year-olds did not. The ability to make advantageous decisions based on a sophisticated evaluations improves during the preschool years, and continues to develop over the course of childhood (Crone & van der Molen, 2004).

This evidence suggests that with development, children are able to make increasingly reflective evaluations and are able to integrate information about the current situation with more accurate predictions of long-term consequences. These more sophisticated evaluations probably result, in part, from more complex reflective processes, instantiated in higher-order prefrontal regions (e.g., *see* McClure et al., 2004). With each iteration of the evaluative system, these more complex processes have the ability to shape evaluative responses to match goals and situational constraints. As new information becomes available, it can only be used to create a more complex representation of a stimulus to the extent that it is incorporated into one's evaluation. With neural maturation, which includes increases in processing speed, children should be able to iterate more quickly and efficiently, allowing for more complex evaluations (*see* Cunningham & Zelazo, 2007). The developmental course of evaluative processing, including its neural and cognitive underpinnings, is an important area for future research.

Conclusion

Humans possess a sophisticated evaluative system, capable of split-second, preconscious judgments, as well as drawn-out, complex, and deliberative decisions. Neuro-imaging research is beginning to unpack the neural correlates of the components of evaluation and, in doing so, contributes to our understanding of the evaluative system. In this chapter, we reviewed our working model of attitudes, in which current evaluations of a stimulus are continually updated and integrated with additional attitudinal, situational, and motivational information to generate increasingly complex evaluations (*see* Cunningham et al., 2009; Cunningham & Zelazo, 2007). Data suggests that this evaluative cycle, supported by affective regions including the amygdala and insula, is sustained and modulated by activity in prefrontal areas, allowing for more reflective, context/goal-appropriate evaluations.

One important implication of our model is that a person's evaluative state varies from moment to moment. The social psychology literature has recently grappled with apparent dissociations between so-called "implicit" and "explicit" measures (e.g., Cunningham, Preacher, & Banaji, 2001; Nosek, 2005). For example, people may report pro-Black attitudes on self-report measures but be shown to have a pro-White bias on a response latency measure of attitudes (Cunningham et al., 2004a; Nosek et al., 2002). Do people have different attitude representations that are activated through different processing routes? Is one the real attitude and the other an artifact? Our model suggests that these differences may reflect evaluations captured at different points in time or at alternate stages of evaluative processing. As such, it may not make a great deal of sense to inquire after an individual's "real" or "true" attitude. To the extent that evaluations shift over time, as attitudinal, motivational, and contextual information is reprocessed and re-integrated, the same individual can be said to possess multiple "real" evaluations of the same stimulus. If we accept the idea that individuals may not have one "true" evaluation, an urgent issue for future research would seem to be understanding how evaluations arising at different points along the iterative continuum relate to different sorts of behavior (e.g., Dovidio et al., 1997, 2002).

Contrary to the hopes of some and the fears of others, the studies described in this chapter suggest that for the present, brain imaging does not provide a direct means of determining the *content* of a person's evaluation. Rather, patterns of neural activity are only interpretable when used in concert with more traditional behavioral measures and when context and motivational states are taken into account. For example, we have seen heightened activity in the amygdala and insula to both positive and negative stimuli, depending on chronic motivational states (Cunningham et al., 2005) as well as evaluative focus (Cunningham et al., 2008). Similarly, we have observed prefrontal activity in response to social targets toward whom participants were motivated to suppress initial negative reactions (Cunningham et al., 2004), as well as in response to targets toward which participants were comfortable expressing negativity (Krendl et al., 2006). Given the current state of our knowledge about neural function, activity in a particular brain region cannot by itself be used as an index of how someone evaluates a particular stimulus.

Finally, our model raises a somewhat philosophical point about the nature of automatic versus reflective evaluations. There is a temptation to view automatic evaluations that arise out of philogenically older limbic areas as more animal-like and less moral than evaluations guided by conscious, reflective thought. Influenced by Freud (1930/2004), lay-theorists tend to believe that human's increased prefrontal capacity exerts a "civilizing" cognitive influence on more primitive emotional responses, holding aggressive and other urges in check. However, the IR model suggests that making a firm distinction between emotional and cognitive evaluations is not possible; rather, processes driven by subcortical and prefrontal regions are highly interactive. Although prefrontal activity may modulate early automatic responses, subcortical regions remain involved in the reprocessing and reconstruction of evaluative states (*see* Cunningham & Zelazo, 2007). Further, as we have seen, the manner in which automatic evaluations are modulated depends on the goals of the perceiver, as well as the social context. Sometimes this may result in a dampening of situationally inappropriate negativity toward social targets; at other times, however, negative responses may be left unchecked or even exacerbated by reflective processing. Indeed, we have suggested that reflective processing may be used, in some circumstances, to justify negative evaluations by embedding them in complex cognitive structures (*see* Crandall & Eshleman, 2003). As such, relatively automatic processes should not be viewed as necessarily resulting in less moral, appropriate, or adaptive responses than relatively controlled processes (*see also* Damasio, 1994; Dijksterhuis, Bos, Nordgren, & van Baaren, 2006; Green & Haidt, 2002).

Given the complexity of the evaluative system, it is likely that our model will require further revision and specification. However, we are excited about the recent contributions neuropsychological research has made to the attitude literature. We are confident that neuro-imaging techniques, in conjunction with traditional social psychological methods, will continue to give rise to an ever-clearer understanding of attitudes and evaluation.

References

Adolphs, R., Tranel, D., Hamann, S., et al. (1999). Recognition of facial emotion in nine individuals with bilateral amygdala damage. *Neuropsychologia, 37*, 1111–1117.

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *Handbook of Social Psychology* (pp. 798–844). Worchester, MA: Clark University Press.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*, 268–277.

Anderson, A. K., Christoff, K., Panitz, D., DeRosa, E., & Gabrieli, J. D. E. (2003a). Neural correlates of the automatic processing of threat facial signals. *Journal of Neuroscience, 23*, 5627–5633.

Anderson, A. K., Christoff, K., Stappen, I., et al. (2003b). Dissociated neural representations of intensity and valence in human olfaction. *Nature Neuroscience, 6*, 196–202.

Armony, J. L., & Dolan, R. J. (2002). Modulation of spatial attention by fear-conditioned stimuli: An event-related fMRI study. *Neuropsychologia, 40*, 807–826.

Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended Thought* (pp. 3–51). New York: Guilford.

Beer, J. S., Heerey, E. A., Keltner, D., Scabini, D., & Knight, R. T. (2003). The regulatory function of self-conscious emotion: Insights from patients with orbitofrontal damage. *Journal of Personality and Social Psychology, 85*, 594–604.

Bechara, A. (2004). The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage. *Brain and Cognition, 55*, 30–40.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50*, 7–15.

Blair, R. J. R. (2004). The roles of orbital frontal cortex in the modulation of antisocial behavior. *Brain Cognition, 55*, 198–208.

Bunge, S., & Zelazo, P. D. (2006). A brain-based account of the development of rule use in childhood. *Current Directions in Psychological Science, 15*, 118–121.

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M. M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science, 280*, 747–749.

Chaiken, S., & Trope, Y. (1999). *Dual-Process Theories in Social Psychology*. New York: Guilford Press.

Cohen, J. D., Botvinick, M., & Carter, C. S. (2000). Anterior cingulate and prefrontal cortex: Who's in control? *Nature Neuroscience, 3*, 421–423.

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin, 129*, 414–446.

Critchley, H. D., Wiens, S., Rotshtein, P., Ohman, A., & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience, 7*, 189–195.

Crone, E. A., & van der Molen, M. W. (2004). Developmental changes in real life decision making: Performance on the gambling task previously shown to depend on the vetromedial prefrontal cortex. *Developmental Neuropsychology, 25*, 251–279.

Cunningham, W. A., Espinet, S. D., DeYoung, C., & Zelazo, P. D. (2005). Attitudes to the right—and left: Frontal ERP asymmetries associated with stimulus valence and processing goals. *NeuroImage, 28*, 827–834.

Cunningham, W. A., & Johnson, M. K. (2007). Attitudes and evaluation: Toward a component process framework. In E. Harmon-Jones, & P. Winkielman (Eds.), *Fundamentals of Social Neuroscience* (pp. 227–245). New York: Guilford Press.

Cunningham, W. A., Johnson, M. K., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2003). Neural components of social evaluation. *Journal of Personality and Social Psychology, 85*, 639–649.

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004a). Seperable neural components in the processing of black and white faces. *Psychological Science, 15*, 806–813.

Cunningham, W. A., Packer, D. J., Kesek, A., & Van Bavel, J. J. (2009). Implicit measurement of attitudes: A physiological approach. In R. E. Petty, R. H. Fazio, & P. Brinol (Eds.), *Insights from the New Implicit Measures* (pp. 485–512). New York, NY: Psychology Press.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability and convergent validity. *Psychological Science, 12*, 163–170.

Cunningham, W. A., Raye, C. L., & Johnson, M. K. (2004b). Implicit and explicit evaluation: fMRI correlates of valence, emotional intensity, and control in the processing of attitudes. *Journal of Cognitive Neuroscience, 16*, 1717–1729.

Cunningham, W. A., Raye, C. L., & Johnson, M. K. (2005). Neural correlates of evaluation associated with promotion and prevention regulatory focus. *Cognitive, Affective, & Behavioral Neuroscience, 5*, 202–211.

Cunningham, W. A., Van Bavel, J. J., & Johnsen, I. (2008). Affective flexibility: Evaluative processing goals shape amygdala activity. *Psychological Science, 19*, 152–160.

Cunningham, W. A., & Zelazo, P. D. (2007). Attitudes and evaluations: A social cognitive neuroscience perspective. *Trends in Cognitive Sciences, 11*, 97–104.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Putnam.

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transcripts of the Royal Society of London, B, Biological Sciences, 352*, 1413–1420.

Davis, M. (1997). Neurobiology of fear responses: The role of the amygdala. *Journal of Neuropsychiatry and Clinical Neurololgy, 9*, 382–402.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5–18.

Dijksterhuis, A., Bos, M. W., Nordgren, L. F., & van Baaren, R. B. (2006). On making the right choice: The deliberation-without-attention effect. *Science, 311*, 1005–1007.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*, 62–68.

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of Experimental Social Psychology, 33*, 510–540.

Eagly, A. H., & Chaiken, S. (1993). *The Psychology of Attitudes*. Forth Worth, TX: Harcourt Brace Jovanovich.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 23, pp. 75–109). New York: Academic Press.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69*, 1013–1027.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*, 229–238.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*, 878–902.

Freud, S. (1930/2004*). Civilization and Its Discontents*. Translated by D. McLintock. Toronto: Penguin Books.

Giedd, J. N., Blumenthal, J., Jeffries, N. O., et al. (1999). Brain development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience, 2*, 861–863.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Science, 6*, 517–523.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*, 4–27.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

Hadjikhani, N., & de Gelder, B. (2003). Seeing fearful body expressions activates the fusiform cortex and amygdala. *Current Biology, 13*, 2201–2205.

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.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist, 52*, 1280–1300.

Iidaka, T., Omori, M., Murata, T., et al. (2001). Neural interaction of the amygdala with the prefrontal and temporal cortices in the processing of facial expression as revealed by fMRI. *Journal of Cognitive Neuroscience, 13*, 1035–1047.

Isenberg, N., Silbersweig, D., Engelien, A., et al. (1999). Linguistic threat activates the human amygdala. *Proceedings of the National Academy of Sciences, USA, 96*, 10,456–10,459.

Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology, 33*, 1–27.

Jung, C. G. (1921/1971). *Psychological Types (Volume 6)*. Princeton, NJ: Princeton University Press.

Katz, D., & Stotland, E. (1959). A preliminary statement to a theory of attitude structure and change. In S. Koch (Ed.), *Psychology: A Study of a Science* (pp. 423–475). Toronto: McGraw-Hill.

Kerr, A., & Zelazo, P. D. (2004). Development of "hot" executive function: The Children's Gambling Task. *Brain and Cognition, 55*, 148–157.

Krendl, A. C., Macrae, C. N., Kelley, W. M., Fugelsang, J. A., & Heatherton, T. F. (2006). The good, the bad, and the ugly. An fMRI investigation of the functional anatomic correlates of stigma. *Social Neuroscience, 1*, 5–15.

Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin, 129*, 522–544.

LeDoux, J. E. (1996). *The Emotional Brain*. New York: Simon & Schuster.

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

Lieberman, M. D., Hariri, A., Jarcho, J. M., Eisenberger, N. I. & Bookheimer, S. Y. (2005). An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. *Nature Neuroscience, 8*, 720–722.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science, 306*, 503–507.

Mischel, W., Ebbesen, E. B., & Zeiss, A. R. (1972). Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology, 21*, 204–218.

Morris, J. S., Frith, C. D., Perrett, D. I., et al. (1996). A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature, 383*, 812–815.

Morris, J. S., Öhman, A., & Dolan, R. J. (1998). Conscious and unconscious emotional learning in the human amygdala. *Nature, 393*, 417–418.

Nagy, Z., Westerberg, H., & Klingberg, T. (2004). Maturation of white matter is associated with the development of cognitive functions during childhood. *Journal of Cognitive Neuroscience, 16*, 1227–1233.

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology, General*, 134, 565–584.

Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics: Theory, Research and Practice, 6*, 101–115.

Ochsner, K. N., Bunge, S. A., Gross, J. J., & Gabrieli, J. D. E. (2002). Rethinking feelings: An fMRI study of the cognitive regulation of emotion. *Journal of Cognitive Neuroscience, 14*, 1215–1229.

Ochsner, K. N., Ray, R. D., Cooper, J. C., et al. (2004). For better or for worse: neural systems supporting the cognitive down- and up-regulation of negative emotion. *NeuroImage, 23*, 483–499.

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. New York: Oxford University Press.

Payne, K. B., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*, 277–293.

Phelps, E. A., O'Connor, K. J., Cunningham, W. A., et al. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience, 12*, 729–738.

Prencipe, A., & Zelazo, P. D. (2005). Development of affective decision-making for self and other: Evidence for the integration of first- and third-person perspectives. *Psychological Science, 16*, 501–505.

Quadflieg, S., Turk, D. J., Waiter, G. D., et al. (2009). Exploring the neural correlates of social stereotyping. *Journal of Cognitive Neuroscience, 21*, 1560–1570.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science, 306*, 443–447.

Rolls, E. T. (2000). The orbitofrontal cortex and reward. *Cerebral Cortex, 10*, 284–294.

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 & Psychiatry, 57*, 1518–1524.

Satpute, A. B., Fenker, D. B., Waldmann, M. R., Tabibnia, G., Holyoak, K. J., & Lieberman, M. D. (2005). An fMRI study of causal judgments. *European Journal of Neuroscience, 22*, 1233–1238.

Sinclair, L., & Kunda, Z. (1999). Reactions to a black professional: Motivated inhibition and activation of conflicting stereotypes. *Journal of Personality and Social Psychology, 77*, 885–904.

Small, D. M., Gregory, M. D., Mak, Y. E., Gitelman, D., Mesulam, M. M., & Parrish, T. (2003). Dissociation of neural representation of intensity and affective valuation in human gestation. *Neuron, 39*, 701–711.

Whalen, P. J. (1998). Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala. *Current Directions in Psychological Science, 7*, 177–188.

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. *The Journal of Neuroscience, 18*, 411–418.

Wheeler, M. E., & Fiske, S. T. (2005). Controlling racial prejudice: Social-cognitive goals affect amygdala and stereotype activation. *Psychological Science, 16*, 56–63.

Zelazo, P. D. (2004). The development of conscious control in childhood. *Trends in Cognitive Science, 8*, 12–17.

Zelazo, P. D., & Cunningham, W. (2007). Executive function: Mechanisms underlying emotion regulation. In J. Gross (Ed.), *Handbook of Emotion Regulation* (pp. 135–158). New York: Guilford.

CHAPTER 11
The Neural Basis of Emotional Decision-Making

Jennifer S. Beer & Jamil P. Bhanji

In the proverbial tale of the six blind men and the elephant, all six men are touching the same elephant but describe it in various forms, depending on whether they are touching the trunk, the torso, and so forth. A true understanding of the elephant could only come from collaboration among the men. This same principle has played an important role in understanding the relation between emotion and reason. In contrast to the traditional view that emotion opposes reason, researchers from psychology and neuroscience both suggest a more favorable role of emotion in decision making. Psychologists have begun to take seriously the idea that emotions may have evolved for adaptive reasons, including shaping cognitive processing such as decision making (e.g., Ekman, 1992; Levenson, 1999). Positive and negative emotion states may be adaptive as they prepare individuals for important information or actions. Neuroscientists have also begun to view emotion as an adaptive force in decision making. The turning point came when researchers discovered that damage to the orbitofrontal cortex, a region of the brain considered important for emotions, also impaired decision making (e.g., Bechara, Damasio, Tranel, & Damasio, 1997).

However, the role of orbitofrontal cortex in emotional decision making has recently been called into question. Recent research that draws on both social psychological and neuroscience approaches (i.e., social neuroscience) suggests that orbitofrontal function may be better characterized as supporting self-insight. From this perspective, the orbitofrontal cortex may have only a distal influence on emotional decision making. Self-insight processes affect which emotions are generated, and these emotions affect subsequent decision making. This research does not refute the theorized adaptive role of emotion in decision making. Instead, this research highlights the need to expand the focus on neural investigations of emotional decision making into systems outside the orbitofrontal cortex. A small number of studies suggest other brain regions that may support the adaptive role of emotion in decision making, but strong conclusions are not currently possible because *(1)* emotional decision making is not the main focus of some studies so it can only be inferred, or *(2)* a lack of behavioral effects make it difficult to interpret the psychological meaning of neural activity. The chapter concludes by proposing future directions for "social neuroscience" investigations of emotional decision making.

The neural basis of emotionally influenced decision making: a focus on the orbitofrontal cortex

The turning point in neuroscientific views of emotion begin with a case study of a patient with orbitofrontal damage. This patient could generate solutions to problems but not prioritize various solutions on the basis of their viability (Saver & Damasio, 1991). For example, he could name a number of ways to address social dilemmas (e.g., two roommates who can not agree on which television program to watch), but he could not distinguish which solutions were most likely to be effective in resolving the dilemma. The orbitofrontal cortex was associated with emotional functions so the discovery of decision-making impairments intrigued neuroscientists. Was it possible that damage to the orbitofrontal cortex impaired decision making because emotion was actually needed to optimize decision making? A number of scientists investigated this question and have described different mechanisms through which the orbitofrontal cortex mediates emotional decision making.

The Somatic Marker Hypothesis

From a Somatic Marker Hypothesis perspective, poor decision making occurs when somatic information (e.g., emotion) is not available to guide decision making (e.g., Bechara, Damasio, & Damasio, 2000; Bechara, Damasio, Tranel, & Damasio, 1997). Orbitofrontal structures are theorized to support learning of associations between complex situations and the somatic changes (i.e., emotional state) usually associated with a particular situation. A distributed network of activity is modulated by the orbitofrontal cortex, and this activity is thought to reflect the brain's attempt to recreate previously experienced associations between internal physiology and external situations. Complex social situations engage the orbitofrontal cortex, which then activates somatic effectors in the amygdala, hypothalamus, and brain-stem nuclei. Somatic markers permit the rapid processing of possible behavioral responses and evaluation of the adaptive value of their associated outcomes. Decision-making can then selectively focus on option–outcome pairings that are potentially rewarding. Therefore, emotion (in the form of somatic markers) will be particularly beneficial in ambiguous situations in which decisions can only be based on similar past experiences (Bechara, Damasio, & Damasio, 2000; Elliot, Dolan, & Frith, 2000).

The Somatic Marker Hypothesis is predominantly based on lesions studies that adopt a gambling paradigm called the Iowa Gambling Task (e.g., Bechara, Damasio, Tranel, & Damasio, 1997). In the Iowa Gambling Task, skin conductance response (SCR) is measured while participants continually draw cards from their choice of four decks. The cards indicate monetary amounts, resulting in either gain or loss. Unbeknownst to the participants, two decks are associated with net winnings—they have low payoffs but have even lower losses. The other two decks are associated with net losses—they have high payoffs but even larger losses. It is up to the participants to figure out how to optimize their winnings by favoring the decks associated with net winnings. Healthy adults and patients with damage outside the orbitofrontal cortex (either within the frontal lobes or outside the frontal lobes) learn the task and gamble in a manner that maximizes winnings. In contrast, orbitofrontal patients learn the task but fail to implement an optimal gambling strategy on a behavioral level. This failure is interpreted as resulting from a parallel deficit in physiological responses to the task. Healthy adults show an increased SCR in anticipation of making a risky gamble. In contrast, patients with orbitofrontal damage show no anticipatory change in SCR. These results are interpreted as indicating that orbitofrontal damage—particularly to the right side (Tranel et al., 2002)—impairs decision making because somatic markers are not triggered and, therefore, cannot guide gambling decisions. These findings are consistent with a lesion study that suggests that patients with orbitofrontal damage may sometimes make better financial decisions because their decisions are not shaped by emotions presumed to arise from the outcome of their last financial decision (Shiv et al., 2005). In this study, participants completed a series of trials in which they could buy a chance to participate in a lottery (e.g., 50–50 chance that they would win $2.50 or lose their payment of $1) or decline to play in that trial (e.g., nothing to gain or lose). Healthy controls tended to decline a chance to play much more often if they had lost than if they had won in the previous trial (i.e., participated 40.5% compared to 61.7% of the time). In contrast, orbitofrontal patients' decisions to buy a chance at the lottery did not differ much as a function of whether they had lost or won in the previous trial (participated 79.8% compared to 79.1% of the time). The researchers suggest that healthy controls experienced negative emotions after a loss, and this emotional state reduced their interest in risking money to participate in the lottery. However, monetary gain is only possible through participation in the lottery, and therefore, orbitofrontal patients' higher rates of lottery participation allowed them to perform better in this task.

Reinforcement and Reversal

Another perspective accounts for the role of orbitofrontal cortex in emotional decision making through reinforcement and reversal processes (Kringelbach & Rolls, 2004; Rolls, 2000). This perspective proposes that medial orbitofrontal cortex (BA 11/12) computes the reward value of stimuli and the lateral orbitofrontal cortex computes the punishing value of stimuli that may lead to a change in behavior (Kringelbach & Rolls, 2004). From this perspective, emotion necessarily arises from reward or punishment. As environmental contexts change, the orbitofrontal cortex learns new reward and punishment associations. Therefore, orbitofrontal patients make poor decisions because they are unable to adjust behavior in reference to changing rewards and punishments (i.e., emotions).

This positions draws on both animal and human research. Single-unit recording from orbitofrontal neurons in rhesus monkeys were collected during a reversal go–no go task (Thorpe, Rolls, & Maddison, 1983). In a go–no go task, participants learn to respond ("go") or withhold a response ("no go") to stimuli based on their association with the delivery of a reward or a punishment. For example, monkeys learn to press a key in response to particular geometric shapes to earn juice rewards and avoid electric shock. In the reversal version, stimuli values are periodically reassigned so that the reward or punishment values of stimuli change throughout the task. Learning the initial stimulus–reinforcement contingencies (e.g., responding to the presence of a reward or responding to the presence of punishment) is associated with orbitofrontal neuronal activity (Thorpe, Rolls, & Maddison, 1983). In cases of reversal, orbitofrontal neurons also respond to a lack of an expected reward and to the presence of an unexpected reward. Therefore, orbitofrontal cortex re-establishes the reward and punishment value of stimuli as contingencies change. In other words, the orbitofrontal cortex suppresses irrelevant emotional responses to stimuli with new emotional meaning. The findings from the monkey literature are consistent with human studies of impaired reversal learning in orbitofrontal patients (Fellows & Farah, 2003; Rolls, Hornak, Wade, & McGrath, 1994). Specifically, orbitofrontal patients are able to learn an initial stimulus–reinforcement association but do not modify behavior once associations are reversed or extinguished. This perseveration is not observed in patients with damage outside of the orbitofrontal cortex or in healthy control participants. One study suggested that errors on reversal and extinction tasks predict the extent of patients' social disinhibition as rated by staff members (Rolls et al., 1994). In other words, social problems are proposed to arise from difficulties making new behavioral decisions as stimulus-reinforcement contingencies (i.e., emotions) change.

Dynamic Filtering Theory

The orbitofrontal region of the prefrontal cortex has also been implicated in integrating emotion and cognitive information through a gating mechanism (Shimamura, 2000). This theory draws on the general executive function of the prefrontal lobes and focuses on the orbitofrontal cortex as a region of control over emotional processing because of its heavy connections to sensory and limbic areas. Patients with orbitofrontal damage may be overwhelmed by their emotions as they are unable to inhibit the neural activity associated with emotional processing. In this case, it would be expected that patients with orbitofrontal damage would show increased emotional biases in decision making, as they are unable to suppress their emotional responses.

The proposal that orbitofrontal damage impairs the ability to suppress emotional responses is supported by an ERP study conducted with orbitofrontal patients (Rule, Shimamura, & Knight, 2002). Participants, including patients with orbitofrontal damage or dorsolateral prefrontal damage and healthy control subjects, were presented with mild shocks or distracting noises while watching a movie. In this task, the shocks and distracting noises were meant to elicit emotional responses. In comparison to the dorsolateral prefrontal group and healthy controls, orbitofrontal patients showed greater P300 amplitudes in both the shock and noise condition. Healthy control subjects eventually habituated for the shock condition, but orbitofrontal patients never did. No significant difference for habituation for the auditory stimuli was found between the orbitofrontal and control groups. These findings suggest that the orbitofrontal cortex is important for regulating neural activity associated with emotional stimuli. Patients with orbitofrontal damage do not habituate to aversive somatosensory stimuli, presumably because they can not suppress their response to the stimuli.

Evaluation of emotional decision making and orbitofrontal function

The neuroscientific research on emotional decision making suggests a number of possible ways the orbitofrontal cortex may support effective decision making driven by emotion. What conclusion can be drawn about the specific role of the orbitofrontal cortex in this process? The opening anecdote about the blind men and the elephant emphasized the importance of multimethod investigations to generate scientific answers. How might a social psychologist interested in the adaptive influence of emotion on decision making evaluate the neuroscience research? What future directions are needed to integrate empirical evidence across these two fields that examine emotional decision making?

A social psychological perspective raises questions about the claim that orbitofrontal cortex mediates emotional influences on cognition through the interpretation of physiological arousal. Research has shown that although physiological arousal may be involved, it is too slow to exclusively account for emotional priming influences on cognition (e.g., Fiske & Taylor, 1991). Studies have also shown that patients with spinal cord injuries, bereft of physiological feedback, report subjective experiences of emotion like those of healthy control participants (Bermond et al., 1991; Chwalisz et al., 1988; but *see* Hohmann, 1966) and do not show impaired performance on gambling tasks (Dunn, Dalgleish, & Lawrence, 2006; North & O'Carroll, 2001). Additionally, research has shown that individuals tend to misattribute the source of their physiological arousal, and consequent decisions may be guided by those misattributions (e.g., Dutton & Aron, 1974). This set of behavioral findings suggests that it is unlikely that patterns of arousal (or mentally represented patterns of arousal) are fundamental for ensuring nonbiased decisions.

A comparison of the neuroscience research and behavioral research also reveals that it is difficult to integrate these two lines of research because of differences in the operationalization of emotion. Differences of emotion measurement are also very evident across the neural studies. In the orbitofrontal studies alone, emotion is operationalized as somatic markers, gains or losses in financial decision-making tasks, positive or negative appraisals, or physical pain. Standardization of emotion measurement will permit better synthesis of findings across the neural and behavioral levels of analysis. Standardized emotion measurement may also help pinpoint the specific role of the orbitofrontal cortex in emotional decision making; different theories may have arisen because some studies involve emotion and others do not. For example, do go–no go paradigms really elicit emotion? In this paradigm, participants are given points to reinforce their behavior as they learn when to produce a response and when to withhold a response. From this perspective, learning that a behavior is "good" is considered to be an emotional process (e.g., Rolls, 2000). However, social psychologists might argue that developing valenced associations for behaviors is better described as attitude or preference formation or that reward and punishment shape motivation to approach or avoid objects. A standardized operationalization of emotion has been developed through empirical investigations by social psychologists. From this perspective, emotion is defined as a short-lived psychological–physiological phenomena that coordinates modes of adaptation to changing environmental demands (Levenson, 1999). Empirical studies suggest that changes at three levels of measurement reflect the presence of emotion: self-report, physiological assessment, and coding of facial expression.

Refining theories of orbitofrontal function from a social psychological perspective: A monitoring hypothesis

The blended approach of social psychological conceptualization and measurement of emotion and decision making with neuroscience methodology (i.e., a social neuroscience approach) characterizes our own research on orbitofrontal function. Our studies suggest that orbitofrontal damage does not impair the ability to generate emotional responses as assessed by physiological, facial muscle movement, and self-report measures. Damage to the orbitofrontal cortex may impair the ability to monitor the contextual relevance of one's behavior and, therefore, may preclude the generation of emotion that is useful for subsequent decision making. The monitoring function of orbitofrontal cortex also extends to evaluating when emotional information should be incorporated into decision-making and when emotional influences should be inhibited. The new perspective on orbitofrontal function does not disprove an adaptive emotional influence on decision making; rather, it highlights the need for future neural research on this question.

Orbitofrontal Cortex and Experimentally Induced Emotionality

To address the discrepancy between emotion measurement in neural studies and behavioral studies, we first examined whether patients with orbitofrontal damage could experience emotions (Beer, 2007). It was possible that orbitofrontal damage might be associated with diminished emotion, as suggested by the Somatic Marker Hypothesis, or particularly intense emotion, as suggested by Dynamic Filtering Theory. We compared the performance of patients with orbitofrontal cortex damage to the performance of age-matched controls in a series of emotion-eliciting tasks (Gross & Levenson, 1995; Keltner, 1995). Participants watched a series of standardized film clips that have been shown to elicit a discrete emotional state (i.e., one film each for amusement, disgust, anger, sadness, or contentment; Gross & Levenson, 1995). Additionally, embarrassment was examined by asking participants to pose a silly face and hold it while viewing themselves on a monitor (Keltner, 1995). Three measures were used to assess emotion: autonomic nervous system physiology, questionnaire, and facial muscle movement (Facial Action Coding System [FACS]; Ekman & Friesen, 1978). We found no significant differences across the groups, with one exception. Orbitofrontal damage was associated with increased self-reports and facial expressions of embarrassment. In an additional series of tasks, we examined whether participants could suppress their emotional facial expression in response to a disgusting film. No significant differences in the ability to suppress facial expressions of disgust were found between the groups. The self-report and autonomic physiological measures also did not differ between the groups. Evidence of suppression in both groups was reflected in the facial twitching typically associated with efforts to suppress facial expressions (Gross & Levenson, 1993). These findings suggest that orbitofrontal cortex does not impair the ability to experience or suppress emotion.

Orbitofrontal Cortex and Spontaneous Emotionality

Although the previous study demonstrated that orbitofrontal patients were able to generate and suppress emotions, it did not examine the effect of orbitofrontal damage on emotion generation in day-to-day life. In a second study, we examined the emotions of participants while they engaged in two social interaction tasks: a teasing task and an overpraise task (Beer et al., 2003). In the teasing task, participants had to make up nicknames for two experiments that they did not know well. The overpraise task required participants to generate a creative title for a paragraph that was read aloud to them. The paragraph intentionally had no content, making it difficult to generate a title that would be considered creative. After the participants made up a title, the experimenters praised them for 2 minutes. When individuals are sincerely praised, they tend to become embarrassed because it violates social norms to pat one's self on the back. This task created a situation in which the praise was clearly undeserved and therefore was intended to be amusing or surprising. The study showed that orbitofrontal patients act inappropriately, and their emotions are unexpected given their inappropriate behavior. In the teasing task, orbitofrontal patients exhibited teasing behavior that was objectively more inappropriate than that of control participants. Rather than being embarrassed by their inappropriate teasing, the orbitofrontal patients were more proud of their behavior. In the overpraise task, the orbitofrontal patients exhibited embarrassment, as if the praise was deserved, whereas the control participants exhibited amusement. In other words, patients with orbitofrontal damage exhibit the kind of emotion that would be expected for individuals who had not acted in an offensive manner and had genuinely excelled at the title task. Together, these studies suggest that orbitofrontal damage is associated with a discrepancy between emotion and behavior.

Orbitofrontal Cortex and Self-Monitoring

One possible explanation for the discrepancy between behavior and emotion is that orbitofrontal patients lack insight into their behavior. In other words, orbitofrontal patients' emotion may reflect an erroneous belief that they had acted appropriately during the task. Without awareness of mistakes, an individual has no reason to become embarrassed, and therefore, the experience of embarrassment cannot motivate the selection of new behaviors to avoid the repetition of the mistake. In this case, orbitofrontal cortex may be important for insight into behavior and only affects emotional decision -making in a distal manner. To examine the association between self-insight, emotion, and orbitofrontal damage, we conducted a study in which we could measure self-insight and emotion and then examine how emotion changed as self-insight became more accurate. Orbitofrontal patients (*see* Fig. 11–1), healthy controls, and brain-damaged controls (i.e., dorsolateral prefrontal damage) took part in a self-disclosure task (Aron et al., 1992). In the self-disclosure task, an experimenter asked each participant a series of questions. Some questions were appropriate to discuss with a stranger (e.g., What would be a perfect day for you?), and some were more appropriate for a discussion with a friend (e.g., If you were going to pass away this evening with no chance to speak to anyone, what would you most regret not having told someone and why haven't you told them yet?). The measurement of appropriate self-disclosure relied on participants' understanding of social norms against excessive disclosure of personal information to strangers. The orbitofrontal patients and both control groups demonstrated equal knowledge of this social norm. After the self-disclosure task, participants reported on their perceptions of their social appropriateness and their emotional experiences during the task. We then manipulated insight into behavior by showing participants a videotape of their task performance and examined how emotion changed.

The study found that orbitofrontal patients disclosed more personal and inappropriate information than the other participants. Before viewing their videotaped behavior, orbitofrontal patients had positively inflated perceptions of their social appropriateness and were not embarrassed by their behavior. In contrast to their initial emotion ratings, orbitofrontal patients' embarrassment significantly increased after viewing their videotaped behavior. These findings support the theory that orbitofrontal cortex mediates online monitoring of behavior (e.g., in reference to social norms). Emotional

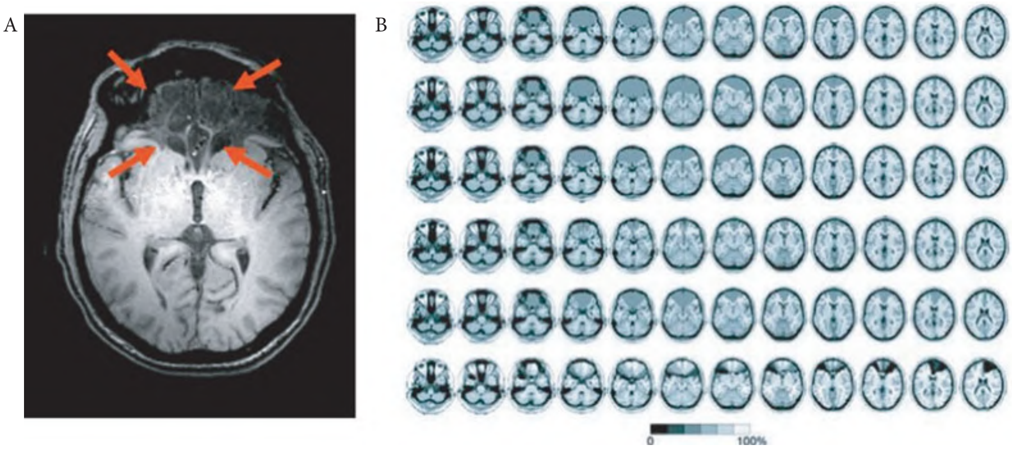


Fig. 11–1 **Panel A: An axial slice through orbitofrontal cortex from an orbitofrontal lesion patient (Patient DH from Beer et al., 2006). Note the lesion completely destroys the lateral and medial orbitofrontal cortex (arrowheads) and spares anterior temporal lobes. Panel B: Five patients with bilateral orbitofrontal cortex damage. The first five rows show the extent of damage in an individual patient as transcribed onto axial templates using 5-mm cuts. The bottom row represents the extent of lesion overlap across subjects.**

decision making may be impacted by orbitofrontal damage because emotional experience may be driven by faulty perceptions of one's behavior. In other words, impaired self-insight may preclude the generation of the emotions needed to guide decision making.

Orbitofrontal Cortex and Monitoring Emotional Influences on Decision Making

If the orbitofrontal cortex does serve a monitoring function, then it may affect emotional decision making aside from impacting self-insight. Although neural models tend to assume that emotional information is either helpful (e.g., Somatic Marker Hypothesis, Reinforcement Model) or hurtful (e.g., Reversal Model), it is clear that emotional influences on decision making may be helpful or hurtful. Emotions influence attention and the amount of cognitive resources we devote to decision making (e.g., Forgas, 2002). The direction of attention and rapidity in decision making can sometimes be advantageous, such as fear motivating the decision to freeze upon seeing a snake on a trail. On the other hand, residual anger from a frustrating commute may motivate snap decision making where deliberation may be more advantageous. The complex role of emotion in decision making and the previous finding that orbitofrontal cortex may serve a monitoring function led us to ask a new question about the involvement of orbitofrontal cortex in emotional decision making. Does the orbitofrontal cortex monitor whether emotion should be incorporated or inhibited in situations of decision making? A series of fMRI studies of healthy individuals has supported the theory that orbitofrontal cortex is involved in mediating emotional influences on decision making by evaluating the relevance of the emotional information (Beer, Knight, & D'Esposito, 2006). Participants were presented with negative neutral pictures as they placed bets in a gambling task (i.e., a roulette game). In the helpful condition, participants were told that the pictures held a clue about the upcoming gamble. Specifically negative pictures indicated high risk in comparison to neutral pictures. Previous studies suggest that individuals are likely to reduce their gambles in relation to fearful pictures because of increased perceptions of risk (e.g., Johnson & Tversky, 1983). In the hurtful condition, participants were told the pictures did not hold a clue about the upcoming bet. This required participants to suppress the normative influence of the negative pictures on betting decisions. Orbitofrontal cortex was recruited for appropriately applying the emotional information to the subsequent gambling decision. In other words, orbitofrontal cortex was recruited for betting that was influenced by helpful emotion cues (e.g., reduced gambling in relation to negative emotion) and recruited for inhibiting the effect of hurtful emotion cues (e.g., gambling decisions that did not differ as a function of cue). These studies suggest that orbitofrontal cortex is involved emotional decision- making by regulating response selection in relation to the helpful or hurtful nature of emotion for a particular decision.

In summary, the social neuroscience approach suggests that the orbitofrontal cortex is important for monitoring functions. These monitoring functions may influence emotional decision making by *(1)* supporting insight into the appropriateness of behavior that affects the emotions that are generated and *(2)* monitoring whether emotional decision making is appropriate. However, the primary function of orbitofrontal cortex is not to apply emotion to decision making. These studies should not be considered evidence against the view that emotion can have an adaptive role in decision-making. Instead the implication is that neural investigations of emotional decision-making should move beyond the focus on orbitofrontal cortex and examine whether other brain regions support adaptive influences of emotion on decision making.

Neural models of emotional decision-making: focus beyond the orbitofrontal cortex

There are a small number of studies that have examined the neural mediation of mood influences on cognition. For example, investigators conducted a PET study to examine the neural activity associated with *(1)* elated and *(2)* depressed mood influences on verbal fluency (Baker et al., 1997). Mood manipulations involved a combination of procedures. For example, the elated mood manipulation required participants to read positive sentences, listen to positive music, and accept a gift certificate before getting in the scanner. The verbal fluency task required participants to nominate as many words as possible that began with a particular letter (e.g., the letter "B"). Although differences in activity were found during the verbal fluency task for the different mood conditions, no behavioral differences were found across the conditions. It will be necessary to replicate this effect in the contexts of behavioral differences to draw strong conclusions about the relation between the mood manipulations and differences in brain activity. Another study examined emotional influences on memory for words and faces (Gray et al., 2002). Participants viewed emotional films and then performed a three-back task for words and faces. A three-back task requires participants to judge whether a currently presented stimuli is the same stimuli that was presented three trials earlier. This task required individuals to remember the sequence of words and faces as they were presented. A marginally significant behavioral effect was found: memory for words was reduced by negative emotion, and memory for faces was enhanced by negative emotion. Activity in the right dorsolateral prefrontal cortex accounted for the differential emotional influences on memory for words and faces. The authors of the study note that the psychological mechanism through which emotion differentially affects memory for words and faces is unclear, and future research is needed to more fully understand the psychological meaning of the dorsolateral prefrontal activity.

Indirect evidence for the neural systems supporting adaptive emotional influences on decision making comes from a recent wave of research combining economic and neuroscience approaches (e.g., Rilling et al., 2002; Sanfey et al., 2003). The main focus of these studies is not emotional decision making, and therefore, emotion is not directly manipulated and then examined in relation to decision making. However, these studies have found that decision-making tasks significantly recruit brain regions previously associated with emotion. For example, one study examined decision making using the Ultimatum Game (Sanfey et al., 2003). In the Ultimatum Game, participants must split a sum of money with another player. In one condition, the other player offers a portion of the sum to the participant and the participant must decide to accept or reject the offer. The offers may be fair (e.g., very close to 50% for each person) or unfair (e.g., 80% for the player and 20% for the participant). The consideration of unfair offers was associated with insula activity. Insula activity has often been associated with negative emotions such as disgust, anger, pain, and distress, suggesting that the participants may have experienced these emotions while considering the offer. From a rational economic perspective, acting on the negative emotion by refusing the offer is maladaptive because the participant gains no money when even a small amount may have been available. However, from a broader perspective, the negative emotional reaction to unfair offers is interpreted as advantageous because the acceptance of unfair offers over time (even from different individuals) may threaten social status. Another study examined decision making in a Prisoner's Dilemma game (Rilling et al., 2002). In a Prisoner's Dilemma game, participants win money as a function of their own decision to cooperate or betray and their partner's decision to cooperate or betray. The choice to cooperate is a double-edged sword; participants win the most if both players choose to cooperate but lose the most if they decide to cooperate and the other player decides to betray. In this study, cooperation was associated with areas associated with reward processing (e.g., nucleus accumbens, orbitofrontal cortex, anterior cingulate, and caudate nucleus). The authors suggest that this activation reflects a positive emotion experience that reinforces prosocial decision making. The positive emotional reaction is adaptive in this case because although the player may gain less in monetary units, social cooperation may prevent ostracism from the group.

Future directions

This chapter has challenged the view that neural investigations of orbitofrontal cortex function provide strong support for the adaptive role of emotion in decision making. Although the orbitofrontal cortex was considered to mediate emotional influences on decision making, research adopting a social neuroscience approach has demonstrated that the orbitofrontal cortex only distally impacts emotional decision making through self-monitoring processes (e.g., insight into online behavior, insight into the relevance of emotional information). This evidence does not refute the adaptive role of emotion in decision making; rather, it suggests that future research is needed to better understand whether areas outside of the orbitofrontal cortex are involved in adaptive emotional decision making. Future research in the domain of emotional decision making might also focus on a number of related questions.

First, emotions can be broadly categorized (i.e, positive, negative, or self-conscious), but behavioral research suggests that studying specific emotions in relation to decision making will be the most powerful avenue of investigation. For example, both fear and anger are negative emotions, but each has a different consequence for decision making (Lerner & Keltner, 2000). Fear reduces risk-taking and anger promotes risk-taking.

Second, does emotion really have a unitary effect on decision making? Emotions are characterized by a number of components, including valence, arousal, and motivational tendencies that might influence decision making. Valence may shift attention toward valence-congruent information, arousal may impact availability of cognitive resources, and motivational tendencies may speed the execution of particular actions. When an emotion influences risky decision making, do separate neural systems compute the influence of valence, arousal, and motivational tendencies or is there an overarching system that computes the influence of all of these factors? Although there are likely to be additional areas recruited for each component, common areas recruited across these factors would support the theory that these factors underlie the overarching psychological construct of emotion (e.g., Berman, Jonides, & Nee, 2006).

Third, another interesting question that will benefit from the neural level of analysis is the mechanism by which mixed emotional experiences influence decision making. Early investigations of mixed emotional experiences examined the daily emotional events of a group of 42 students over a period of 6 weeks (Diener & Iran-Nejad, 1986). The researchers concluded that emotions of the same valence may be reported together, but it was rare for a positive and negative emotion to be reported simultaneously. However, this study also found that the intensity of one emotion did not predict the intensity of a simultaneously felt emotion. This left open the possibility that mixed emotional experiences could occur and that each emotion's intensity was independent of the other. More recently, a series of studies has shown that people report negative and positive emotion simultaneously in certain situations (or at the very least, these emotions cycle so rapidly that they are represented as simultaneous emotion states) (Larsen, McGraw, Mellers, & Cacioppo, 2004; Larsen, McGraw, & Cacioppo, 2001). For example, college graduation may elicit feelings of both sadness and happiness. Do the valence, arousal, and motivational tendencies associated with each emotion become averaged or do they conflict? Examining the neural activity in relation to the influence of mixed emotional experiences (within valence and across valence) on decision making will address this question. However, this was not a study in which an elicited mixed emotion experience influenced decision making. If areas associated with singular emotional influences on decision making are activated for mixed emotions, then it is likely that mixed emotions are averaged or summed in some way when they influence decision making. However, if areas associated with conflict or response competition are activated, then it is likely that mixed emotions provide sources of information that clash in their influence on decision making. One study has examined neural activity in relation to judging mixed emotional stimuli (Simmons, Stein, Matthews, Feinstein, & Paulus, 2006). This study required individuals to make judgments in relation to a "Wall of Faces"—that is, the simultaneous presentation of many different individuals' facial expressions. Participants were asked to judge whether more of the faces expressed a particular emotion, such as amusement (emotion condition), or whether more of the faces belonged to women than to men (gender condition). Although these judgments were straightforward if the faces were distributed unequally across emotion or gender, there was also an ambiguous condition in which 50% of the faces contained the target emotion or gender. Ventromedial prefrontal activity was associated with judgments of emotion compared to gender in the ambiguous condition. However, the study was not designed to examine the influence of a mixed emotional state on decision making, so it is not known if the emotional face condition really elicited mixed emotions in the participants.

Conclusion

A more favorable view of the role of emotion in decision making has recently emerged within psychology and neuroscience. From the neuroscience perspective, the main focus of this line of research has been the orbitofrontal cortex. However, social neuroscience research has shown that the primary function of the orbitofrontal cortex is better characterized by self-monitoring processes. Deficits in these monitoring processes affect emotional decision making, but only in a distal manner. This evidence does not refute the possibility of adaptive emotional influences on decision making but suggests that future neural investigations of systems supporting effective decision making that relies on emotion should focus on a system outside of the orbitofrontal cortex. Many future avenues of neural investigation are suggested by social psychological theory, and paradigms and will promote meaningful synthesis of empirical evidence across the fields of psychology and neuroscience.

References

Adolphs, R. (1999). Social cognition and the human brain. *Trends in Cognitive Sciences*, 3, 469–479.

Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience*, 2, 1032–1037.

Aron, A., Melinat, E., Aron, E. N., Vallone, R. D., & Bator, R. J. (1997). The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and Social Psychology Bulletin*, 4, 363–377.

Baker, S. C., Frith, C. D., & Dolan, R. J. (1997). The interaction between mood and cognitive function studied with PET. *Psychological Medicine*, 27, 565–578.

Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making, and the orbitofrontal cortex. *Cerebral Cortex*, 10, 295–307.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, 275, 1293–1295.

Beer, J. S. (2007). The importance of emotion–cognition interactions for social adjustment: Insights from the orbitofrontal cortex. In E. Harmon-Jones & P. Winkielman (Eds.) *Social Neuroscience: Integrating Biological and Psychological Explanations of Social Behavior* (pp. 15–30). New York: Guilford.

Beer, J. S., Heerey, E. H., Keltner, D., Scabini, D., & Knight, R. T. (2003). The regulatory function of self-conscious emotion: Insights from patients with orbitofrontal damage. *Journal of Personality and Social Psychology*, 85, 594–604.

Beer, J. S, Knight, R. T., & D'Esposito, M. (2006). Integrating emotion and cognition: The role of the frontal lobes in distinguishing between helpful and hurtful emotion. *Psychological Science*, 17, 448–453.

Beer, J. S., Roberts, N. A., Werner, K. H., et al. (2001). Orbitofrontal cortex and self-conscious emotion. *Society for Neuroscience Abstracts*, 27, 1705.

Beer, J. S., Shimamura, A. P., & Knight, R. T. (2004). Frontal lobe contributions to executive control of cognitive and social behavior. In M. S. Gazzaniga (Ed.) *The Newest Cognitive Neurosciences* (3rd ed., pp.1091–1104). Cambridge: MIT Press.

Berman, M. G., Jonides, J., & Nee, D. E. (2006). Studying mind and brain with fMRI. *Social Cognitive and Affective Neuroscience*, 1, 158–161.

Brothers, L. (1996). Brain mechanisms of social cognition. *Journal of Psychopharmacology*, 10, 2–8.

Chwalisz, K., Diener, E., & Gallagher, D. (1988). Autonomic arousal feedback and emotional experience: Evidence from the spinal cord injured. *Journal of Personality and Social Psychology*, 54, 820–828.

Cohen, D., Nisbett, R. E., Bowdle, B. F., & Schwartz, N. (1996). Insult, aggression, and the southern culture of honor: An "experimental ethnography." *Journal of Personality and Social Psychology*. 70, 945–960.

Diener, E., & Iran-Nejad, A. (1986). The relationship in experience between various types of affect. *Journal of Personality and Social Psychology*, 50, 1031–1038.

DeMartino, B., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. *Science*, 313, 684–687.

Drevets, W. C., & Raichle, M. E. (1998). Reciprocal suppression of regional cerebral blood flow during emotional versus high cognitive processess: Implications for interactions between emotion and cognition. *Cognition and Emotion*, 12, 353–385.

Dunn, B. B., Dalgleish, T., & Lawrence, A. D. (2006). The somatic marker hypothesis: A critical evaluation. *Neuroscience and Biobehavioral Reviews*, 30, 239–271.

Dutton, D. G., & Aron, A. P. (1974). Some evidence for heightened sexual attraction under conditions of high anxiety. *Journal of Personality and Social Psychology*, 30, 510–517.

Duval, S., & Wicklund, R. A. (1972). *A Theory of Objective Self-awareness*. New York: Academic Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*. 6, 169–200.

Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The duchenne smile: Emotional expression and brain physiology II. *Journal of Personality and Social Psychology*, 58, 342–353.

Elliott, R., Dolan, R. J., & Frith, C. D. (2000). Dissociable functions in the medial and lateral orbitofrontal cortex: Evidence from human neuroimaging studies. *Cerebral Cortex*, 10, 308–317.

Emery, N. J., & Amaral, D. G. (2000). The role of the amygdala in primate social cognition. In R. D. Lane & L. Nadel (Eds.) *Cognitive Neuroscience of Emotion* (pp. 156–191). New York: Oxford University Press.

Fiske, S. T., & Taylor, S. E. (1991). *Social Cognition* (2nd ed.). San Francisco, CA: McGraw Hill.

Fellows, L. K., & Farah, M. J. (2003). Ventromedial frontal cortex mediates affective shifting in humans: Evidence from a reversal learning paradigm. *Brain*, 126, 1830–1837.

Forgas, J. P. (1995). Mood and judgment: The affect infusion model (AIM). *Psychological Bulletin*, 117, 39–66.

Forgas, J. P. (1998). On being happy and mistaken: Mood effects on the fundamental attribution error. *Journal of Personality and Social Pscyhology*, 75, 318–331.

Forgas, J. P. (2002). Feeling and doing: Affective influences on interpersonal behavior. *Psychological Inquiry*, 13, 1–28.

Gray, J. R., Braver, T. S., & Raichel, M. E. (2002). Integration of emotion and cognition in lateral prefrontal cortex. *Proceedings of the National Academy of Sciences*, 99, 4115–4120.

Gross, J. J., & Levenson, R. W. (1993). Emotional suppression: Physiology, self-report, and expressive behavior. *Journal of Personality and Social Psychology*, 64, 970–986.

Gross, J. J., & Levenson, R. W. (1995). Emotion elicitation using films. *Cognition and Emotion*, 9, 87–108.

Hornak, J., Rolls, E. T., & Wade, D. (1996). Face and voice expression identification in patients the emotional and behavioural changes following ventral frontal lobe damage. *Neuropsychologia*, 34, 247–261.

Isen, A. M., & Geva, N. (1987). The influence of positive affect on acceptable level of risk: The person with a large canoe has a large worry. *Organizational Behavior and Human Decision Processes*, 39, 145–154.

Izard, C. E. (1971). *The Face of Emotion*. New York: Appleton-Century-Crofts.

Johnson, E. J., & Tversky, A. (1983). Affect, generalization, and the perception of risk. *Journal of Personality and Social Psychology*, 45, 20–31.

Keltner, D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*. 68, 441–454.

Keltner, D., Ellsworth, P. C., & Edwards, K. (1993). Beyond simple pessism: Effects of sadness and anger on social perception. *Journal of Personality and Social Psychology*, 64, 740–752.

Keltner, D., & Kring, A. M. (1998). Emotion, social function, and psychopathology. *Review of General Psychology*, 2, 320–342.

Kringelbach, M. L., & Rolls, E. T. (2004). The functional neuroanatomy of the human orbitofrontal cortex: Evidence from neuroimaging and neuropsychology. *Progress in Neurobiology*, 72, 341–372.

Lang, P. J. (1978). A Bio-informational theory of emotional imagery. *Psychophysiology*, 16, 495–512.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1995). *The International Affective Picture System (IAPS): Photographic Slides*. University of Florida: The Center for Research in Psychophysiology.

Larsen, J. T., McGraw, A. P., & Cacioppo, J. T. (2001). Can people feel happy and sad at the same time? *Journal of Personality and Social Psychology*, 81, 684–696.

Larsen, J. T., McGraw, A. P., Mellers, B. A., & Cacioppo, J. T. (2004). The agony of victory and thrill of defeat: Mixed emotional reactions to disappointing wins and relieving losses. *Psychological Science*, 15, 325–330.

Lerner, J. S., & Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology*, 81, 146–159.

Levenson, R. W. (1999). The intrapersonal functions of emotion. *Cognition and Emotion*, 13, 481–504.

Miller, R. S. (1990). Embarrassment and social behavior. In R. Crozier (Ed.), *Shyness and Embarrassment*. New York: Cambridge University Press.

North, N. T., & O'Carroll, R. E. (2001). Decision making in patients with spinal cord damage: Afferent feedback and the somatic marker hypothesis. *Neuropsychologia*, 39, 521–524.

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.

Rahman, S., Sahakian, B. J., Cardinal, R. N., Rogers, R. D., & Robbins, T. W. (2001). Decision making and europsychiatry. *Trends in Cognitive Sciences*, 5, 271–277.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron*, 35, 395–405.

Roberts, N. A., Werner, K. H., Beer, J. S., et al. (2004). The impact of orbital prefrontal cortex damage on emotional reactivity during acoustic startle. *Cognitive, Affective, and Behavioral Neuroscience*, 4, 307–316.

Rolls, E. T. (2000). The orbitofrontal cortex and reward. *Cerebral Cortex*, 10, 284–294.

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.

Rule, R., Shimamura, A., & Knight, R. T. (2002). Orbitofrontal cortex and dynamic filtering of emotions, *Cognitive, Affective, and Behavioral Neuroscience*, 2, 264–270.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, 300, 1755–1758.

Saver, J. L., & Damasio, A. R. (1991). Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage. *Neuropsychologia*, 29, 1241–1249.

Schwarz, N. (1990). Feelings as information: Informational and motivational functions of affective states. In Higgins, E. T., & Sorrentino, R. M. (Eds.), *Handbook of Motivation and Cognition: Foundations of Social Behavior (Vol. 2).* New York: Guilford.

Shaver, P. R., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52, 1061–1086.

Shimamura, A. P. (2000). The role of the prefrontal cortex in dynamic filtering. *Psychobiology*, 28, 207–218.

Shiv, B., Loewenstein, G., Bechara, A., Damasio, H., & Damasio, A. R. (2005). Investment behavior and the negative side of emotion. *Pscyhological Science*, 16, 435–439.

Simmons, A., Stein, M. B., Matthews, S. C., Feinstein, J. S., & Paulus, M. P. (2006). Affective ambiguity for a group recruits ventromedial prefrontal cortex. *NeuroImage*, 29, 655–661.

Sprengelmeyer, R., Rausch, M., Eysel, U. T., & Przuntek, H. (1998). Neural structures associated with recognition of facial expressions of basic emotions. *Proceedings of the Royal Society of London*, 265, 1927–1931.

Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience*, 10, 640–656.

Thorpe, S. J., Rolls, E. T., & Maddison, S. (1983). The orbitofrontal cortex: Neuronal activity in the behaving monkey. *Experimental Brain Research*, 49, 93–115.

Tranel, D., Bechara, A., & Denburg, N. L. (2002). Asymmetric functional roles of right and left ventromedial prefrontal cortices in social conduct, decision making and emotional processing. *Cortex*, 38, 589–612.

Tucker, D. M., Luu, P., & Pribram, K. H. (1995). Social and emotional self-regulation. *Annals of New York Academy of Sciences*, 769, 213–239.

Whalen, P. J. (1998). Fear, vigilance, and ambiguity: Initial neuroimaging studies of the human amygdala. *Current Directions in Psychological Science*, 7, 177–188.

CHAPTER 12
Social Neuroscience of Asymmetrical Frontal Cortical Activity: Considering Anger and Approach Motivation

Eddie Harmon-Jones & Cindy Harmon-Jones

Contemporary dimensional models of emotion regard the positive to negative valence dimension as an important organizing principle (Lang, 1995; Watson, 2000). Over the last three decades, this principle has been used to organize empirical observations of the relationship between left versus right (asymmetrical) frontal cortical activations and emotional experience and expression. In this body of research, positive affect has been found to relate to relatively greater left than right frontal cortical activity, whereas negative affect has been found to relate to relatively greater right than left frontal cortical activity.

The interest in the relationship between asymmetrical frontal brain activity and emotional valence was sparked in part by systematic observations that damage to the left frontal cortex caused depression, whereas damage to the right frontal cortex caused mania (Robinson, Kubos, Starr, Rao, & Price, 1984). Following closely after these observations, research demonstrated that both trait and state positive affect was associated with increased left frontal cortical activity, whereas trait and state negative affect was associated with increased right frontal cortical activity (*see* review by Silberman & Weingartner, 1986). Conceptually similar results have been obtained using a wide variety of neuroscience methods, including lesion studies (Robinson & Downhill, 1995), repetitive transcranial magnetic stimulation (rTMS; van Honk, Schutter, d'Alfonso, Kessels, & de Haan, 2002), positron emission tomography (PET; Thut et al., 1997), fMRI (Canli, Desmond, Zhao, Glover, & Gabrieli, 1998), event-related brain potentials (ERPs; Cunningham, Espinet, DeYoung, & Zelazo, 2005), and EEG (Coan & Allen, 2003). Moreover, these effects have been observed in nonhuman and human animals (Berridge, España, & Stalnaker, 2003).

Until the late 1990s, all studies examining the relationship between asymmetrical frontal cortical activity and emotion confounded affective valence (positive vs. negative affect) with motivational direction. That is, all positive affective states/traits (e.g., joy, interest) that had been empirically examined were approach motivating, whereas all negative affective states/traits (e.g., fear, disgust) were withdrawal motivating. To understand whether these asymmetrical frontal cortical activations resulted from affective valence or motivational direction (approach vs. withdrawal), we needed to examine an emotive state that avoided this confound of valence and motivational direction. To do so, we began investigating the relationship of anger with asymmetrical frontal cortical activity, because past social psychological and animal behavior research suggested that anger is a negative emotion that evokes approach

Portions of the research described within this chapter were supported by a grant from the National Science Foundation (BCS 0350435) and by a grant from the National Institute of Mental Health (R03 MH60747-01).

motivational action tendencies. If asymmetrical frontal cortical activity relates to motivational direction, then *anger should relate to greater left than right frontal activity*, because anger is associated with *approach motivational direction*. On the other hand, if asymmetrical frontal cortical activity relates to affective valence, then *anger should relate to greater right than left frontal activity*, because anger is associated with *negative valence*.

By investigating the relationship of anger with asymmetrical frontal cortical activity, we were in a position to gain a more complete understanding of the psychological and behavioral functions of asymmetrical frontal cortical activity. In addition, basic research on anger and its underlying neural systems can provide insights useful for understanding the relationship of motivational direction and affective valence. Most contemporary theories of emotion assume that positive affects are only related to approach motivation, whereas negative affects are only related to withdrawal motivation (Lang, 1995; Watson, 2000). By exploring anger, we will be in a position to better understand how these two important dimensions are related to each other. Finally, by understanding basic processes involved in anger, we as a society should be in a better position to explain, predict, treat, and control anger when necessary.

> “But anger is problematic above all other negative affects for its social consequences… my anger … threatens violence for you, your family, your friends, and above all for our society. Of all the negative affects it is the least likely to remain under the skin of the one who feels it, and so it is just that affect all societies try hardest to contain within that envelope under the skin …” (Tomkins, 1991, p. 111).

Neuro-imaging methods primer

Much of the research on asymmetrical frontal cortical activity and emotion has used EEG, particularly power (microvolts squared) in the alpha frequency band. The raw EEG signal is a complex waveform that can be decomposed using fast Fourier transforms (FFTs). That is, from the FFT, several frequency bands can be extracted from the raw EEG. Alpha is the frequency range from 8 to 13 Hz (cycles per second). Past research has suggested that alpha power is inversely related to cortical activation using a variety of other measures of cortical activation (Lindsley & Wicke, 1974), such as PET (Cook, O’Hara, Uijtdehaage, Mandelkern, & Leuchter, 1998) and fMRI (Goldman, Stern, Engel, & Cohen, 2002). Although EEG alpha power is inversely correlated with PET and fMRI measures, it may assess different aspects of brain activity (e.g., pre- vs. postsynaptic potentials).

Ultimately, both PET and fMRI rely on blood flow to brain areas recently involved in neuronal activity, although other changes also affect fMRI such as oxygen consumption and blood volume changes. Because both PET and fMRI measure blood flow rather than neuronal activity, the activations are not in real time with neuronal activations; rather, they are blood responses to neuronal responses. Thus, there is a biological limit on the time resolution of the response, such that even in the best measurement systems, the peak blood flow response occurs 6 to 9 seconds after stimulus onset (Reiman, Lane, van Petten, & Bandettini, 2000). However, there are suggestions that experimental methods can be designed to detect stimulus condition differences as early as 2 seconds (Bellgowan, Saad, & Bandettini, 2003). As a consequence, the limitation with fMRI and PET is biological. In contrast, EEG measures electrical activations instantaneously, at sub-millisecond resolution.

The spatial resolution of EEG, the ability to locate which specific areas of the brain generate the signals recorded, is currently not as good as spatial resolution with PET and fMRI. Much work is being conducted to achieve mathematical solutions to this problem, allowing for EEG to have better spatial resolution (e.g., Dien, Spencer, & Donchin, 2003; Pascual-Marqui et al., 1999). EEG research is also much less costly than fMRI and PET research. Finally, PET and EEG permit measurement of tonic (e.g., resting, baseline) activity as well as phasic (e.g., in response to a state manipulation) activity, whereas fMRI permits measurement of phasic but not tonic activity.

Scalp-recorded electrical activity is the result of activity of populations of neurons. The activity can be recorded on the scalp surface because the tissue between the neurons and the scalp acts as a volume conductor. Because the activity generated by one neuron is small, it is thought that the activity recorded at the scalp is the integrated activity of numerous neurons that are active synchronously. Moreover, for activity to be recorded at the scalp, the electric fields generated by each neuron must be oriented in such a way that their effects cumulate. That is, the neurons must be arranged in an open as opposed to closed field. In an open field, the neurons' dendrites are all oriented on one side of the structure, whereas their axons all depart from the other side. Open fields are present where neurons are organized in layers, as in most of the cortex, parts of the thalamus, the cerebellum, and other structures. Because of the need for summation of electrical potentials, the EEG activity is most likely the result of postsynaptic potentials, which have a slower time-course and are more likely to be synchronous and summate than presynaptic potentials.

Testing competing hypotheses: motivational direction versus emotional valence

In 1997, two independent groups observed that trait approach motivation was related to greater left than right frontal activity at resting baseline (Harmon-Jones & Allen, 1997; Sutton & Davidson, 1997). Trait approach motivation was assessed using Carver and White's (1994) behavioral activation and behavioral inhibition scale. The scale was based on Gray's (1987) theory of motivation, which posits that a behavioral activation system (BAS) and behavioral inhibition system (BIS) motivate and guide behavior. In Gray's theory, the BAS is a motivational system that is sensitive to signals of conditioned reward, nonpunishment, and escape from punishment. Its activation causes movement toward goals. The BIS is hypothesized to be sensitive to signals of conditioned punishment, nonreward, novelty, and innate fear stimuli. The BIS inhibits behavior, increases arousal, prepares for vigorous action, and increases attention toward aversive stimuli. Carver and White's (1994) BIS/BAS questionnaire assesses individual differences in BIS and BAS sensitivity. Sample items from the BIS scale include: "I worry about making mistakes," and "I have very few fears compared to my friends (reverse scored)." Sample items from the BAS include: "It would excite me to win a contest"; "I go out of my way to get things I want"; and "I crave excitement and new sensations".

Soon after observing the relationship between trait approach motivation and relative left frontal cortical activity, we noticed that all past studies on asymmetrical frontal cortical activity and emotion had confounded emotional valence (positive, negative affect) with motivational direction (approach, withdrawal motivation). Researchers were claiming that relatively greater left than right frontal cortical activity reflected greater approach motivation and positive affect, whereas relatively greater right than left frontal cortical activity reflected greater withdrawal motivation and negative affect. These claims fit well into dominant emotion theories that associated positive affect with approach motivation and negative affect with withdrawal motivation (Lang, 1995; Watson, 2000).

However, other, older theories suggested that approach motivation and positive affect are not always associated with one another. Anger, for example, is a negatively valenced emotion that evokes behavioral tendencies of approach (e.g., Darwin, 1872; Ekman & Friesen, 1975; Plutchik, 1980; Young, 1943). For example, anger is associated with attack—particularly offensive aggression (e.g., Berkowitz, 1993; Blanchard & Blanchard, 1984). Offensive aggression, associated with anger, can be distinguished from defensive aggression, associated with fear. Offensive aggression leads to attack without attempts to escape, whereas defensive or fear-based aggression leads to attack only if escape is not possible. In demonstrating that organisms evidence offensive aggression and that this is an approach behavior, Lagerspetz (1969) found that under certain conditions mice would cross an electrified grid to attack another mouse.

Lewis et al. (1990; Lewis, Sullivan, Ramsay, & Alessandri, 1992) conditioned infants to pull a string to receive a reward. They found that infants who displayed anger when the reward was withdrawn demonstrated the highest levels of joy, interest, and required arm pull when the learning portion of the task was reinstated. These results suggest that subsequent to frustrating events, anger may maintain and increase task engagement and approach motivation.

Additional support for the idea that anger is associated with approach motivation comes from research testing the conceptual model that integrated reactance theory with learned helplessness theory (Wortman & Brehm, 1975). According to this model, how individuals respond to uncontrollable outcomes depends on their expectation of being able to control the outcome and the importance of the outcome. When an individual expects to be able to control outcomes that are important, and those outcomes are found to be uncontrollable, psychological reactance should be aroused. Thus, for individuals who initially expect control, the first few bouts of uncontrollable outcomes should arouse reactance, a motivational state aimed at restoring control. After several exposures to uncontrollable outcomes, these individuals should become convinced that they cannot control the outcomes and should show decreased motivation (i.e., learned helplessness). In other words, reactance will precede helplessness for individuals who initially expect control. In one study testing this model, individuals who exhibited angry feelings in response to one unsolvable problem had better performance and were presumably more approach motivated on a subsequent cognitive task than did participants who exhibited less anger (Mikulincer, 1988).

Other research has revealed that state anger relates to high levels of self-assurance, physical strength, and bravery (Izard, 1991), inclinations associated with approach motivation. Additionally, Lerner and Keltner (2001) found that anger (both trait and state) is associated with optimistic expectations, whereas fear is associated with pessimistic expectations. Moreover, happiness was associated with optimism, making anger and happiness appear more similar to each other in their relationship with optimism than fear and anger. Although Lerner and Keltner (2001) interpreted their findings as being the result of the appraisals associated with anger, it seems equally plausible that it was the approach motivational character of anger that caused the relationship of anger and optimism. That is, anger creates optimism because anger engages the approach motivational system, which produces greater optimistic expectations.

Other evidence supporting the idea that anger is associated with an approach-orientation comes from research on bipolar disorder. The emotions of euphoria and anger often occur during manic phases of bipolar disorder (Cassidy, Forest, Murry, & Carroll, 1998; Depue & Iacono, 1989; Tyrer & Shopsin, 1982). Both euphoria and anger may be approach-oriented processes, and a dysregulated or hyperactive approach system may underlie mania (Depue & Iacono, 1989; Fowles, 1993). Research suggests that hypomania/mania involves increased left frontal brain activity and approach motivational tendencies. In this research, it has been found that individuals who have suffered damage to the right frontal cortex are more likely to evidence mania (*see* review by Robinson & Downhill, 1995). Thus, this research is consistent with the view that mania may be associated with increased left frontal activity and increased approach tendencies, because the approach motivation functions of the left frontal cortex are released and not restrained by the withdrawal system in the right frontal cortex. Furthermore, lithium carbonate, a treatment for bipolar disorder, reduces aggression (Malone, Delaney, Luebbert, Cater, & Campbell, 2000), suggesting that anger and aggression correlate with the other symptoms of bipolar disorder. In addition, trait anger has been found to relate to high levels of assertiveness and competitiveness (Buss & Perry, 1992).

Other studies have associated anger with trait approach motivation or, more specifically, trait behavioral approach or BAS. In two studies, trait BAS, as assessed by Carver and White's (1994) scale, was positively related to trait anger at the simple correlation level, as assessed by the Buss and Perry (1992) aggression questionnaire (Harmon-Jones, 2003). Carver (2004) has also found that trait BAS predicts state anger in response to situational anger manipulations. These results support the hypothesis that anger is related to approach motivation.

Because of the large body of evidence suggesting that anger is often associated with approach motivation, my colleagues and I examined the relationship between anger and relative left frontal activation to test whether the frontal asymmetry results from emotional valence, motivational direction, or a combination of emotional valence and motivational direction.

Asymmetrical frontal cortical activity and anger

Because much past research from a variety of empirical approaches suggests that anger is associated with approach motivational tendencies, we proposed that by assessing the relationship of anger and asymmetrical frontal cortical activity, we would be better able to determine whether asymmetrical frontal cortical activity related to motivational direction or affective valence. If asymmetrical frontal cortical activity relates to motivational direction, then anger should relate to greater left than right frontal activity, because anger is associated with approach motivational direction. In contrast, if asymmetrical frontal cortical activity relates to affective valence, then anger should relate to greater right than left frontal activity, because anger is associated with negative valence.

Trait Anger

In one of the first studies testing these competing predictions, Harmon-Jones and Allen (1998) assessed trait anger using the Buss and Perry (1992) questionnaire and assessed asymmetrical frontal activity by examining baseline, resting regional EEG activity (alpha power) in a 4-minute period. In this study of adolescents, trait anger related to increased left frontal activity and decreased right frontal activity. In addition, a subset of this sample was comprised of adolescents in a psychiatric in-patient unit for impulsive aggression. Even among these individuals, trait anger related positively with greater left than right frontal activity. Asymmetrical activity in other regions did not relate with anger. The specificity of anger to frontal asymmetries and not other region asymmetries has been observed in all of our studies. Thus, we focus our review on asymmetrical frontal activity.

Other research addressed an alternative explanation for the observation that relative left frontal activity related to anger (Harmon-Jones, 2004). The alternative explanation suggested that persons with high levels of trait anger might experience anger as a positive emotion, and this positive feeling or attitude toward anger could be responsible for anger being associated with relative left frontal activity. After developing a valid and reliable assessment of attitude toward anger, a study was conducted to assess whether resting baseline asymmetrical activity related to trait anger and attitude toward anger. Results indicated that anger related to relative left frontal activity and not attitude toward anger. Moreover, further analyses revealed that the relationship between trait anger and left frontal activity did not result from anger being associated with a positive attitude toward anger.

State Anger

To address the limitations inherent in correlational studies, experiments have been conducted in which anger is manipulated and its effects on regional brain activity are examined. In Harmon-Jones and Sigelman (2001), participants were randomly assigned to a condition in which another person insulted them or to a condition in which another person treated them in a neutral manner. Immediately following the treatment, EEG was collected. As predicted, individuals who were insulted evidenced greater relative left frontal activity than individuals who were not insulted. Additional analyses revealed that within the insult condition, reported anger and aggression were positively correlated with relative left frontal activity. Neither of these correlations was significant in the no-insult condition. These results suggest that relative left frontal activation was associated with more anger and aggression in the condition in which anger was evoked.

More recent experimental evidence has replicated these results and also revealed that state anger evokes both increased left and decreased right frontal activity. In addition, when participants were first induced to feel sympathy for a person who insulted them, this reduced the effects of insult on left and right frontal activity (Harmon-Jones, Vaughn-Scott, Mohr, Sigelman, & Harmon-Jones, 2004). This suggests that the reason experiencing sympathy for another individual reduces aggression toward that individual (e.g., *see* review by Miller & Eisenberg, 1988) may be because sympathy reduces the relative left frontal activity associated with approach-oriented anger.

Independent Manipulation of Approach Motivation Within Anger

In the experiments just described, the designs were tailored in such a way as to evoke anger that was approach-oriented. Although most instances of anger involve approach inclinations, it is possible that not all forms of anger are associated with approach motivation. To manipulate approach motivation independently of anger, Harmon-Jones, Sigelman, Bohlig, and Harmon-Jones (2003) performed an experiment in which the ability to cope with the anger-producing event was manipulated. Based on past research that has revealed that coping potential affects motivational intensity (Brehm & Self, 1996), it was predicted that the expectation of being able to take action to resolve the anger-producing event would increase approach motivational intensity relative to expecting to be unable to take action.

Participants, who were strongly opposed to a tuition increase, were angered by a radio editorial that argued in favor of a 10% tuition increase at their university. To manipulate coping potential or the expectation of acting to change the situation, two conditions differed with regard as to whether it was possible for participants to act to change the event that caused the anger. One condition was led to believe that the tuition increase might not occur and that petitions were being circulated to stop it (action possible condition); the other condition was led to believe that the university administration had already voted in favor of implementing the tuition increase and nothing could be done to change that (action impossible condition). Both conditions evoked significant increases in anger (over baseline) and they were not significantly different from each other. More importantly and consistent with predictions, results indicated that participants who expected to engage in the approach-related action evidenced greater left frontal activity than participants who expected to be unable to engage in approach-related action. Moreover, within the action-possible condition, participants who evidenced greater left frontal activity in response to the angering event also evidenced greater self-reported anger, providing support for the idea that anger is often an approach-related emotional response. In the condition where action was not possible, greater left frontal activity did not relate to greater anger. In our view, this is because, although anger usually leads to approach motivation, when action is not possible, approach motivation remains low, even if angry feelings are high. Finally, within the action-possible condition, participants who evidenced greater left frontal activity in response to the event were more likely to engage in behaviors that would reduce the possibility of the angering event from occurring in the future (i.e., they were more likely to sign a petition to prevent the tuition increase and to take petitions with them for others to sign). This finding suggests that greater approach motivation, as reflected in greater left frontal cortical activity, was associated with more action to correct the negative situation.

The research of Harmon-Jones et al. (2003) suggests that the left frontal region is most accurately described as a region sensitive to approach motivational intensity. That is, it was only when anger was associated with an opportunity to behave in a manner to resolve the anger-producing event that participants evidenced the increased relative left frontal activation. The effect of approach motivation and anger on left frontal activity has recently been produced using pictorial stimuli that evoke anger

(Harmon-Jones, Lueck, Fearn, & Harmon-Jones, 2006). In this experiment, participants low in racial prejudice were shown neutral, positive, and fear/disgust pictures from the International Affective Picture System (Lang, Bradley, & Cuthbert, 2005). Mixed among those pictures were pictures depicting instances of racism and hatred (e.g., neo-Nazis, Ku Klux Klan). Prior to viewing the pictures, half of the participants were informed that they would write an essay on why racism is immoral, unjust, and unfair at the end of the experiment. This manipulation served to increase their anger-related approach motivation. Results revealed that participants showed greater relative left frontal activity to anger pictures than other picture types only when they expected to engage in approach-related behavior. A second study revealed that individuals who scored lower in racial prejudice evidenced even greater relative left frontal activation to the anger-evoking racist pictures in the approach motivation condition.

The above findings may suggest that relatively greater left frontal activity will occur in response to an angering situation only when there is an explicit approach motivational opportunity. However, it is possible that an explicit approach motivational opportunity is not necessary for increased left frontal activity to anger to occur but that it only intensifies left frontal activity. In other words, there may be other features of the situation or person that make it likely that an angering situation will increase approach motivational tendencies and activity in the left frontal cortical region. One possibility along these lines is the personality characteristic of trait anger—that is, individuals who are chronically high in anger may evidence increased left frontal activity (and approach motivational tendencies) in response to angering situations that would not necessarily cause such responses in individuals who are not as chronically angry. This prediction is predicated on the idea that angry individuals have more extensive angry associative networks than less angry individuals and that anger-evoking stimuli should therefore activate parts of the network more readily in these angry individuals (Berkowitz, 1993). In other words, among individuals high in trait anger, even mild anger cues might activate parts of the anger network and, through established associations, lead to angry expressive-motor responses, physiological reactions, feelings, thoughts, and memories.

Along the lines suggested by the cognitive-neo-associative model of aggression (e.g., Berkowitz, 1993), research has revealed that participants high in trait anger show selective perceptual and cognitive biases toward angry words and facial expressions in Stroop-type and visual search tasks (Cohen, Eckhardt, & Schagat, 1998; Eckhardt & Cohen, 1997; van Honk et al., 2001). However, no previous research has tested whether anger-evoking stimuli are more likely to activate neural structures involved in approach motivational tendencies in individuals who are high as compared to low in trait anger. Such results would extend our knowledge of the neural circuitry underlying angry individuals' enhanced likelihood of engaging in angry responses. Therefore, we predicted that individuals high in trait anger would show relatively greater left frontal cortical activation to mild anger cues even when explicit approach motivation opportunities were not made salient.

In this study, participants were exposed to anger-inducing pictures (and other pictures) and given no explicit manipulations of action expectancy. Across all participants, a null effect of relative left frontal asymmetry occurred. However, individual differences in trait anger related to relative left frontal activity to the anger-inducing pictures, such that individuals high in trait anger showed greater left frontal activity to anger-producing pictures (controlling for activity to neutral pictures; Harmon-Jones, 2007). These results suggest that the explicit manipulation or opportunity for approach motivated action may potentiate the effects of approach motivation on relative left frontal activity but may not always be necessary.

Manipulation of Frontal Cortical Activity and Anger Processing

Other research is consistent with the hypothesis that anger is associated with left frontal activity. For example, d'Alfonso et al. (2000) used slow rTMS to inhibit the left or right prefrontal cortex (PFC). Slow rTMS produces inhibition of cortical excitability, so that rTMS applied to the right PFC decreases its activation and causes the left PFC to become more active, whereas rTMS applied to the left PFC causes activation of the right PFC. They found that rTMS applied to the right PFC caused selective attention toward angry faces, whereas rTMS applied to the left PFC caused selective attention away from angry faces. Thus, an increase in left prefrontal activity led participants to attentionally approach angry faces, as in an aggressive confrontation. In contrast, an increase in right prefrontal activity led participants to attentionally avoid angry faces, as in a fear-based avoidance. Conceptually similar results have been found by van Honk and Schutter (2006). The interpretation of these results, which these researchers advanced, concurs with other research that has demonstrated that attention toward angry faces is associated with high levels of self-reported anger and that attention away from angry faces is associated with high levels of cortisol, which is associated with fear (van Honk, Tuiten, de Haan, van den Hout, & Stam, 2001; Van Honk, Tuiten, Van den Hout, Koppeschaar, Thijssen, & de Haan, 1998; van Honk et al., 1999).

We recently extended the work of van Honk and colleagues by examining whether a manipulation of asymmetrical frontal cortical activity would affect behavioral aggression. Based on past research showing that contraction of the left hand increases right frontal cortical activity and that contraction of the right hand increases left frontal cortical activity (Harmon-Jones, 2006), we manipulated asymmetrical frontal cortical activity by having participants contract their right or left hand. Participants then received insulting feedback ostensibly from another participant. They then played a reaction time game on the computer against the other ostensible participant. Participants were told they could give the other participant a blast of 60dB, 70dB, 80dB, 90dB, or 100dB white noise for up to 10 seconds if they were fastest to press the shift key when an image appeared on the screen. Results indicated that participants who squeezed with their right hand gave significantly louder and longer noise blasts to the other ostensible participant than those who squeezed with their left hand (Peterson, Shackman, & Harmon-Jones, 2008).

Research on anger using other brain imaging methods

The reviewed research has revealed that the left frontal cortical region is involved in approach motivated anger. A few studies on anger using brain imaging technologies other than EEG have been conducted. In one, PET (oxygen-15 -labeled carbon dioxide) was measured while men were exposed to mental imagery scripts concerning angering or neutral events that occurred in their own lives. Results revealed that as compared to neutral imagery, anger imagery caused an increase in the left orbital frontal cortex, the right anterior cingulate cortex, the bilateral anterior temporal poles, left precentral gyrus, bilateral medial frontal cortex, and bilateral cerebellum. Dougherty et al. (1999) interpreted the increase in left orbital frontal cortical activity as corresponding "to inhibition of aggressive behavior in the face of anger." (p. 471). Although this interpretation is consistent with some speculations of the role of the left orbital frontal cortex in response inhibition (Mega et al., 1997), it is inconsistent with the EEG results showing that increased left frontal activity is associated with increased aggression and approach behavior (e.g., Harmon-Jones & Sigelman, 2001; Harmon-Jones et al., 2003). The interpretation that the left frontal cortical region is involved in the inhibition of anger and aggression is also inconsistent with lesion data suggesting that mania results from damage to the right frontal region (e.g., Robinson & Downhill, 1995) and results obtained when the left relative to right frontal cortex is activated and angry attentional processes are measured (e.g., d'Alfonso et al., 2000). However, EEG is likely assessing dorsolateral frontal cortical activity and not orbital frontal activity, and left orbital frontal activity may be involved in the inhibition of anger, whereas left dorsolateral frontal activity may be involved in approach motivations like anger. The study by Dougherty et al. did not find an increase in left dorsolateral frontal activity but their manipulation may not have evoked approach-motivated anger.

Of course, anger induced realistically by insulting feedback or goal blocking, as in the EEG experiments, may have very different properties from imagined anger, as produced by imagery. In the imagery experiments, correlations between reported anger and regional brain activity were not reported, whereas in the EEG experiments, self-reported anger has been found to correlate significantly with relative left frontal activity. Examination of correlations between reported emotion and physiological measures assists in determining whether the brain activation is related to emotional experience or some other nonemotional variable.

Comparison of Neural Research on Anger to Neural Research on Violence

The results indicating that relatively greater left frontal cortical activity is associated with increased approach-oriented anger and behavior are seemingly inconsistent with evidence suggesting that violent individuals have reduced frontal lobe function (e.g., Amen et al., 1996; Raine, Stoddard, Bihrle, & Buchsbaum, 1998; Raine et al., 1998). These results, and others, have led some to conclude that the PFC—particularly the left orbital frontal cortex (and frontal activations derived from EEG)—is involved in the regulation of negative affects like anger (Davidson, Putnam, & Larson, 2000).

However, other research has more strongly implicated reduced activity in the right frontal regions in violence (Raine et al., 1998; Raine et al., 2001). These findings are consistent with the reviewed EEG research, which found that approach-oriented anger was associated with reduced right frontal activity (in addition to increased left frontal activity). Other research has revealed that increased activity in the right frontal cortex is associated with withdrawal motivation. Perhaps the violent individuals who showed reduced right frontal activity in the studies of Raine and colleagues (1998, 2001) lack the behavioral constraints engendered by the withdrawal motivation system that may be partially instantiated in the right frontal cortex.

However, other studies have implicated reduced activity in both left and right frontal cortices in violence and aggression. For example, Raine et al. (1998) found in a PET study that affective murderers, as compared to predatory murderers and normal controls, evidenced reduced lateral and medial prefrontal activity in both hemispheres during a continuous performance task. Of course, differences in participant samples may explain the differences in results—that is, most of the anger studies have involved normal individuals, whereas the studies on violence have involved extremely violent individuals. However, in the Harmon-Jones and Allen (1998) study, some participants were adolescents who were in an in-patient psychiatric unit for impulse control disorders. Even among this sample, trait anger was related to greater left frontal activity and reduced right frontal activity at rest (for similar results, *see* Rybak, Crayton, Young, Herba, & Konopka, 2006). These results suggest that the samples may not be the source of the differences. However, the Harmon-Jones and Allen (1998) sample was much younger than samples used in other studies comparing frontal lobe function in violent and nonviolent individuals (e.g., Raine et al., 1998). The longer lifetime of violence in adults may relate to the reductions in frontal cortical activity.

The studies comparing brain function of violent and nonviolent individuals typically assess frontal lobe function at rest or during a cognitive task and not during anger-arousing situations. However, it is not clear that reduced frontal lobe function measured during these tasks causes violence. The violent individuals (as compared to the nonviolent individuals) who display less prefrontal activity during cognitive tasks may simply be less emotively engaged by the relatively unemotional nature of the tasks. The frontal lobe function of violent individuals during a more emotively engaging situation, such as an interpersonal provocation, has not been assessed. It is possible that reduced frontal lobe function might not be seen in these individuals in such a situation. Brain activations during anger-inducing events may be more predictive of violent behavior than brain activations during nonmotivating cognitive tasks. Another difficulty emerges when attempting to compare studies of violent offenders with the studies on anger: anger can be manipulated in the lab, whereas being a violent offender cannot be manipulated. The latter correlational studies are difficult to interpret.

In summary, the idea that anger is associated with approach motivational tendencies is supported by behavioral and neuro-imaging evidence. However, it is possible that some instances of anger, such as anger mixed with fear, may be associated with withdrawal motivational tendencies. Indeed, in one study, we observed such an effect (Zinner, Brodish, Devine, & Harmon-Jones, 2008). In the study, White individuals prepared to interact with a Black person, under the guise of an interest in exploring interracial interactions. In such a context, societal pressure dictates that anger should not be expressed. Thus, in this situation, the experience of anger may coincide with anxiety and a desire to avoid the situation. Cortical activity was measured while White participants anticipated the interracial interaction. Consistent with expectations, self-reported anger was associated with anxiety and relative right frontal cortical activity. In addition, some individuals may have learned to control their angry approach tendencies and may have instead converted these angry tendencies into withdrawal-oriented behaviors (e.g., Hewig, Hagemann, Seifert, Naumann, & Bartussek, 2004). More research is needed to understand whether and how this type of angry expression may emerge.

Discussion

Of course, approach motivations such as anger involve several brain regions, but the reviewed research establishes the importance of the left PFC in approach motivation independent of affective valence. Often in discussions of the functions of the PFC, scientists suggest that the PFC is involved in higher-level cognitive functions, such as working memory and inhibitory processes. Part of the reason scientists reserve the PFC for higher-level cognitive processes is because it is a region that is much larger in humans than nonhuman animals. The logic continues that if the PFC were a relatively recent development in evolution, then it must be the source of those psychological processes that separate us from other animals. This logic is likely at least partially correct but not foolproof. For example, recent single-cell research with rats has revealed that the PFC is involved in aggression and most of the cells activated are not inhibitory cells (Halász, Tóth, Kalló, Liposits, & Haller, 2006). The PFC is a vast territory and is likely involved in a number of psychological processes. Moreover, structures that are involved in certain psychological/behavioral processes in nonhuman animals may be involved in different processes in human animals. For example, many of the anatomical details of components of emotional response circuits are different in rodents and primates. The organization, connectivity, and some functions of amygdala nuclei (Amaral, Price, Pitkanen, & Carmichael, 1992), PFC (Goldman-Rakic, 1987), and anterior cingulate (Bush, Luu, & Posner, 2000) differ between rodents and primates. In addition, evidence suggests that areas throughout the brain are activated during a variety of mental processes, rather than processes being localized in just one brain area. The size, complexity, and activity of the human PFC suggest that it is integrated in many processes.

Humans are better able to plan behavior and control their responses to emotional stimuli than other animals. No doubt the PFC is involved in these processes. However, this planning and execution of behavior is not always in the service of inhibiting destructive motivations. In fact, some behaviors that are said to distinguish humans from other mammals, such as war and genocide, involve planning and control but actually enhance the destructiveness of approach-oriented aggressive motivation.

Indeed, the research on anger and asymmetrical frontal cortical activity, when considered in whole, strongly suggests that the left PFC region is involved in more than inhibition of negative affect, as some have suggested (Davidson, Putnam, & Larson, 2000). That is, relative left frontal activation has been associated with self-reported state anger and behavioral aggression (Harmon-Jones & Sigelman, 2001) and approach-motivated behavior (Harmon-Jones et al., 2003). Individuals with proneness toward mania (Harmon-Jones et al., 2002) and individuals higher in trait anger (Harmon-Jones, 2007) show even greater relative left frontal activation in response to angering events. Moreover, manipulated increases in left frontal activation cause approach-related angry attentional and memory responses (d'Alfonso et al., 2000; van Honk & Schutter, 2006). Finally, even at resting baseline, individuals who are higher in trait anger show greater relative left frontal activity (Hewig et al., 2004; Harmon-Jones & Allen, 1998; Rybak et al., 2006), and this relationship also occurs in adolescents who are in psychiatric in-patient units for impulse control disorders (Harmon-Jones & Allen, 1998; Rybak et al., 2006). It would be illogical to suggest that all of these individuals are inhibiting anger more than individuals without high levels of state anger, trait anger, approach behavior, aggression, or mania.

The approach and withdrawal processes implemented by asymmetrical frontal cortices have been observed in rhesus monkeys (e.g., Kalin, Shelton, Davidson, & Kelley, 2001) and humans as early as 2 to 3 days of age (Fox & Davidson, 1986). In addition, damage to these regions of frontal cortex cause depression versus mania (Robinson & Downhill, 1995), and rTMS manipulations of left versus right cortical regions affects mood and attentional processing in manners consistent with the idea that asymmetrical frontal cortical activity is involved in motivational direction (d'Alfonso et al., 2000; van Honk & Schutter, 2006). Finally, research with organisms as simple as toads has revealed that approach and withdrawal processes are lateralized in a manner similar to that observed in humans (Vallortigara & Rogers, 2005). However, these lateralizations probably involve more structures than the frontal cortex, as amphibians lack such. It is possible that subcortical structures are lateralized for approach and withdrawal motivational processes in amphibians, reptiles, and birds but that these lateralizations are preserved and elaborated into the frontal cortices of primates. Future research will need to explore connections between subcortical and cortical structures in approach and withdrawal motivation. Along these lines, some research suggests activations in the left frontal cortex are related to dopaminergic projections from the striatum associated with the coordination of action with learned reward contingencies (Berridge, Espana, & Stalnacher, 2003). However, it is unlikely that the motivational-related activations observed in the frontal cortices simply result from "propagation" of signals from solely subcortical structures, as source-localization analyses have suggested that approach-withdrawal-related frontal asymmetries reflect changes in dorsolateral prefrontal cortical activity (Pizzagalli et al., 2005).

In conclusion, research suggests that greater left than right frontal cortical activity is associated with approach motivation and not positive affect *per se.* In pursuing a better understanding of the role of asymmetrical frontal cortical activity in emotive processes, the research shed new light on social psychological questions outside the realm of social neuroscience. For example, the neuroscience question regarding the emotive functions of asymmetrical frontal cortical activity prompted a line of research on anger, one of the most socially important emotive states/traits, which has unfortunately been relatively neglected in most major theories of emotion. This research demonstrated that unlike other negative emotions, anger is often associated with approach motivational tendencies. Consequently, major dimensional theories of emotion will need to be modified to incorporate the idea that not all negative affects are associated with withdrawal motivation. Also, the research on anger suggested social situations and individual differences that may cause anger to be associated with withdrawal motivation. This work may have important implications for understanding the inhibition of aggressive behavior as well as the development and/or maintenance of anxiety disorders. Finally, the research on the emotive functions of asymmetrical frontal cortical activity has been extended to assist in understanding the psychological and behavioral functions of guilt (Amodio, Devine, & Harmon-Jones, 2007) as well as cognitive dissonance processes (Harmon-Jones, 2004; Harmon-Jones, Gerdjikov, & Harmon-Jones, 2008). Social neuroscience is an important area of social psychology that has the potential to enhance our understanding of basic social psychological issues and integrate the theories and findings of social psychology into other areas of neuroscientific inquiry.

References

Amaral, D. G., Price, J. L., Pitkanen, A., & Carmichael, S. T. (1992). Anatomical organization of the primate amygdaloid complex. In J. P. Aggleton (Ed.), *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction* (pp. 1–66). New York: Wiley–Liss.

Amen, D. G., Stubblefield, M., Carmichael, B., & Thisted, R. (1996). Brain SPECT findings and aggressiveness. *Annals of Clinical Psychiatry, 8*, 129–137.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2007). A dynamic model of guilt: Implications for motivation and self-regulation in the context of prejudice. *Psychological Science, 18*, 524–530.

Bellgowan, P. S., Saad, Z. S., & Bandettini, P. A. (2003). Understanding neural system dynamics through task modulation and measurement of functional MRI amplitude, latency, and width. *Proceedings National Academy Sciences USA, 100*(3), 1415–1419.

Berkowitz, L. (1993). *Aggression: Its Causes, Consequences, and Control*. Philadelphia, PA: Temple University Press.

Berkowitz, L. (1999). Anger. In T. Dalgleish, & M. J. Power (Eds.), *Handbook of Cognition and Emotion* (pp. 411–428). New York: John Wiley & Sons.

Berkowitz, L. (2000). *Studies in Emotion and Social Interaction*. New York: Cambridge University Press.

Berridge, C. W., España, R. A., & Stalnaker, T. A. (2003). Stress and coping: Asymmetry of dopamine efferents within the prefrontal cortex. In R. J. D. K. Hugdahl (Ed.), *The Asymmetrical Brain*. Cambridge, MA: MIT Press.

Brehm, J. W., & Self, E. (1989). The intensity of motivation. In M. R. Rosenzweig, & L. W. Porter (Eds.), *Annual Review of Psychology* (Vol. 40, pp. 109–131). Palo Alto, CA: Annual Reviews.

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences, 4*, 215–222.

Buss, A. H., & Perry, M. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology, 63*, 452–459.

Canli, T., Desmond, J. E., Zhao, Z., Glover, G., & Gabrieli, J. D. (1998). Hemispheric asymmetry for emotional stimuli detected with fMRI. *Neuroreport, 9*, 3233–3239.

Carver, C. S. (2004). Negative affects deriving from the behavioral approach system. *Emotion, 4*, 3–22.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology, 67*, 319–333.

Cassidy, F., Forest, K., Murry, E., & Carroll, B. J. (1998). A factor analysis of the signs and symptoms of mania. *Archives of General Psychiatry, 55*(1), 27–32.

Coan, J. A., & Allen, J. J. B. (2003). The state and trait nature of frontal EEG asymmetry in emotion. In K. Hugdahl, & R. J. Davidson (Eds.), *The Asymmetrical Brain* (pp. 565–615). Cambridge, MA: MIT Press.

Cohen, D. J., Eckhardt, C. I., & Schagat, K. D. (1998). Attention allocation and habituation to anger-related stimuli during a visual search task. *Aggressive Behavior, 24*(6), 399–409.

Cook, I. A., O'Hara, R., Uijtdehaage, S. H. J., Mandelkern, M., & Leuchter, A. F. (1998). Assessing the accuracy of topographic EEG mapping for determining local brain function. *Electroencephalography and Clinical Neurophysiology, 107*, 408–414.

Cunningham, W. A., Espinet, S. D., DeYoung, C. G., & Zelazo, P. D. (2005). Attitudes to the right- and left: Frontal ERP asymmetries associated with stimulus valence and processing goals. *NeuroImage, 28*, 827–834.

d'Alfonso, A. A. L., van Honk, J., Hermans, E., Postma, A., & de Haan, E. H. F. (2000). Laterality effects in selective attention to threat after repetitive transcranial magnetic stimulation at the prefrontal cortex in female subjects. *Neuroscience Letters, 280*, 195–198.

Darwin, C. (1872/1965). *The Expressions of the Emotions in Man and Animals*. New York: Oxford University Press.

Davidson, R. J., Putnam, K. M., & Larson, C. L. (2000). Dysfunction in the neural circuitry of emotion regulation—a possible prelude to violence. *Science, 289*, 591–594.

Depue, R. A., & Iacono, W. G. (1989). Neurobehavioral aspects of affective disorders. In M. R. Rosenzweig, & L. W. Porter (Eds.), *Annual Review of Psychology* (Vol. 40, pp. 457–492). Palo Alto, CA: Annual Reviews.

Dien, J., Spencer, K. M., & Donchin, E. (2003). Localization of the event-related potential novelty response as defined by principal components analysis. *Cognitive Brain Research, 17*, 637–650.

Dougherty, D. D., Shin, L. M., Alpert, N. M., et al. (1999). Anger in health men: A PET study using script-driven imagery. *Biological Psychiatry, 46*, 466–472.

Eckhardt, C. I., & Cohen, D. J. (1997). Attention to anger-relevant and irrelevant stimuli following naturalistic insult. *Personality and Individual Differences, 23*(4), 619–629.

Ekman, P., & Friesen, W. V. (1975). *Unmasking the Face: A Guide to Recognizing Emotions from Facial Cues*. Oxford: Prentice-Hall.

Fowles, D. C. (1993). Behavioral variables in psychopathology: A psychobiological perspective. In P. B. Sutker, & H. E. Adams (Eds.), *Comprehensive Handbook of Psychopathology* (2nd ed., pp. 57–82). New York: Plenum.

Fox, N. A., & Davidson, R. J. (1986). Taste-elicited changes in facial signs of emotion and the asymmetry of brain electrical activity in human newborns. *Neuropsychologia, 24*(3), 417–422.

Goldman-Rakic, P. S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In V. B. Mountcastle (Ed.), *Handbook of Physiology: 5* (pp. 373–417). Bethesda, MD: American Physiological Society.

Goldman, R. I., Stern, J. M., Engel, J. Jr., & Cohen, M. S. (2002). Simultaneous EEG and fMRI of the alpha rhythm. *Neuroreport, 13*, 2487–2492.

Gray, J. A. (1987). *The Psychology of Fear and Stress*. Cambridge: Cambridge University Press.

Halász, J., Tóth, M., Kalló, I., Liposits, Z., & Haller, J. (2006). The activation of prefrontal cortical neurons in aggression—a double labeling study. *Behavioral Brain Research, 175*, 166–175.

Harmon-Jones, E. (2003). Anger and the behavioural approach system. *Personality and Individual Differences, 35*, 995–1005.

Harmon-Jones, E. (2004). On the relationship of anterior brain activity and anger: Examining the role of attitude toward anger. *Cognition and Emotion, 18*, 337–361.

Harmon-Jones, E. (2007). Trait anger predicts relative left frontal cortical activation to anger-inducing stimuli. *International Journal of Psychophysiology, 66*, 154–160.

Harmon-Jones, E., & Allen, J. J. B. (1998). Anger and frontal brain activity: EEG asymmetry consistent with approach motivation despite negative affective valence. *Journal of Personality and Social Psychology, 74*, 1310–1316.

Harmon-Jones, E., & Allen, J. J. B. (1997). Behavioral activation sensitivity and resting frontal EEG asymmetry: Covariation of putative indicators related to risk for mood disorders. *Journal of Abnormal Psychology, 106*, 159–163.

Harmon-Jones, E., Gerdjikov. T., & Harmon-Jones, C. (2008).The effect of induced compliance on relative left frontal cortical activity: A test of the action-based model of dissonance. *European Journal of Social Psychology, 38*, 35–45.

Harmon-Jones, E., Lueck, L., Fearn, M., & Harmon-Jones, C. (2006). The effect of personal relevance and approach-related action expectation on relative left frontal cortical activity. *Psychological Science, 17*, 434–440.

Harmon-Jones, E., Sigelman J. D., Bohlig A., & Harmon-Jones, C. (2003). Anger, coping, and frontal cortical activity: The effect of coping potential on anger-induced left frontal activity. *Cognition and Emotion, 17*, 1–24.

Harmon-Jones, E., Vaughn-Scott, K., Mohr, S., Sigelman, J., & Harmon-Jones, C. (2004). The effect of manipulated sympathy and anger on left and right frontal cortical activity. *Emotion, 4*, 95–101.

Harmon-Jones, E., & Sigelman, J. (2001). State anger and prefrontal brain activity: Evidence that insult-related relative left prefrontal activation is associated with experienced anger and aggression. *Journal of Personality and Social Psychology, 80*, 797–803.

Hewig, J., Hagemann, D., Seifert, J., Naumann, E., & Bartussek, D. (2004). On the selective relation of frontal cortical asymmetry and anger-out versus anger-control. *Journal of Personality and Social Psychology, 87*, 926–939.

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

Kalin, N. H., Shelton, S. E., Davidson, R. J., & Kelley, A. E. (2001). The primate amygdala mediates acute fear but not the behavioral and physiological components of anxious temperament. *Journal of Neuroscience, 21*, 2067–2074.

Lagerspetz, K. M. J. (1969). Aggression and aggressiveness in laboratory mice. In S. Garattini, & E. B. Sigg (Eds.), *Aggressive Behavior* (pp. 77–85). New York: Wiley.

Lang, P. J. (1995). The emotion probe. *American Psychologist, 50*, 372–385.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2005). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. *Technical Report A–6, University of Florida, Gainesville, FL*.

Lerner, J. S., & Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology, 81*, 146–159.

Lewis, M., Alessandri, S. M., & Sullivan, M. W. (1990). Violation of expectancy, loss of control, and anger expressions in young infants. *Developmental Psychology, 26*(5), 745–751.

Lewis, M., Sullivan, M. W., Ramsay, D. S., & Alessandri, S. M. (1992). Individual and anger and sad expressions during extinction: Antecedents and consequences. *Infant Behavior & Development, 15*(4), 443–452.

Lindsley, D. B., & Wicke, J. D. (1974). The electroencephalogram: Autonomous electrical activity in man and animals. In I. R. T. M. N. Patterson (Ed.), *Bioelectric Recording Techniques* (pp. 3–79). New York: Academic Press.

Malone, R. P., Delaney, M. A., Leubbert, J. F., Cater, J., & Campbell, M. (2000). A double-blind placebo-controlled study of lithium in hospitalized aggressive children and adolescents with conduct disorder. *Archives of General Psychiatry, 57*(7), 649–654.

Mega, M. S., Cummings, J. L., Salloway, S., & Malloy, P. (1997). The limbic system: An anatomic, phylogenetic, and clinical perspective. *Journal of Neuropsychiatry and Clinical Neuroscience, 9*, 315–330.

Mikulincer, M. (1988). Reactance and helplessness following exposure to unsolvable problems: The effects of attributional style. *Journal of Personality and Social Psychology, 54*, 679–686.

Pascual–Marqui, R. D., Lehmann, D., Koenig, T., et al. (1999). Low resolution brain electromagnetic tomography (LORETA) functional imaging in acute, neuroleptic-naive, first-episode, productive schizophrenia. *Psychiatry Research: Neuroimaging, 90*, 169–179.

Peterson, C. K., Shackman, A. J., & Harmon-Jones, E. (2008). The role of asymmetrical frontal cortical activity in aggression. *Psychophysiology, 45*, 86-92..

Plutchik, R. (1980). *Emotion: A Psychoevolutionary Synthesis*. New York: Harpercollins College Division.

Raine, A., Meloy, J. R., Bihrle, S., Stoddard, J., LaCasse, L., & Buchsbaum, M. S. (1998). Reduced prefrontal and increased subcortical brain functioning assessed using positron emission tomography in predatory and affective murderers. *Behavioral Sciences and the Law, 16*, 319–332.

Raine, A., Park, S., Lencz, T., et al. (2001). Reduced right hemisphere activation in severely abused violent offenders during a working memory task: An fMRI study. *Aggressive Behavior, 27*, 111–129.

Raine, A., Stoddard, J., Bihrle, S., & Buchsbaum, M. (1998). Prefrontal glucose deficits in murderers lacking psychosocial deprivation. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology, 11*, 1–7.

Reiman, E. M., Lane, R. D., Van Petten, C., & Bandettini, P. A. (2000). Positron emission tomography and functional magnetic resonance imaging. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of Psychophysiology* (2nd ed., pp. 85–118). New York: Cambridge University Press.

Robinson, R. G., & Downhill, J. E. (1995). Lateralization of psychopathology in response to focal brain injury. In R. J. D. K. Hugdahl (Ed.), *Brain Asymmetry* (pp. 693–711). Cambridge, MA: MIT Press.

Robinson, R. G., Kubos, K. L., Starr, L. B., Rao, K., & Price, T. R. (1984). Mood disorders in stroke patients: Importance of location of lesion. *Brain, 107*, 81–93.

Rybak, M., Crayton, J. W., Young, I. J., Herba, E., & Konopka, L. M. (2006). Frontal alpha power asymmetry in aggressive children and adolescents with mood and disruptive behavior disorders. *Clinical EEG and Neuroscience, 37*, 16–24.

Silberman, E. K., & Weingartner, H. (1986). Hemispheric lateralization of functions related to emotion. *Brain and Cognition, 5*, 322–353.

Sutton, S. K., & Davidson, R. J. (1997). Prefrontal brain asymmetry: A biological substrate of the behavioral approach and inhibition systems. *Psychological Science 8*, 204–210.

Thut, G., Schultz, W., Roelcke, U., et al. (1997). Activation of the human brain by monetary reward. *Neuroreport, 8*, 1225–1228.

Tyrer, S., & Shopsin, B. (1982). Symptoms and assessment of mania. In E. S. Paykel (Ed.), *Handbook of Affective Disorders* (pp. 12–23). New York: Guilford.

Vallortigara, G., & Rogers, L. (2005). Survival with an asymmetrical brain: Advantages and disadvantages of cerebral lateralization. *Behavioral and Brain Sciences, 28*, 575–633.

van Honk, J., & Schutter, D. J. L. G. (2006). From affective valence to motivational direction: The frontal asymmetry of emotion revised. *Psychological Science, 17*, 963–965.

van Honk, J., Schutter, D. J. L. G., d'Alfonso, A. A. L., Kessels, R. P. C., & de Haan, E. H. F. (2002). 1 hz rTMS over the right prefrontal cortex reduces vigilant attention to unmasked but not to masked fearful faces. *Biological Psychiatry, 52*, 312–317.

van Honk, J., Tuiten, A., de Haan, E., van den Hout, M., & Stam, H. (2001). Attentional biases for angry faces: Relationships to trait anger and anxiety. *Cognition and Emotion, 15*(3), 279–297.

van Honk, J., Tuiten, A., van den Hout, M., et al. (1998). Baseline salivary cortisol levels and preconscious selective attention for threat: A pilot study. *Psychoneuroendocrinology, 23*(7), 741–747.

van Honk, J. Tuiten, A., Verbaten, R., et al. (1999). Correlations among salivary testosterone, mood, and selective attention to threat in humans. *Hormones and Behavior, 36*(1), 17–24.

Watson, D. (2000). Mood and temperament. In D. Watson (Ed.), *Emotions and Social Behavior*. New York: Guilford Press.

Wortman, C. B., & Brehm, J. W. (1975). Responses to uncontrollable outcomes: An integration of reactance theory and the learned helplessness model. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 8, pp. 278–336). New York: Academic Press.

Young, P. T. (1943). *Emotion in Man and Animal: Its Nature and Relation to Attitude and Motive*. New York: John Wiley & Sons.

Zinner, L., Brodish, A., Devine, P. G., & Harmon-Jones, E. (2008). Anger and asymmetrical frontal cortical activity: Evidence for an anger-withdrawal relationship. *Cognition and Emotion, 22*, 1081–1093.

CHAPTER 13
Why Symbolic Processing of Affect Can Disrupt Negative Affect: Social Cognitive and Affective Neuroscience Investigations

Matthew D. Lieberman

In the Highland Indian villages of Guatemala, miniature "worry" dolls approximately 1 inch in height and made from small bits of wood, cloth, and string are given from parent to child. According to legend, parents are meant to say the following along with the presentation of the gift: "If you have a problem, then share it with a worry doll. Before going to bed, tell one worry to each doll, then place them beneath your pillow. Whilst you sleep, the dolls will take your worries away!" It is unclear whether these dolls have actually been imbued with the power to whisk away worry, however there is a great deal of evidence to suggest that the process of sharing one's worry, of putting bad feelings into words, can diminish one's emotional distress, at least under certain circumstances. This chapter examines the neurocognitive mechanisms of *disruption effects*, the process by which putting feelings into words can disrupt the feelings being verbalized.

The notion that labeling emotional states can help to dampen down or regulate negative emotional states is hardly new. In commentary on some of the oldest Buddhist texts, it has been written that "The skillful use of labeling … introduces a healthy degree of inner detachment since the act of apostrophizing [i.e. speaking to] one's moods and emotions diminishes one's identification with them" (Analayo, 2003, p. 113). Similarly, a number of western thinkers have written about disruption effects prior to the twentieth century. The philosopher Benedict Spinoza suggested that "An emotion which is a passion, ceases to be a passion as soon as we form a clear and distinct idea thereof" (1675/2000, p. 291). In the *Principles of Psychology*, William James wrote that "The present conscious state when I say … 'I feel angry' is not the…direct state of anger … it is the state of *saying-I-feel-angry*. The act of naming them has momentarily detracted from their force" (1890, p. 190).

In modern psychology, emotions are often thought to be relatively uncontrollable with direct attempts at regulating one's own emotional state often backfiring (LeDoux, 1996; Wegner, Erber, & Zanakos, 1993; Wegner, Shortt, Blake, & Page, 1990). Nevertheless, the legacy of disruption affects lives in various forms of talk therapies. Talk therapies such as cognitive-behavioral therapy and psychoanalysis vary greatly in their approach and the putative mechanisms supporting successful outcomes; however, they all involve individuals putting feelings into words with the hopes of managing or transforming those feelings.

The insight that putting one's feelings into words can have mental and physical health benefits was captured experimentally in work on disclosure through expressive writing (for a review, *see* Lepore & Smyth, 2002). In the 1980s, Pennebaker began a program of research (Pennebaker & Beall, 1986; Pennebaker, 1997) in which participants were asked to write about past negative experiences on four successive days, and these participants were found to have visited the doctor less often over the following half-year compared to those who wrote about trivial experiences. Although numerous studies have shown health benefits of expressive writing across numerous domains, including blood pressure (McGuire, Greenberg, & Gevirtz, 2005), chronic pain (Broderick, Junghaenel, & Schwartz, 2005), cancer-related symptons (Stanton et al., 2002), lung functioning (Smyth et al., 1999), liver functioning (Francis & Pennebaker, 1992), and immune function (Booth, Petrie, & Pennebaker, 1997), a number of other studies have shown that expressive writing leads to improvements in emotional well-being and mental health more generally (Hemenover, 2003; Park & Blumberg, 2002). It is unclear which aspects of the writing produce the physical and mental health benefits (for a review of different accounts, *see* Baikie & Wilhelm, 2005); however, it is clear that merely thinking about negative experiences without being required to organize those thoughts into words does not have the same benefits and can actually be quite detrimental to mental health (Lyubormirsky, Sousa, & Dickerhoof, 2006; Nolen-Hoeksema, 2000).

INTENTIONAL VERSUS UNINTENTIONAL EMOTION REGULATION

Although the effects of expressive writing look like the results of emotion regulation processes, the expressive writing paradigm lacks certain indicators associated with emotion regulation. When one thinks of emotion regulation, one typically thinks of having a very overt intention to change one's emotional experience or at least the outward manifestations of that experience (Gross, 1998). One imagines "grinning and bearing it" when publicly receiving news that someone else received the promotion you were hoping for. Most would also expect that carrying out this intentional emotion regulation would feel effortful (Richards & Gross, 2000). It is unclear to what extent putting feelings into words, either during expressive writing or in other forms, constitutes an intentional or unintentional form of emotion regulation.

This blurred line between intentional and unintentional regulation is present in some of the earliest work on emotion regulation conducted by Lazarus and others. In these studies (Dandoy & Goldstein, 1990; Lazarus & Alfert, 1964; Lazarus, Opton, Nomikos, & Rankin, 1965; Speisman, Lazarus, Mordkoff, & Davison, 1964), subjects' physiological arousal was measured, typically while they watched disturbing films. By providing a verbal narrative explaining the content of the films in different ways, changes in the physiological responses were obtained. For example, telling subjects that the scene they were about to see was created by actors appearing to get injured and that the injuries were fake led to diminished skin conductance responses while subjects watched the scene, relative to subjects not so informed. The framing of the scene changed the appraisal of the scene's meaning and thus had apparent regulatory effects (i.e., diminished skin conductance responses), but it is unclear whether the subjects engaged in anything they would themselves call emotion regulation. Decades of work on placebo effects have a similar phenomenology associated with them (Benedetti, Mayberg, Wager, Stohler, & Zubieta, 2005), such that a belief or appraisal that a pill will prevent pain actually leads to diminished experiences of pain, despite the pill having no active ingredients. More recent fMRI work (Ochsner, this volume) has put this reframing or *reappraisal* process in the hands of subjects and thus made the process fully overt, asking subjects to understand aversive stimuli in ways that make them less aversive.

The expressive writing studies (Pennebaker et al., 1997) and appraisal studies (Lazarus & Alfert, 1964) suggest that verbal processing of emotional content and explicit changes to the framing of emotional content can serve to regulate emotional responses, even when there is no obvious regulatory intent. Nevertheless, these paradigms could both produce spontaneous intentions to regulate one's emotions, and this could be serving as an unmeasured, but mediating, mechanism. Two other lines of research suggest that intention to regulate one's affect is not, in fact, necessary for the disruption of affect to occur for a broader review of unintentional emotion regulation, see Berkman and Lieberman (2009).

For example, Wilson and Schooler (1991; Wilson et al., 1993) conducted a series of studies demonstrating that reflecting upon and writing about one's own affective state disrupted the impact that their affective states would otherwise have had on their decision making. Critically, in these studies, the task was not focused on emotion regulation at all but instead was focused on merely making good decisions by consulting one's own affective response as a guide. In one study, individuals were asked to choose between a number of works of art and were ultimately able to take one art print home with them. Some individuals were also asked to reflect on their feelings about each of the prints *before* announcing their rating. Surprisingly, individuals who reflected on their feelings before choosing were more likely to choose an art print that they themselves would later regret choosing than individuals who did not reflect on their feelings. The authors suggested that some aspects of feeling states are more verbalizable than others, and when making a decision, we weight verbal information in our minds more heavily than nonverbal feelings. Thus, if good decisions are driven by feelings that cannot be easily verbalized, relying on that which can be verbalized will produce suboptimal decisions. It is also possible, however, that verbalizing one's feelings temporarily altered the feeling states themselves by dampening them. Behavioral data alone cannot easily tease these two interpretations apart (i.e., overemphasizing verbal information vs. dampening of affect) and this was actually one of the original incentives for using fMRI to examine this issue, as it may be better suited for teasing apart these interpretations.

Another study by Greenberg, Wortman, and Stone (1996) more directly addresses the issue of whether regulatory intent is critical for the benefits of putting feelings into words. In this study, an expressive writing paradigm similar to Pennebaker's was used except that an additional condition was included. Individuals in this condition were asked to write about a trauma, but one that was imagined rather than real. Despite the imaginary nature of the traumas written about, these individuals showed benefits of expressive writing similar to those seen in previous studies. It is difficult to argue that these benefits derived from any overt attempts at emotion regulation. Instead, merely putting feelings into words—albeit imagined feelings—produced disruption-like effects.

It is important to note here that I am not suggesting that intentional emotion regulation is reducible to putting feelings into words. The understanding that people have of themselves and of those around them guide their emotional lives, and thus new understandings reached through introspection, disclosure, and reappraisal undoubtedly have the power to transform one's emotional responses. I am simply suggesting that *some of the benefits* derived from these therapeutic techniques may result from neurocognitive consequences of merely putting feelings into words. And if this is the case, these benefits could be put to good use therapeutically, even in cases for which an individual is unwilling or unable to engage in emotion regulation.

RVLPFC AS A CANDIDATE MECHANISM

The rest of this chapter is devoted to exploring one possible neurocognitive mechanism by which putting feelings into words could disrupt basic negative affect processes, thereby improving one's affective state. Disruption theory posits that right ventrolateral prefrontal cortex (RVLPFC; *see* highlighted area in Fig. 13–1d) plays a central role in the disruption effects. RVLPFC long been associated with inhibitory processes and more recently it has been identified in studies examining the symbolic processing of affect. With both of these functions associated with RVLPFC activity, RVLPFC emerges as an ideal candidate for disruption effects, as these effects appear to involve symbolic processing of affect, which leads to the inhibition of affective processes. Before turning to the evidence that experimentally combines these functions in RVLPFC, I first review the evidence that links RVLPFC separately to inhibition and to symbolic processing of affect.

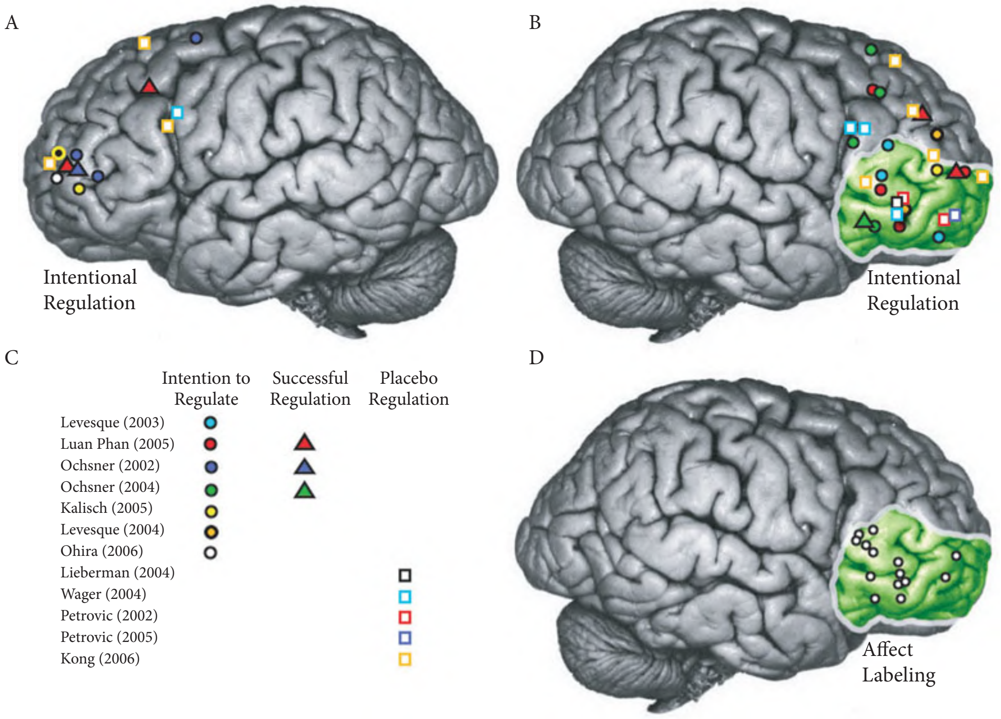


Fig. 13–1 Right ventrolateral prefrontal activity (RVLPFC; highlighted area) in affect labeling and emotion regulation studies. (A) Left lateral and (B) Right lateral activations in studies of emotion regulation and placebo effects. (C) Legend for emotion regulation and placebo effects (D) RVLPFC activations in affect labeling studies.

RVLPFC and Inhibition

Although there is ongoing debate about the full set of neural regions involved in inhibitory processes, RVLPFC would certainly be included in anyone's candidate set. More than a dozen neuro-imaging studies of the Go-NoGo, Flanker, and Stroop tasks have identified RVLPFC activations associated with trying to inhibit a prepotent motor response or trying to ignore task-irrelevant information that would lead to an incorrect response (Asahi, Okamoto, Okada, Yamawaki, & Tokota, 2004; Blasi et al., 2006; Garavan, Ross, & Stein, 1999; Horn, Dolan, Elliott, Deakin, & Woodruff, 2003; Kawashima, 1996; Konishi, 1999; Liddle, Kiehl, & Smith, 2001; Matthews, Simmons, Arce, & Paulus, 2005; Rubia, Smith, Brammer, & Taylor, 2003; Hazeltine, Poldrack, & Gabrieli, 2000; Hazeltine, Bunge, Scanlon, & Gabrieli, 2003; Fan, Flombaum, McCandiss, Thomas, & Posner, 2003; Kemmotsu, Villalobos, Gaffrey, Courchesne, & Muller, 2005; Leung, Skudlarski, Gatenby, Peterson, & Gore, 2000). In addition, these tasks have found that RVLPFC activity is associated with faster reaction times on inhibition trials (Garavan et al., 1999), that RVLPFC activity is greater for successful inhibition trials than unsuccessful inhibition trials (Rubia et al., 2003), and that RVLPFC activity is greater for harder inhibition trials than easy inhibition trials (Matthews et al., 2005). Children with attention deficit hyperactivity disorder (ADHD) show impaired behavioral performance on motor inhibition tasks and also evidence less RVLPFC activity during inhibition tasks than controls (Durston, Mulder, Casey, Ziermans, & van Engeland, 2006; Rubia et al., 1999). One study that observed better motor inhibition in an ADHD sample after neurofeedback training also observed an increase in RVLPFC activity, relative to a sample that did not receive this training (Beauregard & Levesque, 2006). Studies of permanent lesions (Aron, Fletcher, Bullmore, Sahakian, & Robbins, 2003) and temporary lesions to RVLPFC induced by transcranial magnetic stimulation (Chambers et al, 2006) have also found impaired motor inhibition. Finally, pharmacological studies in which participants receive serotonergic agonists, associated with enhanced self-control and diminished impulsivity, observed greater activity in RVLPFC during motor inhibition trials (Anderson et al., 2002; Del Ben et al., 2005; *see also* Rubia et al., 2005, but cf. Vollm et al., 2006).

A fascinating study by Goel and Dolan (2003) suggests that RVLPFC may also be involved in nonmotoric forms of inhibition such as the inhibition of belief. In this study, participants assessed the validity of syllogisms (i.e., Does the conclusion logically follow from the premises?) that were either sound (premises were true) or unsound (one premise was false). Participants had difficulty accurately identifying a valid syllogism as valid if it was unsound and therefore not true. For example, given the premises "All addictive things are expensive" and "Some cigarettes are inexpensive," it is valid to conclude that "Some cigarettes are not addictive" although the first premise and conclusion are false. RVLPFC was the only region of the brain that was more active when participants overcame their belief-bias and indicated that this kind of syllogism was valid. A number of studies on active deception have also suggested a role for RVLPFC in the inhibition of belief (Abe et al., 2006; Spence et al., 2001; Luan Phan et al., 2005; Nunez, Casey, Egner, Hare, & Hirsch, 2005). Across these studies, when individuals were required to inhibit what they knew to be true to say something false, RVLPFC was recruited.

RVLPFC and Symbolic Processing of Affect

There have been many fewer studies examining symbolic processing of affect (SPA) than inhibitory processes, but the percentage of SPA studies implicating RVLPFC is at least as high as that seen in the inhibition literature. SPA refers, roughly, to the explicit linguistic/propositional processing of one's own affect ("I feel sad"), the affect of others ("She looks frightened"), evaluatively valenced categories ("Terrorists are bad"), or the value of response options ("I will lose money if I keep my money in betamax stock"). Across a variety of studies, RVLPFC tends to be more active during SPA than non-SPA, particularly in the case of negatively valenced SPA.

For example, Cunningham and colleagues (Cunningham, Johnson, Gatenby, Gore, & Banaji, 2003) presented participants with famous names like Bill Cosby and Adolph Hitler, who are generally viewed either positively or negatively. On some trials, participants were asked to decide whether the target was alive or dead but on other trials were asked if the target was good or bad. Thus, on all trials, implicit affective responses to the targets should be expected, but explicit SPA should only occur when the targets are evaluated as good or bad. Cunningham et al. (2003) observed that RVLPFC along with medial prefrontal cortex (mPFC) were more active during good/bad judgments than during alive/dead judgments, suggesting that these regions are involved in SPA. They also found that RVLPFC was the region of the brain that was most active during bad judgments relative to good judgments, suggesting a possible selective role in negative SPA.

A number of studies that have focused on explicit judgments about the emotional aspects of pictures (Gorno-Tempini et al., 2001; Gur et al., 2002; Nakamura et al., 1999; Narumoto et al., 2000; Royet, Plailly, Delon-Martin, Kareken, & Segebarth, 2003) and voices (Wildgruber et al., 2004, 2005) demonstrated greater RVLPFC activations to emotional than nonemotional judgments. A study that specifically compared negative emotion judgments to neutral and positive judgments observed greater RVLPFC to negative emotion judgments (Dolcos, LaBar, & Cabeza, 2004), similarly to Cunningham et al. (2003). In addition, multiple studies have observed that reading negatively valenced words is associated with greater RVLPFC than reading neutral or positive words (Cunningham, Espinet, DeYoung, & Zelazo, 2006; Cunningham, Raye, & Johnson, 2004; Kuchinke et al., 2005).

Nomura et al. (2003; *see also* Shaw et al., 2005) compared difficult emotion judgments to easy emotion judgments. Presumably, the difficult judgments required more top-down elaboration of the emotional qualities of the stimulus than the easy judgments and thus would involve more SPA. In this study, participants judged the emotional expression or the gender of target faces. For half of the trials, the critical dimension was ambiguous (e.g., half of the gender trials had faces that were ambiguous with respect to gender). Nomura et al. (2003) found that RVLPFC and the dorsal anterior cingulate cortex (dACC) were the only regions of the brain that were more active during ambiguous trials than unambiguous trials. Importantly, however, the effect in RVLPFC was driven entirely by its response to ambiguous emotion trials, whereas the dACC was equally responsive to both kinds of ambiguity. Thus, one reasonable interpretation of these results is that RVLPFC was recruited on ambiguous emotion trials as participants engaged in explicit hypothesis testing about the emotional expression, which would be consistent with its putative role in SPA.

RVLPFC ANATOMICAL PROJECTIONS TO LIMBIC REGIONS

The preceding sections set up the possibility that SPA in RVLPFC could inhibit activity in limbic regions such as the amygdala, insula, and ACC associated with affective experience. It is important to establish that such a claim is neuro-anatomically plausible. That is, does RVLPFC have the right kinds of neuro-anatomical connections to these other regions to produce these regulatory effects? For the connections to the insula and ACC, the answer is a resounding yes. RVLPFC has strong bidirectional connections with both of these regions (Augustine, 1996; Vogt & Pandya, 1987).

The neuro-anatomical connections from RVLPFC to the amygdala are more complex. On the one hand, there are direct projections from RVLPFC to the amygdala. Carmichael and Price (1995; *see also* Ghashghaei & Barbas, 2002; McDonald, Mascagni, & Guo, 1996) made anterograde tracer injections into area 12l (the region in the rhesus monkey homologous to Brodmann's area 47 in humans) and found evidence of projections from area 12l to the basolateral nucleus of the amygdala (BLA). However, these projections are not particularly dense, calling into question whether these direct projections are sufficient to allow RVLPFC to regulate amygdala responses. As suggested by Phelps, Delgado, Nearing, and LeDoux (2004), RVLPFC could also have its effect on the amygdala indirectly by way of projections from RVLPFC to mPFC, which in turn has dense projections to the amygdala (Carmichael & Price, 1995) and is known to regulate the amygdala in studies of extinction (Phelps et al., 2004; Quirk, Likhtik, Pelletier, & Pare, 2003).

RVLPFC DIMINISHES NEURAL AND SUBJECTIVE NEGATIVE AFFECT

This section reviews research that suggests that RVLPFC not only inhibits motor and cognitive responses but also inhibits negative affective responses both in terms of subjective reports of negative affect and in terms of activity in limbic regions associated with negative affect and distress. In light of the previous sections that establish a major role for RVLPFC in *(1)* inhibitory processes; *(2)* the symbolic processing of negative affect; and *(3)* possessing neuro-anatomical connections to limbic regions, it is perhaps not a giant leap to suggest that RVLPFC may contribute to the inhibition of motoric, cognitive, and affective responses. Nevertheless, establishing this relationship will serve as a critical stepping stone to full-blown disruption effects reviewed in the next section.

RVLPFC is one of the regions that has been associated with increased pain analgesia (Petrovic, Kalso, Petersson, & Ingvar, 2002). More recently, a number of studies have observed that placebo effects appear to be mediated by RVLPFC, along with rostral anterior cingulate cortex (rACC). In one study, we (Lieberman et al., 2004) examined a group of patients with irritable bowel syndrome (IBS), a chronic pain condition associated with heightened pain sensitivity in the limbic system (Naliboff et al., 2006). The IBS patients were scanned prior to and then again after receiving 3 weeks of sham treatment with placebos for their pain. During each scanning session, patients received painful rectal stimulation, simulating the symptoms of IBS and generating a measure of current neural responses to this stimulation. We found that to the extent that participants reported improvements in their pain symptoms at the end of the placebo regimen, compared to before the regimen began, they also showed increased activity in RVLPFC and decreased dACC activity from the first scanning session to the second. Multiple other studies have also observed within session placebo effects associated with increased RVLPFC activity and decreased limbic activity in the domains of physical pain (Petrovic et al., 2002; Wager et al., 2004) and anxiety (Petrovic et al., 2005).

We have also examined the role of RVLPFC in the regulation of "social pain" or the distress associated with social rejection (Eisenberger, this volume; Eisenberger & Lieberman, 2004). In one study (Eisenberger, Lieberman, & Williams, 2003), participants ostensibly played a game of Internet "catch" with two other players, who were actually computer simulations. Part of the way through the game, the other players stopped throwing the ball to the participant and thus excluded the participant for the rest of the game. Numerous behavioral studies have shown that this exclusion manipulation causes considerable distress in participants, even when they know the other players are just computer simulations (Williams, 2007). Our participants also reported being distressed in response to being excluded and showed a pattern of neural activity consistent with the experience of visceral pain (*see also* Eisenberger, Way, Taylor, Welch, & Lieberman, 2007). Most relevant here is that participants produced increased activity in dACC to the extent that they felt; however, to the extent that RVLPFC was active, participants reported feeling less distressed by the episode of exclusion. Moreover, activity in RVLPFC was negatively correlated with dACC activity, and changes in dACC activity mediated the relationship between RVLPFC and distress. In other words, it appears that increased RVLPFC activity may have helped to downregulate dACC responses, which in turn were associated with reduced distress.

In contrast to the social and physical pain studies, fMRI studies of reappraisal explicitly instruct subjects to engage in emotion regulation. Nearly all of the fMRI studies of reappraisal have observed activity in or near RVLPFC along with other prefrontal regions (*see* Fig. 13–1a & 13–1b: Beauregard, Levesque, & Bourgouin, 2001; Kalisch et al., 2005; Levesque et al., 2003; Luan Phan et al., 2005; Ochsner et al., 2004; Schaefer,et al., 2003; cf. Ochsner, Bunge, Gross, & Gabrieli, 2002).

A handful of other studies have implicated RVLPFC in the regulation of emotional behaviors. These studies may be something of a blend between the motor inhibition and emotion regulation paradigms, supporting the notion that RVLPFC is involved in a continuum of regulatory effects. In one study (Small, Zatorre, Dagher, Evans, & Jones-Gotman, 2001), participants were required to eat a piece of chocolate during each of a series of PET scans. After each scan, participants indicated how much they liked eating the chocolate and how much they wanted to have another piece. Predictably, in early scans, participants liked the chocolate and wanted more; however, by the second half of the study, the participants did not like the chocolate anymore and did not want to eat another piece. Activity in RVLPFC was strongly associated with self-reports of not wanting to eat anymore chocolate despite being asked by the experimenter to continue eating it, suggesting that RVLPFC may have been involved in suppressing the desire to reject the chocolate to comply with the requirements of the study (i.e., eating the unwanted chocolate). Note that although not framed as such in this study, the results may have implications for future work on the neural correlates of compliance and conformity.

In another recent study (Tabibnia, Satpute, & Lieberman, 2008b), we examined how individuals overcome the slight of insulting unfair offers in a financial bargaining game to accept financially advantageous offers. Participants played the "responder" role in several one-shot versions of the ultimatum game. In this game, the "proposer" is asked to split a sum of money between him/herself and the responder. Thus, if the proposer has a $10 stake to split, she may propose an even split of $5 and $5 or, perhaps, a more unfair split of $8 for herself and $2 for the responder. The responder then decides whether or not to accept the offer. If the responder accepts, then both the proposer and responder get exactly what the proposer proposed. However, if the responder rejects the proposal, neither participant receives anything. Either way, there is no additional bargaining after the responder responds.

An earlier fMRI study of the ultimatum game (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003) compared the neural responses to fair ($5 out of $10) and unfair ($1 out of $10) offers. The main finding was that unfair offers were associated with increased activity in the anterior insula, a region that has previously been associated with disgust responses. In our study (Tabibnia et al., 2008b), we also included offers that were unfair and yet still financially desirable to undergraduate participants. In the study by Sanfey et al., both kinds of offers presented little conflict as the $5 offers were both fair and desirable, financially, whereas the $1 (and $2) offers were unfair and not that desirable, financially. To create this conflict between fairness and financial desirability, we included offers such as $5 out of $23, which were both insulting and yet also financially desirable. What we found across a number of different analyses is that the tendency to reject unfair but financially desirable offers was associated with activity in the anterior insula, consistent with the results from Sanfey et al. However, the tendency to accept the unfair but financially desirable offers was associated with activity in RVLPFC. Moreover, greater RVLPFC activity on these trials was associated with diminished anterior insula activity, and changes in anterior insula activity mediated the relationship between RVLPFC activity and the tendency to accept unfair offers. These results are consistent with the idea that RVLPFC is involved in dampening the limbic response to the insulting offer, allowing the individual to "swallow one's pride" and accept the unfair offer.

Symbolic processing of affect disrupts affect via RVLPFC

I have established that RVLPFC activity is associated with the inhibition of motor, cognitive, and emotional responses. Additionally, RVLPFC is active in various forms of SPA, particularly negatively valenced SPA (SPA$_{\text{Neg}}$). If SPA$_{\text{Neg}}$ activates RVLPFC and activity in RVLPFC is associated with the inhibition of emotional responses, then it seems plausible that SPA$_{\text{Neg}}$ would be associated with the inhibition of emotional responses and that activity in RVLPFC would be largely responsible for this effect.

Prior to the studies that directly linked SPA with the downregulation of affect, there were also a handful of studies suggestive of this link without overtly assessing it. Hornak, Rolls, and Wade (1996) tested a sample of patients with ventral prefrontal damage and found that these patients were impaired at explicitly recognizing emotional face expressions and voice tones. Of the 11 patients in the sample, 9 had right or bilateral ventral damage, and 8 of these were impaired on one or both SPA tests. Of the 2 "left-only" ventral prefrontal patients, one performed well above the mean of the nonventral controls. Additionally, the extent of impairment in SPA tasks was correlated with disinhibition of emotional behavior, suggesting that impaired ability to engage in SPA is associated with more emotional behavior and that this association may be related to ventral prefrontal impairment.

Hariri, Bookheimer, and Mazziotta (2000) produced the first evidence of the complete pathway from SPA to RVLPFC activity to reduced amygdala activity. In their study, participants judged the emotional identity of a target's facial expression, however, the trials varied with respect to whether symbolic processing was required to make the judgment. In the SPA condition ("affect label"; *see* Fig. 13–2a), a target face was presented at the top of the screen along with two emotion words (e.g., "angry," "surprised") at the bottom of the display, and participants had to choose which of

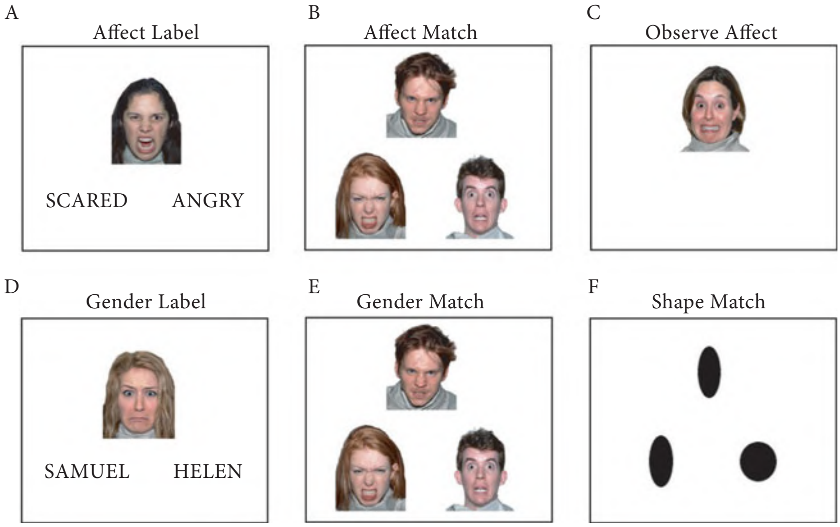


Fig. 13–2 Sample trials from an affect labeling study (Lieberman et al., 2007).

words best described the target's emotion. In the non-SPA condition ("affect match"; *see* Fig. 13–2b), a target face was presented at the top of the screen along with two other emotional faces at the bottom of the display and participants had to choose which of these were showing the same emotion as the target face. According to Hariri et al., in the non-SPA condition participants could "match the faces based on perceptual characteristics, such as wide eyes, furrowed brow or clenched teeth, but need not judge or interpret the information" (p. 44). Indeed, when viewing these stimuli, there is a strong sense of "pop-out" in the non-SPA stimuli in which the faces that match seem to automatically pop-out together.

In the non-SPA condition, there was significant amygdala activity relative to a shape-matching control condition ("shape match"; *see* Fig. 13–2f); however, there was no amygdala activity observed during the SPA condition. Instead, SPA was associated with activity in RVLPFC and the fusiform "face" area, the latter presumably indicating that the target face was still being attended to in the SPA condition. In the direct comparison of SPA and non-SPA trials, greater RVLPFC and diminished amygdala activity was observed during the SPA trials. Thus, two different forms of emotional processing—one symbolic and one non-symbolic—appear to be routed through distinct neural systems. Given that the amygdala has been shown in multiple studies of affective processing to be activated by conditions that would allow only automatic processing (i.e., subliminal presentations and binocular rivalry studies; Morris et al., 1998; Whalen et al., 1998; Pasley, Mayes, & Schultz, 2004; Villeumier, Armony, Driver, & Dolan, 2001), it is quite surprising to see the amygdala not responding under conditions that would allow both automatic and controlled processing.

In a follow-up study, we (Lieberman, Hariri, Jarcho, Eisenberger, & Bookheimer, 2005) compared SPA and non-SPA processing in the context of race. Rather than using different facial expressions of emotions, we used all neutral expression faces that varied by race. In the United States, the stereotypes of Blacks are evaluatively negative, particularly when assessed implicitly (Devine, 1989). Indeed, even U.S. Blacks have more negative implicit stereotypes of Blacks than of Whites (Nosek, Banaji, & Greenwald, 2002; Livingston & Brewer, 2002). Consistent with these behavioral findings, a number of neuro-imaging studies have observed greater amygdala activity to Black faces than to White faces, at least to the extent that participants possessed strong anti-Black implicit stereotypes (Phelps et al., 2004; Cunningham, Johnson, Raye, Gatenby, Gore, & Banaji, 2004). We reasoned that because a neutrally expressive Black face produces a similar amygdala response as a negatively expressive White face, engaging in SPA by labeling the race of Black target faces might disrupt this race-related amygdala activity in much the same way that affect labeling disrupts the amygdala response to negatively expressive faces. It is worth noting that another reasonable hypothesis is that race labels would focus attention onto the negative stereotyped aspect of the targets (i.e., race) rather than on other more neutral or positive aspects (i.e., gender) and would therefore produce greater activity in the amygdala.

As in other race fMRI studies, we observed greater amygdala to Black faces than White faces when participants performed a "race-match" task (visually analogous to the trial shown in Fig. 13–2b) that did not require SPA. In fact, we observed this separately for both our White and Black participants. That is, Black participants produced greater amygdala activity to Black faces than White faces, consistent with the previous behavioral findings of Blacks displaying negative implicit stereotypes towards Blacks (Nosek et al., 2002; Livingston & Brewer, 2002).

In contrast to the non-SPA condition, when participants performed the "race-label" task (analogous to Fig. 13–2a), there was no differential amygdala activity to Black and White faces, and the amygdala responses to Black faces diminished compared to the amygdala response during race matching of Black faces and even compared to the control task that did not involve faces at all. As predicted, there was greater RVLPFC activity during race labeling of Black faces (SPA_{Neg}) but not during the race labeling of White faces (SPA_{Pos}). Additionally, there was a strong negative correlation between RVLPFC and amygdala activity during race labeling of Black faces such that the individuals who activated RVLPFC the most during these blocks also tended to activate the amygdala the least. Finally, all of these effects were evident for both the Black and White participants.

Disruption effects redux

The advantage of the affect labeling paradigm over previous SPA studies is that during both matching and labeling conditions, attention is focused on the emotional aspects of the stimulus, with only the need to engage in SPA varying across the conditions. Affect labeling requires SPA, whereas affect matching does not, although affect matching does not prevent spontaneous SPA. Additionally, by using verbal labels that appear in different positions across trials, participants cannot learn a stimulus response mapping between, say, perceptual cues of fear and a right button press. Participants need to read the labels on each trial to see which options are available.

Despite these advantages, there are some inferential limitations present in the original formulation of the affect labeling paradigm. Although the comparison of the affect label to the affect match conditions represents a comparison of SPA and non-SPA, this distinction is confounded with other differences between the conditions. First and foremost, affect match trials present three faces, of which at least two are posing negative emotional expressions on most trials. In contrast, the affect label trials never present more than a single negatively expressive face. Thus, one could argue that greater amygdala activity is present in the affect matching condition because there are more amygdala activating stimuli present on those trials. This argument is not entirely satisfactory given that a single negatively expressive face, even presented subliminally, is usually sufficient to produce amygdala activity (Morris et al., 1998; Whalen et al., 1998), whereas neither of the two affect labeling studies reported the presence of amygdala activity during the affect labeling condition.

Another possibility is that affect labeling is not really affecting amygdala activity, but rather, affect matching leads to hyper-amygdala responses and thus the difference between the two conditions emerges. This criticism does not address the issue of why there has been no amygdala activity observed during the affect labeling condition, but it does raise the important issue that affect matching has different task requirements than tasks that typically provoke amygdala activity such as passive observation of faces or making gender judgments of faces. It is unknown how much the difference between the labeling and matching conditions results from each of these factors because a passive observation condition has not been included.

The last criticism of the paradigm acknowledges that the labeling condition is indeed modulating amygdala activity but takes issue with the source of this modulation. Although we have characterized the affect labeling task in terms of SPA and non-SPA, one could just as easily label them as cognitive and perceptual processes more generally without making any claims about the affective component of these tasks. In other words, perhaps any kind of cognitive or verbal labeling process will diminish the amygdala response to these emotional stimuli.

To address all of these concerns, we ran a modified version of the affect labeling task that included a number of control conditions (Lieberman et al., 2007). All of the conditions of this study are shown in Figure 13–2. We included a passive observation condition (Fig. 13–2c) during which subjects were presented with a single negative emotional target face on each trial and simply attended to the face. This condition was used to construct regions of interest (ROIs) in the amygdala, which could then be compared across all conditions to examine the modulatory effect of other forms of stimulus processing. Then, in addition to the standard affect label and affect match conditions, we included gender label and gender match conditions (Fig. 13–2d & 13–2e). The comparison of affect label and gender label is the most critical comparison as both conditions present only a single target face and both involve labeling—albeit different kinds of labeling (affective vs. non-affective).

As can be seen in Figure 13–3, affect match, gender match, and gender label each produced amygdala activity that was statistically equivalent to that produced during the passive observation of emotional faces ("observe"). Only affect labeling produced significantly less amygdala than the observe condition. Affect labeling also produced less amygdala activity than gender labeling or affect matching, indicating that this effect really resulted from SPA rather than the number of faces on each trial or cognitive processes more generally. Incidentally, in whole-brain analyses, a number of limbic and paralimbic structures were also less active during affect labeling than gender labeling, including dorsal ACC, subgenual ACC, posterior insula, and ventromedial PFC.

In contrast, only a single region of the brain, RVLPFC, was more active during affect labeling

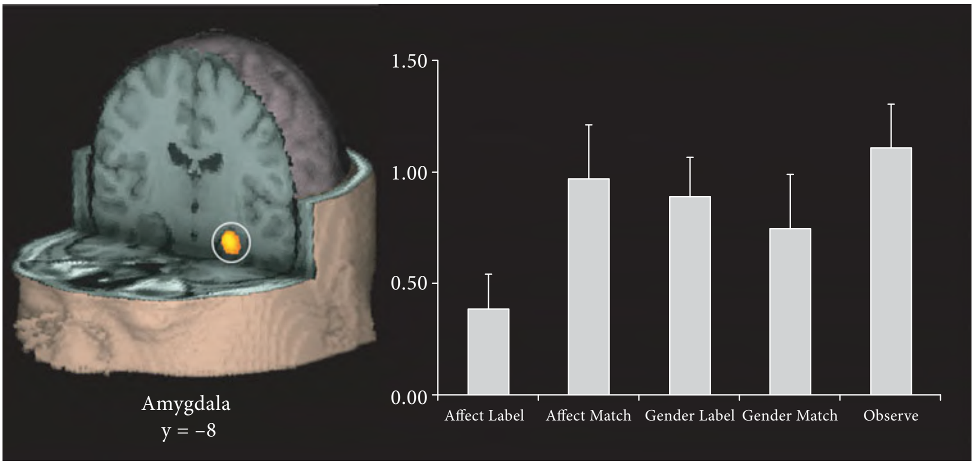


Fig. 13–3 Amygdala response under various processing conditions. Only affect labeling produced a lower level of amygdala activity than simply observing a negative emotional face.

than gender labeling. In addition, after running a correlational analysis using the amygdala cluster from the comparison of affect and gender labeling as a seed, we found that RVLPFC was one of only two regions that had a negatively correlated pattern of activity during this comparison. In other words, if one wanted to know which subjects produced the least amygdala activity during affect labeling, relative to gender labeling, finding the subjects who had the most activity in RVLPFC would be the way to do this. Interestingly, mPFC in BA10 was the only other region of the brain to show this pattern. This is interesting because mPFC has been identified as a possible mediator of RVLPFC effects on the amygdala. Additionally, mPFC is critical to extinction processes and the regulation of the amygdala in this context (Phelps et al., 2004; Quirk et al., 2003) and has been associated with reflective emotional processes (Lane et al., 1997; Taylor, Phan, Decker, & Liberzon, 2003). In running a mediational analysis, we found support for the RVLPFC→mPFC→amygdala pathway effect such that the relationship between RVLPFC and the amygdala during affect labeling was significantly mediated by mPFC activity.

In a psychophysiological follow-up, we found similar results for skin conductance, paralleling the amygdala findings in the fMRI research. In this study (Crockett, Lieberman, & Tabibnia, unpublished manuscript), subjects performed the affect label, affect match, gender label, and shape match tasks while skin conductance responses (SCR) were measured. Across the entire sample, affect labeling was associated with smaller SCRs than affect matching and equivalent SCRs to the shape-matching control task. Gender labeling produced SCRs between the levels observed for affect labeling and affect matching but was not significantly different from either. One reason these latter effects may not have been significant is that a number of subjects did not show reliable SCRs in any of the conditions, which dampened the statistical power of the entire sample. This may have occurred because face stimuli are not as emotionally provocative as other stimuli known to produce strong SCRs (Britton, Taylor, Sudheimer, & Liberzon, 2006), such as the images from the International Affect Picture System (IAPS; Lang, Bradley, & Cuthbert, 1999). When we separated the sample into high and low neurotics, a clearer picture emerged. Non-neurotics, who tend to be less reactive to negatively valenced stimuli, showed no reliable SCR differences across any of the conditions. Those high in neuroticism, however, produced strong SCR responses to affect match and gender label trials and much weaker SCR responses to affect label and shape match trials. Thus, for those that were showing SCR responses at all to the emotional stimuli, the disruption hypotheses were fully supported.

Affect labeling and behavioral inhibition

RVLPFC activity is associated with reduced activity in limbic regions, such as the amygdala and dACC, and SPA is associated both with increased RVLPFC activity and decreased limbic activity. One of the core reasons for pursuing this line of work is the established role of RVLPFC in motor and behavioral inhibition. In light of these various effects, it is reasonable to ask whether SPA, which activates RVLPFC, also has inhibitory effects on behavior. Perhaps RVLPFC produces various forms of inhibition simultaneously (although past studies have typically looked at motor, cognitive, or affective inhibition alone), and perhaps SPA sets the various forms of inhibition in motion. This would certainly be consistent with claims of Goethe, Emerson, Dewey, Arendt and others that thought paralyzes action. In a recent study, Robinson and Wiklowski (2006) found behavioral evidence indicating that SPA_{Neg} leads to motor inhibition, observing that reading negatively valenced primes, but not neutral or positive prime words, led to longer reaction times on a simple motor response task.

We conducted an fMRI study (Lieberman, Eisenberger, & Crockett, unpublished manuscript) to examine the effects of priming a negative stereotype on walking speed. We adapted the classic "automatic behavior" study (Bargh, Chen, & Burrows, 1996) in which priming the "elderly" stereotype leads to slower walking, for use in the scanner environment. We reasoned that reading sentences related to the negative valenced stereotype of the "elderly" constitutes a form of SPA_{Neg} just as labeling the race of Black targets did in our previous study (Lieberman et al., 2005). If true, this would be expected to activate RVLPFC and diminish the activation in limbic structures and possibly inhibit motor processes as well, which could promote slower walking.

This is exactly what we found. After being primed with sentences related to the elderly stereotype in the scanner, participants walked more slowly than they did before scanning. Although part of this effect was no doubt the result of the general sluggishness felt after scanning, we were interested in how neural activity during the sentence priming related to the changes in walking speed from pre- to post-scanning. We found that RVLPFC was the only region of the brain for which greater activity during the priming of the elderly stereotype was associated with more slowing from pre- to post-scan walking measurements. As in our previous studies, we also observed greater increases in RVLPFC associated with reductions in limbic areas, including the amygala and dACC. However, greater activity in RVLPFC was also associated with less activity in the cerebellar vermis, a region that has been associated with motor processes related to walking and lower limb control (Jahn et al., 2004; Martin, 1996). Moreover, during the presentation of sentences related to the elderly, compared to control sentences, this same region of cerebellum was less active. Thus, in this study, SPA_{Neg} not only activated RVLPFC and attenuated limbic responses but also attenuated activity in a region linked to motor preparation and to walking behavior, suggesting that SPA_{Neg} may, in fact, produce motor inhibition as well as emotion regulation. It should also be noted that the RVLPFC-limbic effects occurred in this study despite any plausible impetus for subjects to intentionally engage in emotion regulation. Consequently, it appears that the desire to regulate one's emotional responses may not be necessary to receive the regulatory benefits of activating RVLPFC, consistent with previous research on the benefits of writing about imaginary traumas (Greenberg et al., 1996).

Clinical applications

Given that SPA appears to regulate limbic responses without the intention to do so, this would provide a mechanism by which putting feelings into words would have benefits for regulating emotional distress and for mental health more generally. In an initial attempt to bridge between disruption studies and clinical therapy, we have conducted a series of studies that integrate a SPA manipulation into an analogue of exposure therapy.

In one study, Tabibnia, Lieberman, and Craske (2008a) presented participants with a number of different high-arousal negative images from the IAPS (Langet al., 1999) on Day 1 while SCR was measured. Each of the pictures was presented a total of six times throughout the session to mimic the repeated exposure involved in exposure therapy (Foa & Kozak, 1986). Some of the pictures were presented alone on each trial, whereas others were presented and then followed by either a neutral or negatively valenced word on each trial. Once a picture was presented alone, with a negative word, or with a neutral word, the picture was presented the same way for all the trials. However, the specific words used varied with each presentation, such that a picture presented with negative words would be presented with six different negative words across the six presentations, thus preventing strong associations to a particular word. Exposure therapy is based on the premise that allowing individuals to fully experience an emotional response to a feared stimulus on multiple occasions will allow that emotional response to subside over time. In light of this, the temporal placement of the affect labels was deemed critical. We presented the words 3.5 seconds after the pictures to allow a full physiological response to emerge. Because disruption theory posits that the labels can reduce these responses, simultaneous presentation of pictures and words might actually prevent exposure effects from occurring.

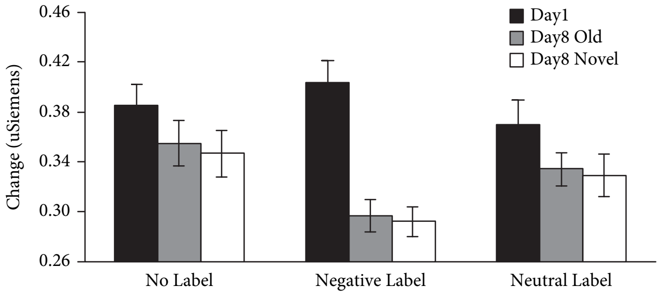


Fig. 13–4 Spider phobic skin conductance responses to spider images as a function of day and initial encoding condition. Higher bars indicate greater reactivity. For the labeling conditions (Negative Label, Neutral Label), the labels were present on Day 1, but on Day 8, pictures were presented without labels for all conditions.

A week after the first session, participants returned for a second session. On Day 8, participants were again shown the same pictures from Day 1 while SCR was measured; however, on Day 8, no words were shown for any of the conditions. By comparing SCR to pictures in each condition across the two sessions, we could determine the extent to which repeated exposure on Day 1 led to diminished SCR a week later and also whether the addition of affect labels enhanced this effect. As predicted, pictures that had been presented alone on Day 1 produced diminished SCRs on Day 8. This was also true for pictures that were presented with negative words on Day 1; however, pictures presented with neutral words on Day 1 only showed a trend in this direction. Critically, although both pictures shown alone and pictures shown with negative words showed diminished SCRs on Day 8, the reduction for the negative word condition was greater than the reduction for the no word condition.

This effect was replicated in a second study (Tabibnia et al., 2008a), examining the SCRs of individuals with spider fears to pictures of spiders. In this between-groups study, individuals saw pictures of spiders in one condition only (no words, negative words, or neutral words). In each condition, participants produced smaller SCRs to spider pictures on Day 8 than on Day 1 and replicating the first study, this effect was significantly greater in the negative words condition than the no words condition (*see* Fig. 13–4). Interestingly, the effects of the negative words shown on Day 1 generalized to new pictures of spiders that were not shown on Day 1 and had never been paired with words. Thus, these results suggest that pairing affect labels with repeated exposures of feared stimuli can lead to long-term reductions in the emotional responses to those stimuli.

More generally, these results point to the benefits of examining how specific symbolic processes unique to humans can benefit mental health processes. There has been a great deal of work in the past decade to translate the animal research on extinction processes into the human domain and demonstrating that these processes do translate from rodent to human. At the same time, humans have specific capacities that we do not share with other animals and these undoubtedly modulate the ways in which the lower processes operate within humans (Davey, 1992).

Social cognitive implications

Automaticity and Control

In addition to the applied clinical applications of disruption theory, this work also has important implications for both theory and methods within social cognition. First, the findings from this work suggest that our basic definitions of automaticity and control, a core distinction within social cognition (Chaiken & Trope, 1999), need to be revisited (cf. Bargh, 1989). One of the gold standards for determining whether a process is automatic is to observe whether the process still occurs when the eliciting stimulus is presented subliminally (Monahan, Murphy, & Zajonc, 2000; Murphy & Zajonc, 1993). Thus, if a trait word is presented subliminally and influences subsequent personality judgments, all would agree that this represents automatic or implicit priming. A second standard that has been used has been the amount of time a mental process takes to occur. Generally speaking, the effects of a prime word on the processing of a second word that follows within 300 milliseconds of the prime word are thought to be automatic (Neely, 1977). Finally, processes that produce the same outputs when a person is under cognitive load (i.e., mental distraction usually caused by a concurrent task), as when there is no cognitive load, are also considered to be automatic (Gilbert, 1989).

By the first two of these definitions, the amygdala response to emotional images is an exemplary case of automaticity. Multiple studies have demonstrated that the amygdala responds to subliminal presentations of emotional images (Morris et al., 2000; Whalen et al., 1999) and also that the amygdala responds within 150 milliseconds of stimulus presentation. Clearly, no conscious mental resources are needed to produce the amygdala's response to emotional stimuli. Indeed, the race-matching task, which produced the greatest amount of amygdala activity in a comparison with race labeling (Lieberman et al., 2005), was performed at the same speed with a concurrent working memory task as without this task.

Nevertheless, when individuals are asked to process affect labels while looking at negative emotional images, the amygdala response either disappears or is significantly attenuated. Here, the presence of a particular kind of concurrent controlled processing task (i.e., affect labeling) modulates what would otherwise be an automatic response in the amygdala. This runs counter to the dogma of standard dual-process models that controlled processes cannot affect automatic processes. How could a process that can occur during subliminal presentations when an individual has no awarness at all of the eliciting stimulus possibly be prevented or attenuated by conscious processing?

Once a cognitive neuroscience approach to automatic and controlled social cognition is taken (Lieberman, 2007), the answer is actually quite straightforward. One possibility is that the amygdala performs its operations automatically as has often been supposed (Pasley et al., 2004). On this account, the amygdala in no way depends on cognitive resources or controlled processing to perform its computations. However, the amygdala receives inputs from various regions of prefrontal cortex (Ghashghei & Barbas, 2002), and the functional effect of some of these inputs may be inhibitory (Quirk et al., 2003; Rosenkranz & Grace, 2002). Affect labeling may interfere with amygdala processing not because they compete for a limited pool of cognitive resources (as is assumed to be the case for competing controlled processes) but because affect labeling just happens to activate a prefrontal region that has inhibitory inputs to the amygdala. Thus, processes internal to the amygdala may well be automatic, and yet at the same time, other brain structures may be capable of modulating or inhibiting these processes. On the one hand, this suggests that some individual neural mechanisms may follow the standard principles of automaticity, but on the other hand it suggests that at a system level, our understanding of automaticity and control may be far too simplistic.

Semantic versus Embodied Emotion

A second issue for social cognition is the use of word-and-picture primes in experimental studies. It is not uncommon for social psychological research to use word-and-picture primes interchangeably (e.g., Dasgupta & Greenwald, 2001; Devine, Plant, Amodio, Harmon-Jones, & Vance, 2002; Galinsky & Moskowitz, 2000; Lowery, Hardin, & Sinclair, 2001; Wittenbrink, Judd, & Park, 2001). This may be a result of assuming that there are unified representations in the mind and that any stimulus relevant to that mental construct is going to activate this unified representation. From this perspective, it might seem that the same representation can be activated implicitly or explicitly, but ultimately the same representation is invoked. The cognitive neuroscience of memory has demonstrated that not only are there implicit and explicit memory processes (i.e., ways of using and invoking mental representations) but also that there are distinct neural mechanisms that retain different aspects of our past experience in qualitatively distinct representations (e.g., episodic, semantic, conditioning). In the context of affect labeling research, it seems that negative emotional stimuli can also activate distinct representations and processes depending to some extent on whether the eliciting stimuli are words or pictures. Negatively valenced pictures reliably activate the amygdala and also lead to SCR increases, suggesting that *embodied* emotional processes are invoked. Alternatively, negatively valenced words produce neither of these effects and instead activate RVLPFC. Thus, it is possible that these words are producing thoughts about emotion but are not producing, or may even be inhibiting, basic emotional responses. In a pure social cognition task with word-only primes, this effect may be overlooked as negatively valenced words will presumably activate a *semantic* network of emotion representations (Robinson & Clore, 2002). It appears that it would be a mistake, however, to infer from the activation of this semantic network that more basic and embodied emotional processes have also been activated. Although this distinction has yet to be fully fleshed out, it does suggest that we may not be priming what we think we're priming in affect priming studies.

Conclusion

Numerous philosophers and psychologists have noted over the years that thinking about affect has the capacity to alter and even dampen the affect that is being thought about. This has been used to good effect in various forms of therapy, from formal psychotherapies to informal social support networks in which people talk about their feelings with friends. The reason that putting feelings into words helps has remained elusive and somewhat mystical. The work presented here describes a neurocognitive process focused on RVLPFC that provides the beginnings of an answer. Putting feelings into words activates a region of the brain that is capable of inhibiting various aspects of immediate experience, including affective distress. Although we cannot say why the brain evolved such that putting feelings into words has this neurocognitive effect, knowing that it does allows us to probe various aspects of this process in the future and examine its contribution to various social and affective experience in healthy and clinical populations.

Acknowledgments

Although the references have been updated, this chapter was written a few years ago and with fast developments in social cognitive neuroscience, more recent reviews can be found in Berkman and Lieberman (2009), Cohen and Lieberman (2010), and Lieberman (2010).

References

Abe, N, Suzuki, M., Tsukiura, T., et al. (2006). Dissociable roles of prefrontal and anterior cingulate cortices in deception. *Cerebral Cortex, 16*, 192–199.

Analayo (2003). *Satipatthana: The Direct Path to Realization*. Birmingham, UK: Windhorse publications.

Anderson, I. M., Clark, L., Elliott, R., Kulkarni, B., Williams, S. R., & Deakin, J. F. W. (2002). 5-HT2c receptor activation by m-chlorophenylpiperazine detected in humans with fMRI. *Neuroreport, 13*, 1547–1551.

Aron, A. R., Fletcher, P. C., Bullmore, E. T., Sahakian, B. J., & Robbins, T. W. (2003). Stop-signal inhibition disrupted by damage to right inferior frontal gyrus in humans. *Nature Neuroscience, 6*, 115–116.

Asahi, S., Okamoto, Y., Okada, G., Yamawaki, S., & Yokota, N. (2004). Negative correlation between right prefrontal activity during response inhibition and impulsiveness: A fMRI study. *European Archives of Clinical Neuroscience, 254*, 245–251.

Augustine, J. R. (1996). Circuitry and functional aspects of the insular lobe in primates including humans. *Brain Research Reviews, 22*, 229–244.

Baikie, K. A. & Wilhelm, K. (2005). Emotional and physical health benefits of expressive writing. *Advances in Psychiatric Treatment, 11*, 338–346.

Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended Thought* (pp. 3–51). New York: Guilford.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*, 230–244.

Beauregard, M., & Levesque, J. (2006). Functional magnetic resonance imaging investigation of the effects of neurofeedback training on the neural bases of selective attention and response inhibition in children with attention-deficit/hyperactivity disorder. *Applied Psychophysiological Biofeedback, 31*, 3–20.

Beauregard, M., Levesque, J., & Bourgouin, P. (2001). Neural correlates of conscious self-regulation of emotion. *Journal of Neuroscience, 21*, 6993–7000.

Bennedetti, F., Mayberg, H. S., Wager, T. D., Stohler, C. S., & Zubieta, J. K. (2005). Neurobiological mechanisms of the placebo effect. *Journal of Neuroscience, 25*, 10,390–10,402.

Berkman, E., & Lieberman, M. D. (2009). Using neuroscience to broaden emotion regulation: Theoretical and methodological considerations. *Social and Personality Psychology Compass, 3*, 475–493.

Blasi, G., Goldberg, T. E., Weickert, T., et al. (2006). Brain regions underlying response inhibition and interference monitoring and suppression. *European Journal of Neuroscience, 23*, 1658–1664.

Booth, R. J., Petrie, K. J., & Pennebaker, J. W. (1997). Changes in circulating lymphocyte numbers following emotional disclosure: evidence of buffering? *Stress Medicine, 13*, 23–29.

Britton, J. C., Taylor, S. F., Sudheimer, K. D., & Liberzon, I. (2006). Facial expressions and complex IAPS pictures: Common and differential networks. *Neuroimage, 31*, 906–919.

Broderick, J. E., Junghaenel, D. U., & Schwartz, J. E. (2005). Written emotional expression produces health benefits in fibromyalgia patients. *Psychosomatic Medicine, 67*, 326–334.

Carmichael, S. T., & Price, J. L. (1995). Limbic connections of the orbital and medial prefrontal cortex in the macaque monkey. *Journal of Comparative Neurology, 363*, 615–641.

Chaiken, S., & Trope, Y. (1999). *Dual-process Theories in Social Psychology*. New York: Guilford Press.

Chambers, C. D., Bellgrove, M. A., Stokes, M. G., et al. (2006). Executive "brake failure" following deactivation of human frontal lobe. *Journal of Cognitive Neuroscience, 18*, 444–455.

Cohen, J. R. & Lieberman, M. D. (2010). The common neural basis of exerting self-control in multiple domains. To appear in Y. Trope, R. Hassin, & K. N. Ochsner (eds.), *Self-control* (pp. 141–160). Oxford University Press.

Critchley, H., Daly, E., Phillips, M., et al. (2000). Explicit and implicit neural mechanisms for processing of social information from facial expressions: A functional magnetic resonance imaging study. *Human Brain Mapping, 9*, 93–105.

Crockett, M. J., Lieberman, M. D., & Tabibnia, G. (unpublished manuscript). Affect labeling attenuates skin conductance responses to emotionally evocative faces in neurotics. University of California, Los Angeles.

Cunningham, W. A., Espinet, S. D., Deyoung, C. G., & Zelazo, P. D. (2006). Attitudes to the right- and left: Frontal ERT asymmetries associated with stimulus valence and processing goals. *NeuroImage, 28*, 827–834.

Cunningham, W. A., Johnson, M. K., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2003). Neural components of social evaluation. *Journal of Personality and Social Psychology, 85*, 639–649.

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological Science, 15*, 806–813.

Cunningham, W. A., Raye, C. L., & Johnson, M. K. (2004). Implicit and explicit evaluation: fMRI correlates of valence, emotional intensity, and control in the processing of attitudes. *Journal of Cognitive Neuroscience, 16*, 1717–1729.

Dandoy, A. C., & Goldstein, A. G. (1990). The use of cognitive appraisal to reduce stress reactions: A replication. *Journal of Social Behavior and Personality, 5*, 275–285.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81*, 800–814.

Davey G. C. L. (1992): Classical conditioning and the acquisition of human fears and phobias: A review and synthesis of the literature. *Advances in Behavioral Research and Therapy, 14*, 29–66.

Del-Ben, C. M., Deakin, J. F. W., Mckie, S., et al. (2005). The effect of citalopram pretreatment on neuroanl responses to neuropsychological tasks in normal volunteers: An fMRI study. *Neuropsychopharmacology, 30*, 1724–1734.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*, 5–18.

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology, 82*, 835–848.

Dolcos, F., LaBar, K. S., & Cabeza, R. (2004). Dissociable effects of arousal and valence on prefrontal activity indexing emotional evaluation and subsequent memory: An event-related fMRI study. *Neuroimage, 23*, 64–74.

Durston, S., Mulder, M., Casey, B. J., Ziermans, T., & van Engeland, H. (2006). Activation in ventral prefrontal cortex is sensitive to genetic vulnerability for attention-deficit hyperactivity disorder. *Biological Psychiatry, 60*, 1062–1070.

Eisenberger, N. I. (2006). Identifying the neural correlates underlying social pain: Implications for developmental processes. *Human Development, 49*, 273–293.

Eisenberger, N. I., & Lieberman, M. D. (2004). Why rejection hurts: A common neural alarm system for physical and social pain. *Trends in Cognitive Sciences, 8*, 294–300.

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science, 302*, 290–292.

Eisenberger, N. I., Way, B., Taylor, S. E., Welch, W. T., Lieberman, M. D. (2007). Understanding genetic risk for aggression: Clues from the brain's response to social exclusion. *Biological Psychiatry, 61*, 1100–1108.

Fan, J., Flombaum, J. I., McCandiss, B. D., Thomas, K. M., & Posner, M. I. (2003). Cognitive and brain consequences of conflict. *NeuroImage, 18*, 42–57.

Foa, E. B., & Kozak, M. J. (1986). Emotional processing of fear: Exposure to corrective information. *Psychology Bulletin, 99*, 20–35.

Francis, M. E., & Pennebaker, J. W. (1992). Putting stress into words. The impact of writing on physiological, absentee, and self-reported emotional well-being measures. *American Journal of Health Promotion, 6*, 280–287.

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favortism. *Journal of Personality and Social Psychology, 78*, 708–724.

Garavan, H., Ross, T. J., & Stein, E. A. (1999). Right hemispheric dominance of inhibitory control: An event-related functional MRI study. *Proceedings of the National Academy of Science, 96*, 8301–8306.

Ghashghaei, H. T., & Barbas, H. (2002). Pathways for emotion: Interactions of prefrontal and anterior temporal pathways in the amygdala of the rhesus monkey. *Neuroscience, 115*, 1261–1279.

Gilbert, D. T. (1989). Thinking lightly about others. Automatic components of the social inference process. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended Thought* (pp. 189–211). New York: Guilford.

Goel, V., & Dolan, R. J. (2003). Reciprocal neural response within lateral and ventral medial prefrontal cortex during hot and cold reasoning. *Neuroimage, 20*, 4314–4321.

Gorno-Tempini, M. L., Pradelli, S., Serafini, M., et al. (2001). Explicit and incidental facial expression processing. An fMRI study. *NeuroImage, 14*, 465–473.

Greenberg, M. A., Wortman, C. B., & Stone, A. A. (1996). Emotional expression and physical health: revising traumatic memories or fostering self-regulation? *Journal of Personality and Social Psychology, 71*, 588–602.

Gross, J. J. (1998). Antecedent- and response-focused emotion regulation: Divergent consequences for experience, expression, and physiology. *Journal of Personality and Social Psychology, 74*, 227–237.

Gur, R. C., Schroeder, L., Turner, T., et al. (2002). Brain activation during facial emotion processing. *NeuroImage, 16*, 651–662.

Hariri, A. R., Bookheimer, S. Y., & Mazziotta, J. C. (2000). Modulating emotional response: Effects of a neocortical network on the limbic system. *NeuroReport, 11*, 43–48.

Hazeltine, E., Bunge, S. A., Scanlon, M. D., & Gabrieli, J. D. E. (2003). Material-dependent and material-indpendent selection processes in the frontal and parietal lobes: An event-related fMRI investigation of response competition. *Neuropsychologia*, 41, 1208–1217.

Hazeltine, E., Poldrack, R., & Gabrieli, J. D. E. (2000). Neural activation during response competition. *Journal of Cognitive Neuroscience, 12*, 118–129.

Hemenover, S. H. (2003). The good, the bad, and the healthy: Impacts of emotional disclosure of trauma on resilient self-concept and psychological distress. *Personality and Social Psychology Bulletin, 29*, 1236–1244.

Horn, N. R., Dolan, M. Elliott, R., Deakin, J. F. W., & Woodruff, P. W. R. (2003). Response inhibition and impulsivity: An fMRI study. *Neuropsychologia, 41*, 1959–1966.

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.

Jahn, K., Deutschlander, A., Stephan, T., Strupp, M., Wiesmann, M., & Brandt, T. (2004). Brain activation patterns during imagined stance and locomotion in functional magnetic resonance imaging. *Neuroimage, 22*, 1722–1731.

James, W. (1890/1950). *The Principles of Psychology*. New York: Dover.

Kalisch, R., Wiech, K., Critchley, H. D., et al. (2005). Anxiety reduction through detachment: Subjective, physiological, and neural effects. *Journal of Cognitive Neuroscience, 17*, 874–883.

Kawashima, R., Satoh, K., Itoh, H., et al. (1996). Functional anatomy of GO/NO-GO discrimination and response selection—a PET study in man. *Brain Research, 728*, 79–89.

Kemmotsu, N., Villalobos, M. E., Gaffrey, M. S., Courchesne, E., & Muller, R. A. (2005). Activity and functional connectivity of inferior frontal cortex associated with response conflict. *Cognitive Brain Research, 24*, 335–342.

Konishi, S., Nakajima, K., Uchida, I., Kikyo, H., Kameyama, M., & Miyashita, Y. (1999). Common inhibitory mechanism in human inferior prefrontal cortex revealed by event-related functional MRI. *Brain, 122*, 981–999.

Kuchinke, L., Jacobs, A. M., Grubich, C., Vo, M. L., Conrad, M., & Hermann, M. (2005). Incidental effects of emotional valence in single word processing. An fMRI study. *NeuroImage, 28*, 1022–1032.

Lane, R. D., Fink, G. R., Chau, P. M.-L., & Dolan, R. J. (1997). Neural activation during selective attention to subjective emotional responses. *NeuroReport, 8*, 3969–3972.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1999). *International Affective Picture System (IAPS): Instruction Manual and Affective Ratings*. Gainsville: University of Florida, The Center for Research in Psychophysiology.

Lange, K., Williams, L. M., Young, A. W., et al. (2003). Task instructions modulate neural response to fearful facial expressions. *Biological Psychiatry, 53*, 226–232.

Lazarus, R. S., & Alfert, E. (1964). Short-circuiting of threat by experimentally altering cognitive appraisal. *Journal of Abnormal and Social Psychology, 69*, 195–205.

Lazarus, R. S., Opton, E. M., Nomikos, M. S., & Rankin, N. O. (1965). The principle of short-circuiting of threat: Further evidence. *Journal of Personality, 33*, 622–635.

LeDoux, J. E. (1996). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon & Schuster.

Lepore, S. J., & Smyth, J. M. (2002). *The Writing Cure: How Expressive Writing Promotes Health and Emotional Well-being*. Washington D.C.: American Psychological Association.

Leung, H., Skudlarski, P., Gatenby, J. C., Peterson, B. S., & Gore, J. C. (2000). An event-related functional MRI study of the Stroop color word interference task. *Cerebral Cortex, 10*, 552–560.

Levesque, J., Eugene, F., Joanette, Y., et al. (2003). Neural circuitry underlying voluntary suppression of sadness. *Biological Psychiatry, 53*, 502–510.

Liddle, P. F., Kiehl, K. A., & Smith, A. M. (2001). Event-related fMRI study of response inhibition. *Human Brain Mapping, 12*, 100–109.

Lieberman, M. D. (2007). The X- and C-sytems: The neural basis of reflexive and reflective social cognition. In E. Harmon-Jones & P. Winkelman (eds.), *Fundamentals of Social Neuroscience* (pp. 290–315). New York: Guilford.

Lieberman, M. D. (2010). Social cognitive neuroscience. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of Social Psychology* (5th ed., pp. 143–193). New York, NY: McGraw-Hill.

Lieberman, M. D., Eisenberber, N. I., & Crockett, M. J. (unpublished manuscript). An fMRI study of automatic behavior: Comparing ideomotor and disruption accounts. University of California, Los Angeles.

Lieberman, M. D., Eisengerger, N. I., Crockett, M. J., Tom, S. M., Pfeifer, J. H., & Way, B. M. (2007). Putting feelings into words: Affect labeling disrupts amygdala activity to affective stimuli. *Psychological Science, 18*, 421–428.

Lieberman, M. D., Hariri, A., Jarcho, J. J., Eisenberger, N. I., & Bookheimer, S. Y. (2005). An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. *Nature Neuroscience, 8*, 720–722.

Lieberman, M. D., Jarcho, J. M., Berman, S., et al. (2004). The neural correlates of placebo effects: A disruption account. *NeuroImage, 22*, 447–455.

Livingston, R. W., & Brewer, M. B. (2002). What are we really priming? Cue-based versus category-based processing of facial stimuli. *Journal of Personality and Social Psychology, 82*, 5–18.

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence on automatic racial prejudice. *Journal of Personality and Social Psychology, 81*, 842–855.

Luan Phan, K., Magalhaes, A., Ziemlewicz, T. J., Fitzgerald, D. A., Green, C., & Smith, W. (2005). Neural correlates of telling lies: A functional magnetic resonance imaging study at 4 tesla. *Academic Radiology, 12*, 164–172.

Lyubomirsky, S., Sousa, L., & Dickerhoof, R. (2006). The costs and benefits of writing, talking, and thinking about life's triumphs and defeats. *Journal of Personality and Social Psychology, 90*, 692–708.

Martin, J. H. (1996). *Neuroanatomy* (2nd ed.). Stamford, CT: Appleton & Lange.

Matthews, S. C., Simmons, A. N., Arce, E., & Paulus, M. P. (2005). Dissociation of inhibition from error processing using a parametric inhibitory task during functional magnetic resonance imaging. *Neuroreport, 16*, 755–760.

McDonald, A. J., Mascagni, F., & Guo, L. (1996). Projections of the medial and lateral prefrontal cortices to the amygdala: A *phaseolus vulgaris* leucoagglutinin study in the rat. *Neuroscience, 71*, 55–75.

McGuire, K. M. B., Greenberg, M. A., & Gevirtz, R. (2005). Autonomic effects of expressive writing in individuals with elevated blood pressure. *Journal of Health Psychology, 10*, 197–209.

Monahan, J. L., Murphy, S. T., & Zajonc, R. B. (2000). Subliminal mere exposure: Specific, general, and diffuse effects. *Psychological Science, 11*, 462–466.

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

Morris, J. S., DeGelder, B., Weiskrantz, L., & Dolan, R. J. (2001). Differential extrageniculostriate and amygdala responses to presentation of emotional faces in a cortically blind field. *Brain, 124*, 1241–1252.

Murphy, S. T., & Zajonc, R. B. (1993) Affect, cognition, and awareness: Affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology, 64*, 723–739.

Nakamura, K., Kawashima, R., Ito, K., et al. (1999). Activation of the right inferior frontal cortex during assessment of facial emotion. *Journal of Neurophysiology, 82*, 1610–1614.

Naliboff, B. D., Berman, S., Suyenobu, B., et al. (2006). Longitudinal change in perceptual and brain activation response to visceral stimuli in irritable bowel syndrome patients. *Gastroenterology, 131*, 352–365.

Narumoto, J., Yamada, H., Iidaka, T., et al. (2000). Brain regions involved in verbal or nonverbal aspects of facial emotion recognition. *NeuroReport, 11*, 2571–2576.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited capacity attention. *Journal of Experimental Psychology: General, 106*, 226–254.

Nolen-Hoeksema, S. (2000). The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. *Journal of Abnormal Psychology, 109*, 504–511.

Nomura, M., Iidaka, T., Kakehi, K., et al. (2003). Frontal lobe networks for effective processing of ambiguously expressed emotions in humans. *Neuroscience Letters, 348*, 113–116.

Nosek, B.A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice, 6*, 101–115.

Nunez, J. M., Casey, B. J., Egner, T., Hare, T., & Hirsch, J. (2005). Intentional false responding shares neural substrates with response conflict and cognitive control. *Neuroimage, 25*, 267–277.

Ochsner, K. N., Bunge, S. A., Gross, J. J., & Gabrieli, J. D. E. (2002). Rethinking feelings: An fMRI study of the cognitive regulation of emotion. *Journal of Cognitive Neuroscience, 14*, 1215–1229.

Ochsner, K. N., Ray, R. D., Cooper, J. C., et al. (2004). For better or for worse: Neural systems supporting the cognitive down- and up-regulation of negative emotion. *Neuroimage, 23*, 483–499.

Park, C. L., & Blumberg, C. J. (2002). Disclosing trauma through writing: Testing the meaning-making hypothesis. *Cognitive Therapy and Research, 26*, 597–616.

Pasley, B. N., Mayes, L. C., & Schultz, R. T. (2004). Subcortical discrimination of unperceived objects during binocular rivalry. *Neuron, 42*, 163–172.

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science, 8*, 162–166.

Pennebaker, J. W., & Beall, S. K. (1986). Confronting a traumatic event. Toward an understanding of inhibition and disease. *Journal of Abnormal Psychology, 95*, 274–281.

Petrovic, P., Dietrich, T., Fransson, P., Andersson, J., Carlsson, & Ingvar, M. (2005). Placebo in emotional processing—induced expectations of anxiety relief activate a generalized modulatory network. *Neuron, 46*, 957–969.

Petrovic, P., Kalso, E., Petersson, K. M., & Ingvar, M. (2002). Placebo and opioid analgesia imaging a shared neuronal network. *Science, 295*, 1737–1740.

Phelps, E. A., Delgado, M. R., Nearing, K. I., & LeDoux, J. E. (2004). Extinction learning in humans: Role of the amygdala and vmPFC. *Neuron, 43*, 897–905.

Quirk, G. J., Likhtik, E., Pelletier, J. G., & Pare, D. (2003). Stimulation of medial prefrontal cortex decreases the responsiveness of central amygdala output neurons. *Journal of Neuroscience, 23*, 8800–8807.

Richards, J. M., & Gross, J. J. (2000). Emotion regulation and memory: The cognitive costs of keeping one's cool. *Journal of Personality and Social Psychology, 79*, 410–424.

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: evidence for an accessibility model of emotional self-report. *Psychological Bulletin, 128*, 934–960.

Robinson, M. D., & Wilkowski, B. M. (2006). Loving, hating, vacillating: Agreeableness, implicit self-esteem, and neurotic conflict. *Journal of Personality, 74*, 935–977.

Rosenkranz, J. A., & Grace, A. A. (2002). Cellular mechanisms of infralimbic and prelimbic prefrontal cortical inhibition and dopaminergic modulation of basolateral amygdala neuronsin vivo. *Journal of Neuroscience, 22*, 324–337.

Royet, J. P., Plailly, J., Delon-Martin, C., Kareken, D. A., & Segebarth, C. (2003). fMRI of emotional responses to odors: Influence of hedonic valence and judgments, handedness, and gender. *NeuroImage, 20*, 713–728.

Rubia, K., Lee, F., Cleare, A. J., et al. (2005). Tryptophan depletion reduces right inferior prefrontal activation during response inhibition in fast, event-related fMRI. *Psychopharmacology, 179*, 791–803.

Rubia, K., Overmeyer, S., Taylor, E., et al. (1999). Hypofrontality in attention deficit hyperactivity disorder during higher-order motor control: A study with functional MRI. *American Journal of Psychiatry, 156*, 891–896.

Rubia, K., Smith, A. B., Brammer, M. J., & Taylor, E. (2003). Right inferior prefrontal cortex mediates response inhibition while mesial prefrontal cortex is responsible for error detection. *NeuroImage, 20*, 351–358.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). Theneural basis of economic decision-making in the Ultimatum Game. *Science, 300*, 1755–1758.

Schaefer, A., Collette, F., Philippot, P., et al. (2003). Neural correlates of "hot" and "cold" emotional processing: A multilevel approach to the functional anatomy of emotion. *NeuroImage, 18*, 938–949.

Shaw, P., Bramham, J., Lawrence, E. J., Morris, R., Baron-Cohen, S., & David, A. S. (2005). Differential effects of lesions of amygdala and prefrontal cortex on recognizing facial expressions of complex emotions. *Journal of Cognitive Neuroscience, 17*, 1410–1419.

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.

Smyth, J. M., Stone, A. A., & Hurewitz, A. (1999). Effects of writing about stressful experiences on symptom reduction in patients with asthma or rheumatoid arthritis. A randomized trial. *Journal of the American Medical Association, 281*, 1304–1309.

Speisman, J. C., Lazarus, R. S., Mordkoff, A. M., & Davison, L. A. (1964). Experimental reduction of stress based on ego-defense theory. *Journal of Abnormal and Social Psychology, 68*, 367–380.

Spence, S. A., Farrow, T. F. D., Herford, A. E., Wilkinson, I. D., Zheng, Y., & Woodruff, P. W. R. (2001). Behavioural and functional anatomical correlates of deception in humans. *Neuroreport, 13*, 2849–2852.

Spinoza, B. (1675/2000). *Ethics*. New York: Oxford University Press.

Stanton, A. L., Danoff-Burg, S., Sworowski, L. A., et al. (2002). Randomized, controlled trial of written emotional expression and benefit finding in breast cancer patients. *Journal of Clinical Oncology, 15*, 4160–4168.

Tabibnia, G., Lieberman, M. D., & Craske, M. (2008a). The lasting effect of words on feelings: Words may facilitate exposure effects to threatening images. *Emotion, 8*, 307–317.

Tabibnia, G., Satpute, A. B., & Lieberman, M. D. (2008b). The sunny side of fairness: Preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychological Science, 19*, 339–347.

Taylor, S. F., Phan, K. L., Decker, L. R., & Liberzon, I. (2003). Subjective rating of emotionally salient stimuli modulates neural activity. *NeuroImage, 18*, 650–659.

Vogt, B. A. & Pandya, D. N. (1987). Cingulate cortex of the rhesus monkey: II. Cortical afferents. *Journal of Comparative Neurology, 262*, 271–289.

Vollm, B., Richardson, P., McKie, S., Elliott, R., Deakin, J. F. W., & Anderson, I. M. (2006). Serotonergic modulation of neuronal responses to behavioural inhibition and reinforcing stimuli: An fMRI study in healthy volunteers. *European Journal of Neuroscience, 23*, 552–560.

Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2001). Effects of attention and emotion on face processing the human brain: An event-related fMRI study. *Neuron, 30*, 829–841.

Wager, T. D., Rilling, J. K., Smith, E. E., et al. (2004). Placebo-induced changes in fMRI in the anticipation and experience of pain. *Science, 303*, 1162–1167.

Wegner, D. M., Erber, R., & Zanakos, S. (1993). Ironic processes in the mental control of mood and mood-related thought. *Journal of Personality and Social Psychology, 65*, 1093–1104.

Wegner, D. M., Shortt, J. W., Blake, A. W., & Page, M. S. (1990). The suppression of exciting thoughts. *Journal of Personality and Social Psychology, 58*, 409–418.

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.

Wildgruber, D., Hertrich, I., Riccker, A., et al. (2004). Distinct frontal regions subserve evaluation of linguistic and emotional aspects of speech intonation. *Cerebral Cortex, 14*, 1384–1389.

Wildgruber, D., Riecker, A., Hertrich, I., et al. (2005). Identification of emotional intonation evaluated by fMRI. *NeuroImage, 24*, 1233–1241.

Williams, K. D. (2007). Ostracism. *Annual Review of Psychology, 58*, 425–452.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology, 60*, 181–192.

Wilson, T. D., Lisle, D. J., Schooler, J. W., Hodges, S. D., Klaaren, K. J., & LaFleur, S. J. (1993). Introspecting about reasons can reduce post-choice satisfaction. *Personality and Social Psychology Bulletin, 19*, 331–339.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated attitudes. *Journal of Personality and Social Psychology, 81*, 815–827.

CHAPTER 14
Emotion in Social Neuroscience

Elizabeth A. Phelps

Perhaps more than other topics in the emerging field of social neuroscience, *emotion* spans the many discrete disciplines that comprise the study of human behavior. This is in part because research on emotion does not necessarily entail a social interaction or social process but is an important component of all aspects of behavior. Traditions within psychology led to the study of emotion primarily being investigated by researchers specializing in the social or clinical domains. This was not because emotion is unimportant in other domains of the study of human behavior, such as cognition or economics, but rather because the traditional approach taken within these disciplines generally excluded its consideration. For example, the field of cognitive psychology was heavily influenced by the computer metaphor or the "information processing" approach (Miller, 2003). Viewing human thought as analogous to the processing of information by a computer did not encourage a consideration of emotion. Similarly, the study of economics was—and still is—heavily influenced by rational choice theory, which fails to consider emotion as an important factor in decision making (Kahneman, 2003). Because of these disciplinary approaches, the experimental study of human emotion has been explored primarily by social psychologists, who have no such theoretical or historical constraints. In contrast, the study of behavioral neuroscience, which for ethical and methodological reasons was historically conducted with nonhuman animals, has consistently struggled with the concept of motivation, an important component of emotion. To compel animals to engage in behaviors for study, they had to be motivated. Emotion and motivation have been key components of animal studies of behavior and neuroscience.

For these reasons, as techniques to study the human brain have developed, human neuroscience studies of emotion fall squarely within the domain of social neuroscience. Although cognitive neuroscientists and neuro-economists are increasingly investigating emotion, our psychological models of human emotion are drawn from social psychology and our neural models are drawn from behavioral neuroscience. The social neuroscience approach of combining insights across these disciplines is resulting in neural models of human emotion that are informed by both animal models of the brain and a long history of psychological theory. Recent investigations of the role of emotion in other fields of human neuroscience rely on these emerging social neuroscience models. The chapters in this section represent a broad range of approaches and questions in the social neuroscience emotion. By doing this, they demonstrate how the interdisciplinary approach of social neuroscience can provide insight into important experimental and theoretical questions across a number of disciplines of human behavior.

The chapter by Harmon-Jones and Harmon-Jones demonstrates how neuroscience data can inform theoretical debates about the structure of emotion. It has long been recognized by philosophers and psychologists that emotion is not a unitary construct but, rather, a compilation of processes that share some common principles. The theoretical challenge has been how to appropriately characterize the range of functions that fall under the rubric "emotion." One approach has been to characterize dimensions of emotion that capture a subset of emotional experience. One dimension is valence, which highlights whether the experience is positive or negative. Studies of the validity of valence as a primary dimension of emotion have generally relied on self-report measures of emotional experience. Another dimension is motivational tendency. Some emotional responses, such as sadness or fear, cause one to withdraw from a situation, whereas others cause one to approach, such as happiness or surprise. It has been suggested that valence and motivation represent a single dimension of emotion, in that most positive emotions cause one to approach and most negative emotions cause one to withdraw. The problem with this unified dimensional approach is that it fails to capture anger. Anger is an emotion most people report as negative, yet it results in a motivational tendency to approach. Because of anger, it is clear that the unified dimensional approach does not accurately represent the range of emotional experience. To date, psychological and philosophical investigations have not produced clear data suggesting one dimensional approach captures experience better than the other. In their chapter, Harmon-Jones and Harmon-Jones explore how neural data might inform this debate.

Some of the earliest research in human affective neuroscience used EEG to examine electrical activity from the scalp to examine differences in hemispheric processing of emotion states and traits (e.g., Davidson & Fox, 1982). This research led to the general conclusion that positive affect traits and states are related to greater left relative to right frontal activity, whereas negative affect traits and states are related to relatively greater right than left frontal activity. Although this basic finding is robust and has been demonstrated across development and species, it has generally failed to consider whether this effect results from the valence or motivational dimension of emotion. By addressing this question, one can inform the theoretical debate about which dimensional approach may more fully capture human experience. In their review of a long series of studies, Harmon-Jones and Harmon-Jones demonstrate how anger, unlike other negative emotions, is represented by relatively greater left than right frontal EEG activity, consistent with the conclusion that the motivational dimension of emotion is captured by this cerebral asymmetry. Their studies not only explore correlations among reported states, traits, and brain activity but also experimental manipulations of anger that result in changes in state anger. They also conclusively demonstrate how the strength of the left frontal activity may be specifically linked to approach motivation by specifically manipulating the opportunity to act on the emotion. Their review explores how this cerebral asymmetry represents affect in social and nonsocial situations, suggesting, as in the other chapters, that these emotional responses are linked to social interaction but are not uniquely social. Finally, this chapter examines how data derived from other approaches examining the neural basis of emotion may, or may not, be consistent with their general conclusions.

Whereas the Harmon-Jones and Harmon-Jones chapter highlights how neuroscience tools can inform a psychological and philosophical debate about emotion, the chapter by Beer and Bhanji demonstrates how nuanced social psychological paradigms can inform our understanding of the function of a specific neural structure, especially as it relates to neuroeconomic theories of emotion and decision making. The brain structure they examine is the orbitofrontal cortex. Recent investigations of role of emotion in economic decision making have highlighted the role of the orbitofrontal cortex in the integration of emotion into economic choices (Damasio, 1994). This research, which resulted in the Somatic Marker hypothesis, suggests that bodily states—particularly arousal—play an important role in determining appropriate choices, especially those that are more automatic and less amenable to conscious reflection. It is suggested that the orbitofrontal cortex codes these somatic markers, resulting in impaired emotion and decisions in patients with orbitofrotnal cortex damage. Although there are only a few correlational studies supporting the Somatic Marker hypothesis of decision making, it has received wide recognition in neuro-economics, perhaps because emotion has not been widely studied in economic decision making and it is one of the few neural models characterizing the relation between emotion and choice.

In their chapter, Beer and Bhanji critically evaluate this hypothesis as it relates to the function of the orbitofrontal cortex. By examining the larger literature of orbitofrontal cortex function, they demonstrate that the relatively discrete role assigned to this region in economic decision making is not supported by the larger literature. Critically, patients with orbitofrontal cortex damage show few emotion deficits across a range of tasks. Rather these patients show very specific deficits in updating reinforcement contingencies in decision tasks, which is a confound that studies examining the Somatic Marker hypothesis failed to address. However, Beer and Bhanji do not argue that the orbitofrontal cortex plays no role in emotion and decision making, but rather, it has a more discrete and nuanced role. They review a series of studies using clever social psychological paradigms to show that the orbitofrontal cortex is involved in self-monitoring and self-insight. Impaired self-monitoring can lead to problems with emotion but does not necessarily. They also highlight how this self-monitoring deficit can specifically influence emotion's impact on economic decisions, whether it is helpful or hurtful, even if the emotional reaction is unimpaired. Finally, they review the limited means by which emotion has been investigated in neuro-economic research to date and suggest how future research might benefit from an expanded view of both emotion and neural systems, informed with the social neuroscience approach.

Beer and Bhanji emphasize how a social neuroscience perspective can inform research in another domain of human neuroscience—namely, neuro-economics. In contrast, the chapter by Parker, Kesek, and Cunningham examines a topic at the core of social psychology and shows how neural models can enhance social psychological theory. The topic they examine is the representation of attitudes. Research on attitudes over the last several decades has increasingly highlighted the complexity of evaluations and their expression. A relatively recent distinction that has emerged is the differentiation of automatic/implicit and controlled/explicit attitudes. In their chapter, Perker, Kesek, and Cunningham review this literature and argue that this simple dichotomous approach may not be sufficient.

In this chapter, Parker, Kesek, and Cunningham introduce the Iterative Reprocessing (IR) model of attitudes and link it to a proposed neural circuitry. This model distinguishes between attitude, which refers to preexisting, valenced information, and evaluation, which refers to the current state of the evaluative system. The IR model proposes that there are more automatic aspects of the expression of attitudes but that these are influenced by reflective processes they may represent more complex aspects of the current situation. These two processes interact in a repetitive, iterative manner to represent the current evaluative state. In support of the IR model, they draw on fMRI data examining the evaluation of stimuli and expression of attitudes. Specifically, they review data suggesting that the amygdala and insula may play a role in the more automatic evaluation of stimuli but that the prefrontal cortex may be involved in the construction of more complex evaluations. Importantly, they review studies suggesting that these neural systems may influence each other, reflecting IR in the formation of attitudes. Finally, they examine how these functions may emerge over development as the prefrontal cortex slowly matures. By using neuroscience data to inform social psychological theory, Parker, Kesek, and Cunningham highlight how the social neuroscience approach is valuable to our basic understanding of social processes.

The final chapter in this section by Lieberman examines the alteration of emotion, a topic of critical importance to the treatment of psychological disorders, as well as informative to other research domains of human behavior. Studies of emotion have often described it as an automatic response to a stimulus or event. However, more recent research has renewed interest in the idea that the generation of emotion is strongly influenced by the situation and interpretation of events. The chapter by Lieberman explores one means by which emotion may be changed, which is verbalizing the emotion and, as a result, reducing its impact or intensity. The idea that verbalizing emotions can alter them is a core component of psychotherapy. Lieberman reviews the psychological literature suggesting that emotion is altered by the Symbolic Processing of Affect (SPA), or putting feelings into words, and links this effect to a neural model. Specifically, his model suggests that the right ventrolateral prefrontal cortex (RVLPFC) plays a key role in inhibition in general, not just SPA. However, in situations where SPA dampens the emotional response, the RVLPFC has its influence by altering responses primarily in the amygdala, which in turn influences the affective response. By reviewing human neuroscience studies from a number of domains, Lieberman shows how the RVLPFC is involved in inhibition broadly, how it may influence the amygdala through its connectivity, and how a number of studies, including SPA paradigms, result in increased RVLPFC activation and decreased amygdala activation. In introducing his model, Lieberman cites studies from affective neuroscience, neuro-economics, clinical neuroscience, as well as social neuroscience. Because inhibition and affect are important not only in clinical psychology and talk therapy, but across all of these domains, Lieberman highlights the potential implications of this model in social and cognitive psychology, as well as its clinical implications.

In his chapter, Lieberman explicitly highlights the interdisciplinary nature of research on emotion, but taken together all of the chapters emphasize different benefits of the social neuroscience approach. Harmon-Jones and Harmon-Jones demonstrate how brain data can inform classic theoretical debates about the nature of emotion. Beer and Bhanji highlight both how social psychological paradigms can significantly enhance our understanding of brain function and inform the emerging field of neuro-economics. Parker, Kesek, and Cunningham demonstrate how neuroscience methods can provide insight into core social psychological questions. Finally, Lieberman demonstrates how looking across disciplines with a social neuroscience perspective can provide scientific support for a traditional clinical approach advocated over century ago by Freud.

The interdisciplinary nature of social neuroscience, and particularly the study of emotion, is emblematic of a larger shift in the study of human behavior. The first 100 years of the experimental study of human mental function resulted in dividing behavior into different domains, such as social, clinical, cognitive, and economic. At the time, this seemed like the only logical means to tackle its complexity. This resulted in discrete disciplines that rarely communicated, even if the questions being addressed overlapped. Recent methodological advances in human neuroscience, however, have created a common currency, which is now pulling these discrete disciplines back together. As research in social neuroscience and other disciplines of human neuroscience progresses, we will necessarily have more complex models of human social behavior. However, these models will hopefully more accurately reflect the detailed and nuanced influence of social and emotional factors on our mental lives.

References

Damasio AR. (1994). *Descartes Error: Emotion, Reason and the Human Brain*. New York: G.P. Putnam's Sons.

Davidson RJ, & Fox NA. (1982). Asymmetrical brain activity discriminates between positive and negative affective stimuli in human infants. *Science, 218*, 1235–1237.

Kahneman D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*, 697–720.

Miller GA. 2003. The cognitive revolution: A historical perspective. *Trends in Cognitive Science, 7*, 141–144.

This page intentionally left blank

PART IV

Navigating Social Life

This page intentionally left blank

CHAPTER 15
The Social Brain in Interactive Games

James K. Rilling

In this chapter, I discuss the advantages and disadvantages of an approach to social cognitive neuroscience that involves imaging brain function in subjects who are immersed in genuine social interactions. I also discuss what this approach can and cannot reveal about one of the fundamental questions in social neuroscience: whether the human brain has domain-specific neural systems that are specialized for social cognition.

Imaging studies of social interactions have emerged relatively recently within cognitive neuroscience. Many early fMRI studies presented subjects with face stimuli, given the obvious importance of faces in human social interactions. Typically, stimuli were static, two-dimensional pictures of faces that subjects were instructed to either passively view or judge on some attribute. Other studies examined face processing deficits in patients with damage to specific brain regions like the amygdala or the fusiform gyrus using similar types of stimuli (reviewed in Adolphs, 2001; Adolphs, 2003). Still others have attempted to probe social cognition by asking subjects to read stories or view cartoons and make judgments about these hypothetical scenarios. For example, the neural correlates of both mentalizing (Brunet et al., 2000; Fletcher et al., 1995; Gallagher et al., 2000; Vogeley et al., 2001) and moral reasoning (Greene et al., 2001; Greene et al., 2004; Moll et al., 2002) have been probed with this methodology. These studies have yielded valuable insights with respect to the neural underpinnings of human social cognition. However, for each, one can raise questions about the ecological validity of the stimuli. Does the pattern of brain activation in response to static, two dimensional face stimuli accurately reflect the brain's response to the dynamic, embodied faces that we encounter in everyday life? Is the pattern of brain activation in response to reasoning about hypothetical, fictitious scenarios the same as when grappling with real-life, consequential social problems?

One approach to improving the ecological validity of experiments in social cognitive neuroscience is to image brain function as subjects interact with other people in real social interactions. In recent years, a number of such studies have been conducted, and these are reviewed below.

Across the primate order, there is a positive correlation between relative neocortex size and the size of the social group to which an individual of a given species belongs (Dunbar, 1998). This observation has led to the "Social Brain Hypothesis," according to which the need to navigate complex social environments associated with larger groups selected for increased neocortical size throughout primate evolution. If this is true, then the function of this enlargement is enhanced social cognition, and we might well expect domain-specific neural circuitry for social cognition to have evolved in primates. Given how much larger the human neocortex is compared with other primates (Rilling & Insel, 1999), humans may be an extreme manifestation of this trend, and the human brain may be in large part a social cognitive organ.

An alternative possibility is that social cognition does not have its own dedicated neural circuitry and instead makes use of more basic systems designed for general purpose tasks. In principal, neuro-imaging investigations of social interactions could be used to adjudicate between these two competing hypotheses. If there are dedicated neural systems for social cognition, then interacting with a human partner should yield a different pattern of neural activation than interacting with a computer partner, even after standardizing the behavior of those partners. However, interpreting similarities would be more ambiguous. If computer interactions activate the same set of regions as human interactions, is it because a domain general network is being recruited, or is it because humans reflexively anthropomorphize computers and recruit specialized social cognition systems for these interactions as well? With these ambiguities in mind, I next review published neuro-imaging studies of social interactions and discuss what they reveal about the domain-specificity of social cognition, as well as what they reveal about social neuroscience more generally.

Probing the neural correlates of social emotions

One goal of social cognitive neuroscience should be to map the neural correlates of the social emotions. Interactive tasks are particularly useful in this effort because of their effectiveness in provoking social emotions. For example, Eisenberger et al. (Eisenberger et al., 2003) investigated the neural correlates of social exclusion by scanning subjects as they played a virtual ball-tossing game from which they were ultimately excluded by their partners. They found that subjects' self-reported distress in response to exclusion was positively correlated with activation in a region of the anterior cingulate cortex (ACC) known to be involved in the affective response to painful physical stimuli (Fig. 15–1). Based on this result, it is suggested that social and physical pain share a common neural basis. This conclusion is consistent with the idea that social cognition draws on domain-general neural systems that initially evolved for more basic functions. The authors also report a negative correlation between self-reported distress and activation in the ventrolateral prefrontal cortex (VLPFC), as well as a negative correlation between ACC and VLPFC activity, suggesting that VLPFC may regulate the distress of social exclusion by disrupting ACC activity.

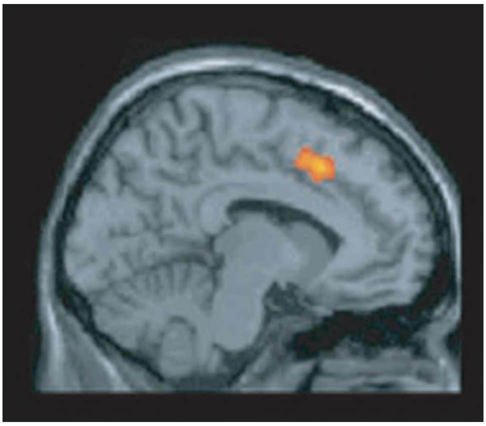


Fig. 15–1 Region of anterior cingulate cortex showing a positive correlation with self-reported distress in response to social exclusion in a virtual ball tossing game. From Eisenberger NI, Lieberman MD, & Williams KD. (2003): Does rejection hurt? An FMRI study of social exclusion. *Science 302,* 290–292. Reprinted with permission.

Another example of interactive tasks being used to probe the neural correlates of social emotions is a study by Sanfey et al. (Sanfey et al., 2003), in which subjects were scanned with fMRI as they received fair and unfair offers from partners in an ultimatum game (UG). In the UG game, two subjects—say player A and player B—are asked to split a sum of money. Player A is asked to propose how to divide the sum. Player B can either accept or reject the proposal. If player B accepts, the money is divided as specified by player A. On the other hand, if the player B rejects the offer, neither player receives any money. The indignation one feels upon being made an unfair offer often motivates the irrational decision to reject the offer and receive nothing rather than something. In the study by Sanfey et al., the subject in the scanner was always in the role of player B and received an offer from each of 10 players met previously. In reality, offers were predetermined by a computer algorithm such that five of them were fair (5:5 split) and five were unfair (two 7:3, two 8:2, and one 9:1). Receiving an unfair offer was associated with activation in three brain regions: anterior insular cortex, dorsolateral prefrontal cortex (DLPFC), and anterior cingulate cortex (ACC) (Fig. 15–2). When activation within the anterior insula was stronger than activation in DLPFC, subjects were more likely to reject than accept unfair offers, whereas subjects were more likely to accept unfair offers when DLPFC activation exceeded anterior insula activation. Based on these results, a model was proposed in which the emotional response to an unfair offer, represented in anterior insula, motivated rejection of the offer, whereas activation in DLPFC maintained the rational goal of maximizing earnings by accepting the offer. The conflict between these two competing motives is represented in the ACC, a region implicated in cognitive conflict (Botvinick et al., 1999; MacDonald et al., 2000).

In addition to receiving fair and unfair offers from alleged human partners, subjects received the exact same series of offers from alleged computer partners. Unfair offers from computer partners were also associated with activation within anterior insula, suggesting that either the anterior insula is part of a domain-general neural system that responds to aversive stimuli in general or that human subjects were anthropomorphizing their computer partners. In this case, both explanations likely apply. The insula is, in fact, known to activate to a wide range of aversive stimuli (Sanfey et al., 2003), and the fact that unfair offers from computer partners were occasionally rejected raises the possibility that subjects were anthropomorphizing their computer partners. It should also be noted that anterior insula activation in response to unfair offers was stronger for human than computer partners, suggesting that humans are a more potent stimulus for this neural system.

Singer et al. (Singer et al., 2004; Singer et al., 2006) probed the neural correlates of empathy and their modulation by the reputation of the person with whom one empathized. Interactive tasks were used to establish these reputations. In part one of their experiment, subjects played a sequential Prisoner's Dilemma Game with each of two human confederates. Subjects were always the first movers in the game, a role that involved choosing to either send 10 monetary units to their partner, in which case the points would be tripled and the partner would have the option of returning a portion, or to keep the 10 monetary units to themselves. One of the confederates played fairly and returned large amounts of money to the player, whereas the other played unfairly, returning small

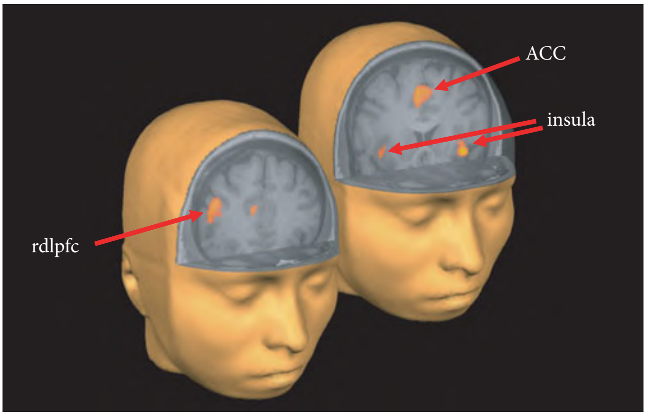


Fig. 15–2 Activated brain regions in response to receiving an unfair (vs. fair) offer in the Ultimatum Game.

amounts of money. In part two of the experiment, brain activity was measured with fMRI in these same subjects as they witnessed the fair and unfair confederates receive painful electric shocks. Both male and female subjects exhibited empathy-related activation in pain-related brain areas like the fronto-insular and ACC. However, in males, the magnitude of these empathy-related responses during pain was significantly attenuated when the unfair confederates were being shocked (Fig. 15–3). Men also showed stronger activation in the ventral striatum, a putative reward processing region, when unfair as compared with fair confederates received painful shocks, and the magnitude of this activation scaled with the reported desire for revenge expressed in post-experiment questionnaires. These data are consistent with the possibility that men derive pleasure from punishing unfair social partners, a finding echoed in another study discussed below.

The neural correlates of positive and negative social emotions have also been explored in the context of an iterated Prisoner's Dilemma (PD) game (Rilling et al., 2002). In this study, subjects were scanned with fMRI as they played several consecutive rounds of a simultaneous choice PD game with real or assumed human partners who were outside the scanner. The iterated PD game models relationships based on reciprocal altruism, or the reciprocal exchange of favors. In the PD game, two players simultaneously and independently choose to either cooperate with each other or not (i.e., defection) and receive a payoff that depends on the interaction of their two choices. The largest payoff occurs when the player defects and the partner cooperates (DC = $3). The second largest payoff is for mutual cooperation (CC = $2), followed by mutual defection (DD = $1), and finally player cooperation combined with partner defection (CD = $0). Each outcome corresponds to a different outcome of a social interaction and typically elicits a different set of social emotions. Mutual cooperation is often associated with friendship, love, trust, or obligation; mutual defection is associated with feelings of rejection and hatred; and cooperation by one and defection by the other typically results in the cooperator feeling anger or indignation and in the defector feeling anxiety, guilt, or elation from successfully exploiting the partner to their advantage. CC outcomes were associated with activation in anteroventral striatum and orbitofrontal cortex, and CD outcomes were associated with deactivation in these same areas. This finding recurred in a single-shot version of the PD game where subjects played just one round of the game with each of 10 different assumed human partners (Rilling et al., 2004; Fig. 15–4). These regions are targets of mesencepahlic dopamine projections thought to be involved in representing rewards and calculating reward prediction errors (Montague et al., 1996; Schultz, 2002). The results are therefore consistent with the hypothesis that mesolimbic dopamine projection sites carry information about errors in reward prediction that allow us to learn who can and cannot be trusted to reciprocate favors.

Like the study by Sanfey and colleagues described above, these PD studies included

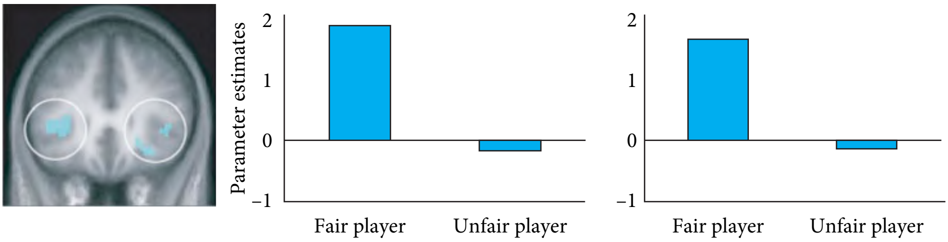


Fig. 15–3 Activation within left and right frontoinsular cortices in male subjects as they observed fair and unfair confederates receiving painful electric shocks. From Singer T, Seymour B, O'Doherty JP, Stephan KE, Dolan RJ, & Frith CD. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature 439*, 466–469. Reprinted with permission.

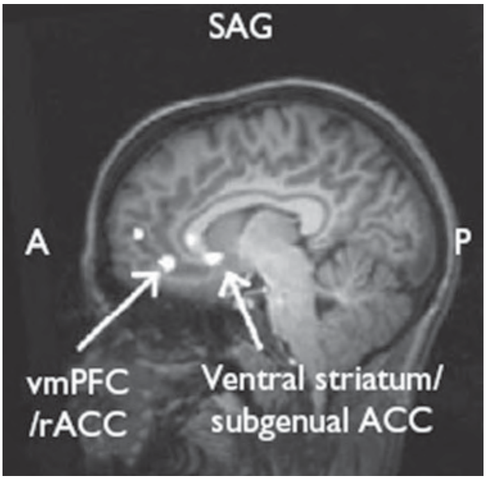


Fig. 15–4 Areas that activate in response to reciprocated cooperation and deactivate in response to unreciprocated cooperation in a single-shot PD game. From Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, & Cohen JD. (2004). Opposing BOLD responses to reciprocated and unreciprocted altruism in putative reward pathways, *NeuroReport 15*, 2539–2543.

trials with putative computer partners for comparison with putative human partners. CD outcomes with computer partners were also associated with negative BOLD responses in the striatum, but CC outcomes with computer partners were not associated with positive BOLD responses in the striatum, suggesting that striatal responses to mutual cooperation are specific to interactions with human partners.

In contrast to Singer et al., who used a trust game (i.e., sequential PD game) outside the scanner to establish confederate reputations for later scanning, Delgado et al. (2005) scanned subjects with fMRI while playing an iterated trust game. Subjects were given $1, which they could choose to either keep or transfer to a partner, in which case it would triple in value and the partner would have the option of returning either half (share) or none (keep). The contrast between the partner opting to share versus keep revealed activation in the ventral caudate in a similar location to that observed by Rilling et al. for the contrast between reciprocated and unreciprocated cooperation (Rilling et al., 2002; Fig. 15–5). However,

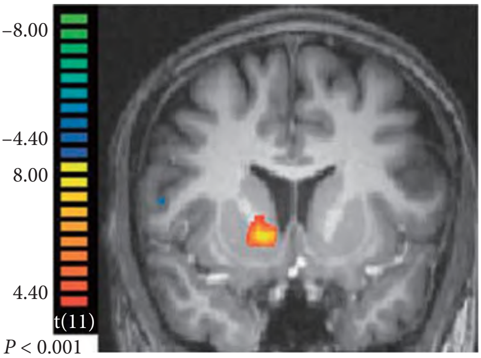


Fig. 15–5 Ventral caudate activation for the contrast between partner reciprocation and non-reciprocation in an iterated trust game. Reprinted by permission from Delgado MR, Frank RH, & Phelps EA. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience 8*, 1611–1618.

Delgado et al. added an interesting manipulation in which subjects played the game with each of three different partners who had pre-existing reputations of praiseworthy, neutral, or morally suspect. Despite the fact that all three of these partners behaved identically (sharing 50% of the time), subjects chose to transfer money to the praiseworthy partner more frequently. Further, the neural response for the share versus keep contrast differed as a function of the moral reputation of the partner. When playing with either bad or neutral partners, the share versus keep contrast yielded significant activation in the ventral caudate. However, there was no significant difference in ventral caudate activation between share and keep when playing with praiseworthy partners. The authors interpret these findings to suggest that prior moral perceptions can diminish reliance on feedback mechanisms in the neural circuitry of trial and error learning.

Finally, Decety et al. (Decety et al., 2004) scanned subjects while playing a different type of interactive game in which subjects either worked with or against a partner who was trying to reproduce a specified pattern on a computerized gameboard. As in the iterated PD study discussed above, cooperating with the partner was associated with activation in

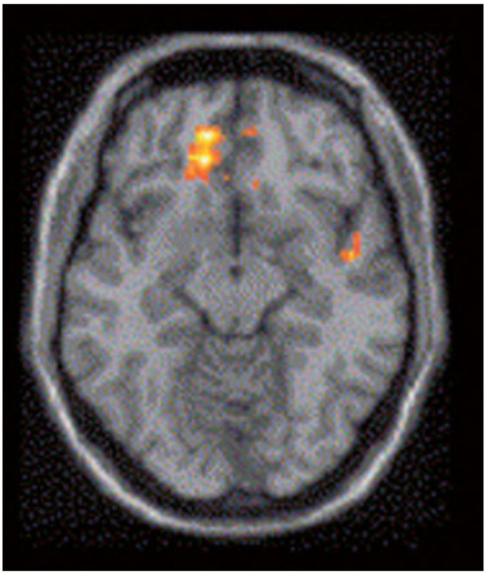


Fig. 15–6 Orbitofrontal cortex activation for the contrast between cooperating and competing with a partner in a computerized board game. Decety J, Jackson PL, Sommerville JA, Chaminade T, & Meltzoff AN (2004). The neural bases of cooperation and competition: an fMRI investigation, *Neuroimage* 23, 744–751. Reprinted with permission from Elsevier.

medial orbitofrontal cortex, which the authors interpreted to suggest that cooperation is more rewarding than competing (Fig. 15–6).

More potent social stimuli

In addition to provoking social emotions, interactive tasks can be used as more potent stimuli for provoking social cognition. For example, previous studies have investigated the neural correlates of Theory of Mind (ToM) by asking subjects to make inferences about characters in hypothetical scenarios (Brunet et al., 2000; Fletcher et al., 1995; Gallagher et al., 2000; Saxe & Kanwisher, 2003; Vogeley et al., 2001). But people likely invoke their mentalizing abilities most when attempting to understand the minds of other individuals with whom they are directly interacting, especially when those interactions have real consequences for them. Thus, at least three groups have attempted to probe the neural correlates of ToM by immersing subjects in meaningful social interactions. Gallagher et al. (Gallagher & Frith, 2003) imaged subjects with PET as they played the game "stone, paper, scissors" with an assumed human partner outside the scanner and with an assumed computer partner. To control for partner behavior in these two conditions, a random sequence of choices was surreptitiously inserted during the PET scanning epoch in each condition. Playing with an assumed human partner was associated with stronger activation within anterior paracingulate cortex compared to playing with an assumed computer partner (Fig. 15–7), suggesting that this brain region may be specialized for making inferences about the mental states of other humans. This result was consistent with earlier non-interactive ToM imaging studies that had consistently reported anterior paracingulate cortex as one of three brain areas reliably activated by ToM tasks (Gallagher & Frith, 2003). Thus, this study did not reveal an obvious advantage of the interactive task paradigm.

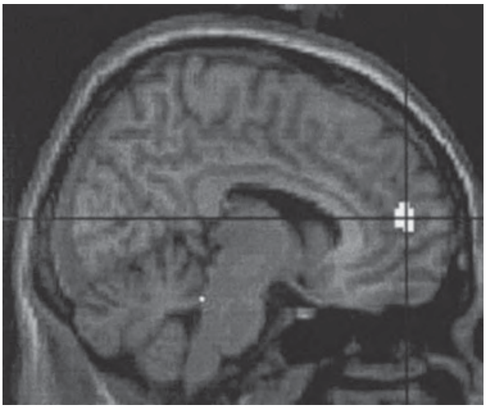


Fig. 15–7 Anterior paracingulate activation for the contrast between playing with human and computer partners in the game "stone, paper, scissors." Gallagher H, Jack A, Roepstorff A, & Frith CD. (2002). Imaging the intentional stance in a competitive game. *Neuroimage 16*, 814–821. Reprinted with permission from Elsevier.

A subsequent study by Rilling et al. (2004) examined brain regions that were activated in response to partner feedback in both the UG and PD game. In this experiment, subjects met a group of 10 behavioral confederates prior to scanning to reinforce their belief that they were interacting with real people. Additionally, digital photographs of confederates were presented to subjects while they played the game in the scanner. In both games, partner feedback revealed something about the partner's intentions and was therefore expected to invoke ToM processing. Specifically, in the UG, revelation of the partner's offer revealed whether the partner was generous or greedy. In the PD game, revelation of the partner's decision to cooperate or defect revealed whether the partner was cooperative or selfish. In contrast to non-interactive ToM studies in which subjects reasoned about fictitious scenarios, in this experiment subjects' interactions with their partners were consequential because they were compensated as a function of the actual game outcomes, and most subjects' motivation for participating was in fact monetary compensation. In both tasks, classic ToM areas, such as the anterior paracingulate cortex and the right posterior superior temporal sulcus, were activated (Fig. 15–8). However, several additional areas that have not been previously reported in ToM studies were also activated, including mid-superior temporal sulcus (mid-STS), posterior cingulate and precuneus, and hippocampus.

In monkeys, mid-STS contains neurons tuned to facial expressions, direction of eye gaze, and purposeful body movements of other monkeys and may be involved more generally with detecting the intentions of other social beings (Adolphs, 2001; Allison et al., 2000). A recent fMRI study contrasting activity when subjects watched movies of a lone actor versus watching a video of two people interacting (Iacoboni et al., 2004) found activation along the anterior-posterior extent of the STS for the interactive condition, prompting the authors to suggest that the anterior STS activations may have been driven by the use of complex stimuli that were closer to real-life situations, as would also be the case in the present study. As for the other activations, posterior cingulate activation has consistently been associated with emotionally salient stimuli (Maddock, 1999) and may relate to emotional arousal in response to receiving feedback from an assumed human partner in this experiment. Finally, hippocampus has been linked with episodic memory encoding (Squire & Zola, 1996; Zola et al., 2000), suggesting that subjects may be encoding the identity of cooperative and noncooperative partners. Thus, some of the activated

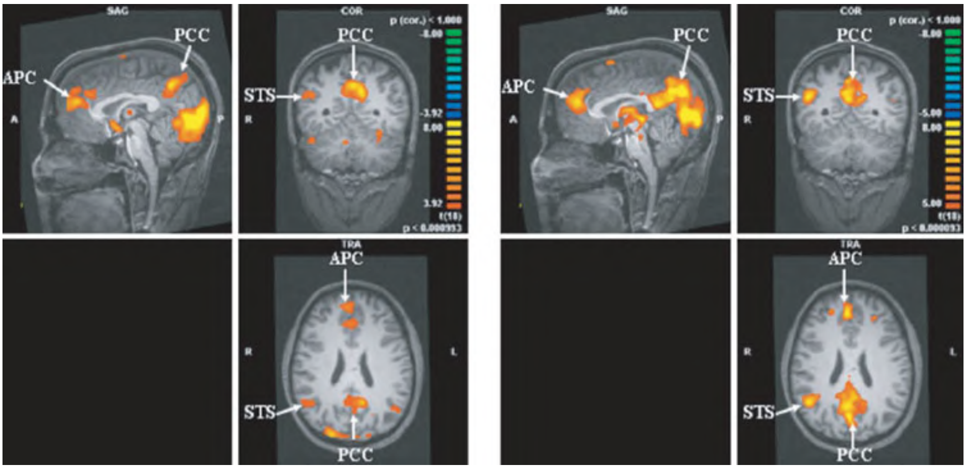


Fig. 15–8 Activations within classic theory of mind regions in response to receiving feedback from partners in both Ultimatum (left) and PD (right) games. From Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, & Cohen JD. (2004). Opposing BOLD responses to reciprocated and unreciprocted altruism in putative reward pathways. *NeuroReport 15*, 2539–2543.

regions are not involved in ToM processing *per se* but in related processes that reliably accompany ToM processing when subjects are immersed in genuine social interactions. This study therefore provides a more complete picture of brain regions that are likely to be engaged when subjects engage ToM processing in the context of everyday social interactions.

Once again, both human and computer partners were included in this study. Receiving feedback from a computer partner was associated with activation in anterior paracingulate cortex, right posterior STS, and posterior cingulate/precuneus; however, in each case, the activations were weaker compared with receiving feedback from human partners. In contrast to receiving feedback from human partners, neither mid-STS nor hippocampus were significantly activated when receiving feedback from computer partners, raising the possibility of a specialized neural system for processing feedback from human partners in social interactions.

Finally, in the computerized board game study by Decety et al. mentioned above, stronger activation was observed in anterior paracingulate cortex when subjects were competing with their partners compared to when they were cooperating with them, which the authors suggest reflects greater mentalizing demands for competition compared with cooperation.

Probing the Neural Correlates of Social Decision Making

Interactive tasks are also useful insofar as they allow us to proceed beyond probing the neural correlates of perception and judgment of social stimuli to the decision-making processes that guide social behavior. For example, McCabe et al. (McCabe et al., 2001) scanned subjects as they played a trust game with real human partners who were outside the scanner and with computer partners. Subjects who cooperated more often showed stronger activation in medial prefrontal cortex (mPFC) during the decision-making epoch when playing with human compared with computer partners (Fig. 15–9). The activation in mPFC is near the anterior paracingulate focus of ToM imaging studies, leading the authors to speculate that this region binds joint attention to mutual gains with the inhibition of immediate reward gratification to allow cooperative decisions. These results are consistent with those of Gallagher et al. and Rilling et al. in suggesting that mPFC is a domain-specific brain region, specialized for making inferences about the mental states of other humans.

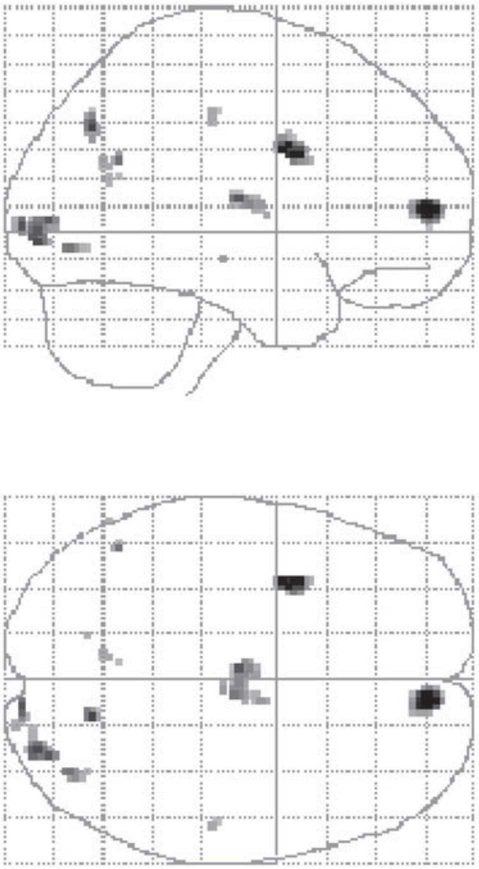


Fig. 15–9 Activation within medial prefrontal cortex in cooperative subjects during decision-making for the contrast between playing with human vs. computer partners. From McCabe K, Houser D, Ryan L, Smith V, & Trouard T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *PNAS 98*, 11,832–11,835. Copyright (2001) *National Academy of Sciences, USA*. Reprinted with permission.

de Quervain et al. (2004) also examined neural activity in a trust game, but they used PET rather than fMRI. As in the study by Delgado et al. (2005) discussed above, scanned subjects were in the role of first mover. If subjects chose to transfer money to their non-scanned partner, the money quadrupled and the partner could then decide to either return half the sum (share) to the subject or keep all of it. In this version of the game, scanned subjects were given the opportunity to pay to punish non-reciprocating partners. Subjects were scanned for 1 minute after learning that their partner had chosen to keep all the money, while they were deciding whether and how much to punish their partner. Effectively punishing a non-reciprocating partner was associated with activation in the caudate nucleus (dorsal striatum), a region implicated in processing rewards that accrue as a result of goal directed actions (Fig. 15–10). Moreover, subjects with stronger activation in the dorsal striatum were willing to incur greater costs to punish the partner more severely. Like the Singer et al. study, this finding supports the hypothesis that people derive satisfaction from punishing defectors in social exchange, or perhaps from achieving revenge more generally.

Like Delgado et al., King-Casas et al. (2005) scanned subjects with fMRI while playing an iterated trust game with human partners, but with the interesting addition that both participants were scanned simultaneously with a new technology known as hyperscanning (Montague et al., 2002). Within the dorsal striatum (head of the caudate nucleus), a stronger neural response to the revelation of the partner's investment predicted future increases in reciprocity by the player (Fig. 15–11). As the game progressed, this intention to trust signal shifted forward by 14 seconds such that it actually occurred during the partner's decision-making epoch, before the decision was actually revealed. This finding is reminiscent of analogous shifts of reward prediction error signals from reinforcement learning that have recently been identified with fMRI in caudate and putamen and are thought to involve outputs of midbrain dopaminergic systems (McClure et al., 2003; O'Doherty et al., 2003). So it seems that early in the game when the subject is uncertain as to the partner's intentions, reciprocity elicits a positive reward predication error. However, as the game progresses and the partner's behavior becomes more predictable, the positive reward

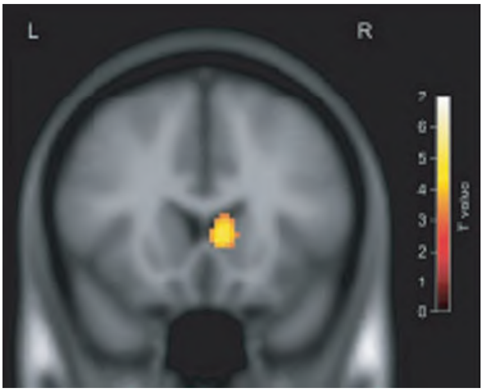


Fig. 15–10 Activation within caudate nucleus related to effectively punishing a non-reciprocating partner in a trust game. From de Quervain DJ, Fischbacher U, Treyer V, Schellhammer M, Schnyder U, Buck A, & Fehr E. (2004). The neural basis of altruistic punishment. *Science 305*, 1254–1258. Reprinted with permission from AAAS.

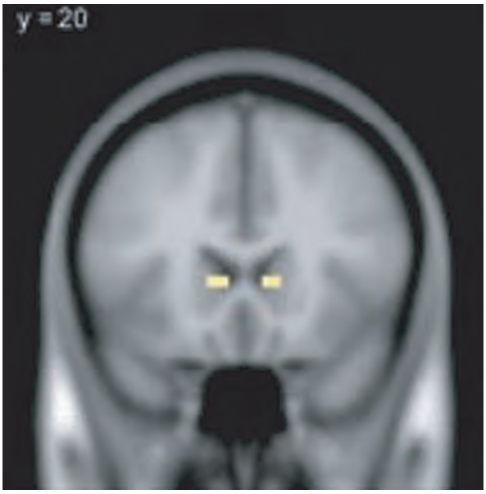


Fig. 15–11 Region of caudate nucleus where the magnitude of activation in response to revelation of the partner's investment positively predicted future increases in reciprocity in a trust game. From King-Casas B, Tomlin D, Anen C, Camerer CF, Quartz SR, & Montague PR (2005): Getting to know you: reputation and trust in a two-person economic exchange. *Science 308*, 78–83. Reprinted with permission from AAAS.

prediction error is transferred to an earlier point in time, which presumably reliably predicts reciprocity.

The finding of this study is echoed in the iterated PD fMRI results of Rilling et al. (2002), where the magnitude of activation within the anteroventral striatum in response to CC outcomes predicted the likelihood of the subject persisting with a strategy of mutual cooperation. That is, stronger activation was positively correlated with the probability of a CC outcome in the next round of the game.

The hyperscanning methodology pioneered by Montague and colleagues has obvious advantages in terms of data collection efficiency (i.e., collecting twice as much data in the same amount of time) but will also open new vistas in social cognitive neuroscience. For example, it will allow imaging of coordinated patterns of brain activity in people who are effectively working together toward a common goal. It could also be used to evaluate simulation theories of empathy, according to which we understand others by reproducing their neural states. Further applications for this method will undoubtedly emerge in the future.

Conclusions

In summary, interactive tasks have been productively used to provoke and image the neural correlates of social emotions and feelings such as indignation, empathy, trust, and social exclusion. Interactive tasks have also been used to image the neural correlates of mentalizing and competition and to probe the neural correlates of social decision making in the realm of cooperation and altruistic punishment. Studies involving interactive tasks are often more challenging to conduct because they often involve more subjects, more instruction of subjects, computer platforms that support interactive tasks, and less predictability and control of stimuli. For example, it is impossible to know in advance when a partner will choose to reciprocate cooperation or not, so statistical design matrices need to be tailored to individual subjects. Nevertheless, the extra effort can yield significant payoffs. These studies have already significantly advanced our knowledge of the neuroscience of human social behavior and will continue to do so in the future. To take just one example of what has been learned with this approach, the above studies have revealed a pervasive role of the caudate nucleus in human social interactions involving reciprocity, from distinguishing between reciprocation and non-reciprocation (Delgado et al., 2005; Rilling et al., 2002) to predicting future reciprocity (King-Casas et al., 2005; Rilling et al., 2002) to marking effective punishment of non-reciprocators (de Quervain et al., 2004; Singer et al., 2006). Using this approach, we can look forward to additional insights as the future of social cognitive neuroscience unfolds.

These studies have also yielded findings relevant to the issue of whether the human brain has domain-specific neural systems that are specialized for social cognition. One consistent finding is that the anterior paracingulate cortex shows stronger activation to interactions with human compared to computer partners. Perhaps not coincidentally, this region also contains a special type of neuron known as a Von Economo neuron, found at much higher density in the human brain than great ape brains and not present at all in the brains of monkeys (Allman et al., 2005). Given that humans undoubtedly have greater ToM capabilities than any other primate, these neurons in anterior paracingulate cortex could be part of a specialization for theory of mind. If anterior paracingulate cortex has a more basic, domain-general function, it is not immediately clear what that would be.

On the other hand, other brain regions that were mentioned above as being particularly responsive to human interactions, such as the insula, the anterior cingulate cortex, and the caudate nucleus, do have quite obvious domain-general functions. The insula and anterior cingulate cortex are responsive to painful stimuli, and the caudate is responsive to food reward and reward predication errors. Thus, the neural systems that respond to unfair treatment (i.e., insula) and reciprocated cooperation (caudate) likely originally evolved to solve fundamental problems such as avoidance of noxious physical stimuli and adaptive foraging, respectively (Panskepp, 1998). When social skills became more crucial with the evolution of primates, these systems were exapted (Gould & Vrba, 1982) for new functions such as detecting harmful social stimuli and learning when and to whom altruism should be dispensed. However, in the process of adapting these old systems to novel demands, the systems may well have been modified and social pressures may have left their imprint. Indeed, this is the manner in which the evolutionary process typically unfolds. For example, some mammals such as otters have evolved into aquatic niches. But rather than evolving completely novel structures to propel them through water, they modified their feet, the primary function of which is locomotion on land, to make them better suited to also travel through water. The end result is limbs designed for terrestrial locomotion that also have webbed feet designed for aquatic locomotion. Analogously, human brain systems may be designed for a combination of social and nonsocial functions.

Finally, cognitive neuroscientists are often challenged by others outside the field to state precisely how neuro-imaging has improved our knowledge of social cognition. For many of us, this question is misguided, as the primary goal of our research is not to better understand social cognition but to better understand the human brain. Understanding how the human brain gives rise to social cognition is an end unto itself.

References

Adolphs, R. (2001). The neurobiology of social cognition. *Current Opinion in Neurobiology 11*, 231–239.

Adolphs, R. (2003). Cognitive neuroscience of human social behaviour. [Review] [157 refs]. *Nature Reviews Neuroscience 4*, 165–178.

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences 4*, 267–278.

Allman, J. M., Watson, K. K., Tetreault, N. A., & Hakeem, A. Y. (2005). Intuition and autism: a possible role for Von Economo neurons. *Trends in Cognitive Sciences 9*, 367–373.

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature 402*, 179–181.

Brunet, E., Sarfati, Y., HardyBayle, M.-C., & Decety, J. (2000). A PET investigation of the attribution of intentions with a nonverbal task. *NeuroImage 11*, 157–166.

de Quervain, D. J., Fischbacher, U., Treyer, V., et al. (2004). The neural basis of altruistic punishment. *Science 305*, 1254–1258.

Decety, J., Jackson, P. L., Sommerville, J. A., Chaminade, T., & Meltzoff, A. N. (2004). The neural bases of cooperation and competition: an fMRI investigation. *Neuroimage 23*, 744–751.

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience 8*, 1611–1618.

Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology 6*, 178–190.

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An FMRI study of social exclusion. *Science 302*, 290–292.

Fletcher, P. C., Happe, F., Frith, U., et al. (1995). Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition 57*, 109–128.

Gallagher, H., Happe, F., Brunswick, N., Fletcher, P., Frith, U., & Frith, C. (2000). Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia 38*, 11–21.

Gallagher, H. L., & Frith, C. D. (2003). Functional Imaging of "theory of mind". *Trends in Cognitive Sciences 7*, 77–83.

Gould, S. J., & Vrba, E. S. (1982). Exaptation: a missing term in the science of form. *Paleobiology 8*, 4–15.

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgement. *Science 293*, 2105–2108.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron 44*, 389–400.

Iacoboni, M., Lieberman, M. D., Knowlton, B. J., et al. (2004). Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. *NeuroImage 21*, 1167–1173.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science 308*, 78–83.

MacDonald, A. W., 3rd, Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science 288*, 1835–1838.

Maddock, R. J. (1999). The retrosplenial cortex and emotion: new insights from functional neuroimaging of the human brain.[comment]. *Trends in Neurosciences 22*, 310–316.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *PNAS 98*, 11,832–11,835.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron 38*, 339–346.

Moll, J., Oliveira-Souza, R. D., Eslinger, P., et al. (2002). The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of Neuroscience 22*, 2730–2736.

Montague, P. R., Berns, G. S., Cohen, J. D., et al. (2002). Hyperscanning: simultaneous fMRI during linked social interactions. *Neuroimage 16*, 1159–1164.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience 16*, 1936–1947.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron 38*, 329–337.

Panskepp, J. (1998). *Affective Neuroscience: the Foundations of Human and Animal Emotions*. New York: Oxford University Press.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron 35*, 395–405.

Rilling, J. K., & Insel, T. R. (1999). The primate neocortex in comparative perspective using magnetic resonance imaging. *Journal of Human Evolution 37*, 191–223.

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). Opposing BOLD responses to reciprocated and unreciprocted altruism in putative reward pathways. *NeuroReport 15*, 2539–2543.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science 300*, 1755–1758.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage 19*, 1835–1842.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron 36*, 241–263.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. [see comment]. *Science 303*, 1157–1162.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature 439*, 466–469.

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences of the United States of America 93*, 13,515–13,522.

Vogeley, K., Bussfeld, P., Newen, A., et al. (2001). Mind reading: neural mechanisms of theory of mind and self-perspective. *Neuroimage 14*, 170–181.

Zola, S. M., Squire, L. R., Teng, E., Stefanacci, L., Buffalo, E. A., & Clark, R. E. (2000). Impaired recognition memory in monkeys after damage limited to the hippocampal region. *Journal of Neuroscience 20*, 451–463.

CHAPTER 16
Social Pain: Experiential, Neurocognitive, and Genetic Correlates

Naomi I. Eisenberger

> "There is much suffering in the world…from hunger, from homelessness, from all kinds of diseases. But the greatest suffering is being lonely, feeling unloved, having no one. I have come more and more to realize that it is being unwanted that is the worst disease that any human being can ever experience."—Mother Theresa

Mother Theresa's statement comes as no surprise to most observers of human nature, trained and untrained alike. Experience suggests that the pain of being socially estranged can be just as (if not more) distressing as the pain of hunger or the pain of cold. In fact, the "need to belong" has been identified by social psychologists as a fundamental human motivation that, when unsatisfied, leads to a variety of negative consequences, such as poor health and compromised well-being (Baumeister & Leary, 1995). However, is the pain that results from feeling unloved or unwanted the same kind of pain as that which results from feeling cold or hungry, or is Mother Theresa being metaphorical when she describes a lack of social connection as being "painful?" Can a lack of social connection actually lead to real pain experience, in the same manner that a lack of other basic needs can lead to pain experience? In the present chapter, I suggest, like others have previously (Baumeister & Leary, 1995), that the need for social connection is a fundamental need and that like other basic needs, a lack of social connection can feel "painful," an experience that has been termed "social pain" (Eisenberger & Lieberman, 2004, 2005; MacDonald & Leary, 2005).

The notion that a lack of social connection can lead to painful experience is not new. Rather, it is based on the hypothesis that over the course of mammalian evolution, the social attachment system, responsible for maintaining social connection, may have piggybacked directly onto the physical pain system, borrowing the pain signal to signify and thus prevent the danger of social separation (Panksepp, 1998). Because most mammals are born relatively immature without the capacity to feed or fend for themselves, it is necessary for mammalian infants to maintain close social contact with a caregiver to acquire the appropriate nourishment and protection. An overlap in the neural systems that support physical and social pain experience may have proved invaluable in this endeavor. To the extent that being separated from a caregiver threatens the survival of the infant, feeling "hurt" by separation from a caregiver may be an adaptive way to prevent future separation.

A review of the literature supports this hypothesized physical–social pain overlap and suggests that physical and social pain may share more than just metaphorical similarity. Observational, pharmacological, and neuropsychological evidence together suggest that physical and social pain processes share similar experiential, behavioral, and neural underpinnings.

Perhaps the most accessible source of data supporting a physical–social pain overlap comes from the English language. When individuals feel rejected or left out, they often describe their feelings with physical pain words, complaining of "*hurt* feelings," "*broken* hearts," or "feeling *crushed*." In fact, the English language has no direct synonym for these "hurt feelings," suggesting that the only way that English speakers can describe these feelings of social estrangement are with physical pain words. Indeed, the use of physical pain words to describe episodes of social estrangement is common to many languages (MacDonald & Leary, 2005), highlighting a potentially universal phenomenon.

Pharmacological research also supports the notion that physical and social pain share common substrates by showing that certain drugs have similar effects on both types of pain. For example, opiate-based medications (such as morphine or codeine), which are thought of primarily as "painkillers," also alleviate social pain (Herman & Panksepp, 1978; Kalin, Shelton, & Barksdale, 1988; Panksepp, 1998; Panksepp, Herman, Conner, Bishop, & Scott, 1978). Similarly, antidepressants, which are typically prescribed to treat anxiety and depression (often related to social stressors) are also effective in alleviating physical pain (Nemoto et al., 2003; Shimodozono, Kamishita, Ogata, Tohgo, & Tanaka, 2002; Singh, Jain, & Kulkarni, 2001) and are now commonly prescribed to treat chronic pain conditions.

Research from health psychology supports a physical–social pain overlap as well, demonstrating that changes in one type of pain experience correspond with changes in the other. For example, individuals with more social support (who should presumably experience less social pain) experience less cancer pain (Zaza and Baine, 2002), are less likely to suffer from chest pain following coronary artery bypass surgery (King, Reis, Porter, & Norsen, 1993; Kulik and Mahler, 1989), report less labor pain, and are less likely to use epidural anesthesia during childbirth (Chalmers, Wolman, Nikodem, Gulmezoglu, & Hofmeyer, 1995; Kennell, Klaus, McGrath, Robertson, & Hinkley, 1991). In addition, an experimental study has shown that compared to unsupported individuals, individuals who received social support from either a friend or stranger reported experiencing less pain during a cold pressor task, a task in which the participant's arm is submerged in ice water (Brown, Sheffield, Leary, & Robinson, 2003).

Finally, neuropsychological and neuroimaging research suggests that some of the same neural structures may underlie both physical and social pain. For example, the dorsal portion of the anterior cingulate cortex (dACC) is one neural region that seems to be involved in both forms of pain.[1] With regard to physical pain, the dACC is associated with the *affective* as opposed to the *sensory* component of pain. For example, following cingulotomy for chronic pain, a procedure in which a portion of the dACC is removed, patients report still being able to feel the intensity of pain but report that the pain no longer bothers them (Foltz & White, 1968), highlighting the role that this structure plays in registering the distressing, rather than the purely sensory, component of the pain experience. In line with this, several neuroimaging studies have shown that dACC activity correlates with perceived pain unpleasantness, whereas primary somatosensory cortex activity correlates with perceived pain intensity from cutaneous stimulation (Peyron, Laurent, & Garcia-Larrea, 2000; Ploghaus, et al., 1999; Rainville, Duncan, Price, Carrier, & Bushnell, 1997, Sawamoto et al., 2000). Thus, the dACC seems to be involved in the "distressing," or what is sometimes referred to as the "suffering," component of painful experience.

Although human research has focused on the role of the dACC in physical pain processes, animal research highlights a role for the ACC in social pain processes, such as those involved in preventing social estrangement and promoting

[1] The dACC has also been shown to play a role in more purely cognitive processes, such as "conflict monitoring," when behavioral response tendencies or expectations conflict (Botvinick, Cohen, & Carter, 2004), or "error detection" (Brown & Braver, 2005). These different roles will be discussed more fully at the end of the chapter.

social connection. Specifically, in nonhuman mammals, the ACC has been shown to play a role in the production of "distress vocalizations," a type of vocalization that is produced by infants upon separation from a caregiver. Distress vocalizations are considered to be the most primitive and basic mammalian vocalization with the original purpose of maintaining mother–infant contact (MacLean, 1985). Although it is impossible to determine whether these vocalizations are the product of painful or distressing experiences for the animal that is producing them, these vocalizations represent a behavioral indicator of sensitivity to social separation, which in humans may be a precursor for social pain experience.

To demonstrate the role that the ACC plays in distress vocalizations specifically, it has been shown that ablation of the ACC in squirrel monkeys leads to decreases in distress vocalizations but not other kinds of vocalizations (Hadland, Rushworth, Gaffan, & Passingham, 2003; MacLean & Newman, 1988), whereas electrical stimulation of the ACC in rhesus monkeys leads to the spontaneous production of distress vocalizations (Robinson, 1967; Smith, 1945). In addition, highlighting the specific role of the ACC rather than other neural regions in producing distress vocalizations, stimulation of the area corresponding to Broca's area, an area known to be involved in speech production, elicits movement of the vocal chords but no distress vocalizations in monkeys and apes (Leyton & Sherrington, 1917; Ploog, 1981). Thus, distress vocalizations seem to be uniquely related to ACC activation and not to the activation of neural regions typically involved in speech production. Finally, the cingulate gyrus (of which the ACC is a part) appears for the first time, phylogenetically, in mammalian species (MacLean, 1985) and thus may play a role in certain behaviors that also appear for the first time in mammals, such as those aimed at maintaining close social contact by producing distress or distress-related behaviors upon separation.

In sum, these lines of evidence support the notion that physical and social pain processes overlap by demonstrating that both types of pain rely on common experiential, behavioral, and neural substrates. However, there are still questions that remain. First, although it seems clear from the preceding review that the dACC is involved in the distress of physical pain experience in humans as well as in separation distress in nonhuman mammals, it is not clear if the dACC is also involved in socially painful experience in humans. Moreover, although there is some suggestion that physical and social pain share similar computational substrates and thus similar sensitivities, the extent to which sensitivity to social pain directly relates to sensitivity to physical pain has not been fully explored.

In the next section, I will review some of our own work that has examined these questions more closely. Two of these studies utilized functional neuroimaging methodologies to examine whether the dACC is sensitive to: *(1)* the experience of social pain in humans (social exclusion; Eisenberger, Lieberman, & Williams, 2003) and *(2)* cues that predict social pain in humans ("disapproving facial expressions;" Burklund, Eisenberger, & Lieberman, 2007). A third study examined the extent to which sensitivity to one type of pain relates to sensitivity to the other, as well as whether activating one type of pain heightens sensitivity to the other (Eisenberger, Jarcho, Lieberman, & Naliboff, 2006).

I will then highlight some of the extensions of this work by reviewing three studies that examined whether neural responses to social pain relate to and can help us understand real-world social phenomena. In other words, these studies utilized neural responses to social pain to help elucidate several unresolved questions regarding specific socio-emotional processes. The first study examined whether neural responses to experimental social rejection related to real-world feelings in social interactions, such as how rejected or accepted individuals tended to feel on a daily basis or the extent to which these feelings impacted more global judgments of social standing (Eisenberger, Gable, & Lieberman, 2007). The second study used neural sensitivity to social rejection, along with measures of social support and physiological stress reactivity, to better understand why social support is consistently related to reduced physiological stress reactivity and positive health outcomes (Eisenberger, Taylor, Gable, Hilmert, & Lieberman, 2007). The final study used neural sensitivity to social rejection to help understand the possible socio-emotional mechanisms that linked a specific genetic polymorphism to aggressive or antisocial behavior (Eisenberger, Way, Taylor, Welch, & Lieberman, 2007). Following this, I identify some of the questions that remain for understanding the neural correlates of social pain experience. I also highlight some key areas that will be critical for future research on social pain.

Investigating the Physical–Social Pain Overlap in Humans

The "Pain" of Social Exclusion

Based on the involvement of the dACC in physical pain distress in humans and in separation distress in nonhuman mammals, we investigated whether this neural region was also involved in the distress associated with social exclusion in humans. At the time that this study was conducted, no work had investigated the neural correlates associated with socially painful experience in human subjects.

In this study (Eisenberger, Lieberman, & Williams, 2003), participants were led to believe that they would be playing a virtual ball-tossing game called "Cyberball" (Williams, Cheung, & Choi, 2000) with two other players over the Internet while in the fMRI scanner. During one scan, participants played with the two computer players for the entire duration of the game. In a subsequent scan, participants were excluded from the ball-tossing game partway through the game when the two computer players stopped throwing the ball to them.

Upon being excluded from the game, compared to when being included, participants reported feeling significant levels of social distress (e.g., "I felt rejected," "I felt invisible") and showed increased activity in a region of the dACC, very similar to the region of the dACC associated with the unpleasantness of physical pain experience. Moreover, the magnitude of dACC activity correlated significantly with self-reports of social distress felt during the exclusion episode, such that individuals who showed greater dACC activity in response to social rejection also reported feeling more distressed by the rejection episode. Participants also showed increased activity in the insula, a region known to be involved in processing visceral sensation (e.g., visceral pain) as well as negative affective states (Aziz, Schnitzler, & Enck, 2000; Cechetto & Saper, 1987; Lane, Reiman, Ahern, Schwartz, & Davidson, 1997; Phan, Wager, Taylor, & Liberzon, 2004; Philips et al., 1997); however, insular activity did not correlate significantly with self-reported social distress in this study.

In addition, in response to social exclusion (vs. inclusion), participants showed increased activity in the right ventral prefrontal cortex (RVPFC), a region of the brain typically associated with regulating physical pain experience or negative affect (Hariri, Bookheimer, Mazziotta, 2000; Lieberman et al., 2004; Lieberman, Eisenberger, Crockett, Tom, Pfeifer, & Way, 2007; Petrovic & Ingvar, 2002). Consistent with this region's role in regulatory processes, greater activity in the RVPFC was associated with lower levels of self-reported social distress in response to the ball-tossing game, suggesting that this region may also be involved in regulating the distress of being socially excluded. Finally, we found that the dACC was a significant mediator of the RVPFC–distress relationship, such that RVPFC may be related to lower levels of social distress by downregulating the activity of the dACC.

Thus, neural responses to an episode of social exclusion recruited some of the same neural regions that are involved in the distress (dACC) and regulation (RVPFC) of physical pain experience. In fact, when comparing the neural activations in this study of social pain with those from a study of physical pain in patients with irritable bowel syndrome (Lieberman et al., 2004), very similar regions of activation in the dACC and RVPFC are observed (*see* Fig. 16–1; the left panel displays *social* pain, the right panel displays *physical* pain). Moreover, these two studies also demonstrate similar patterns of correlations between neural activity and pain distress, such that in both cases, greater dACC

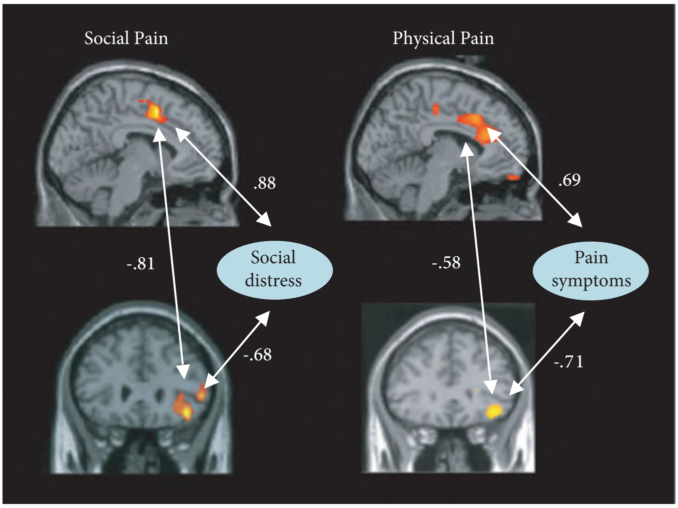


Fig. 16–1 **The left side of the panel displays the neural activity during social exclusion, compared to social inclusion, that correlates with self-reported social distress. (From Eisenberger NI, Lieberman MD, & Williams KD (2003). Does rejection hurt? An fMRI study of social exclusion. *Science, 302*, 290–292. Reprinted with permission from AAAS.). The right side of the panel displays the neural activity during painful visceral stimulation, compared to baseline, that correlates with self-reported pain experience. (From Lieberman, Jarcho, Berman, Naliboff, Suyenobu, Mandelkern, & Mayer, 2004).**

activity is associated with greater reports of social pain or physical pain distress, whereas greater RVPFC activity is associated with lower reports of distress and less dACC activity. Thus, not only do physical and social pain recruit some of the same neural regions, but for both types of pain, these neural regions relate to painful or distressing experience in similar ways.

As further evidence that social pain processes recruit pain-related neural regions, additional work has shown that other types of socially painful experience, such as bereavement or relationship dissolution, can lead to dACC activation as well. In one study (Gundel, O'Connor, Littrell, Fort, & Lane, 2003), bereaved participants were scanned while viewing pictures of their deceased first-degree relative or a stranger. In response to viewing pictures of the deceased, compared to pictures of a stranger, participants showed greater activity in regions of the dACC and insula. Similarly, in a study investigating the neural responses associated with grieving a romantic relationship, women whose romantic relationship ended within the preceding 4 months showed greater activity in several neural regions, including the dACC, when thinking about their relationship compared to when thinking about another individual (Najib, Lorberbaum, Kose, Bohning, & George, 2004). However, there were many neural regions activated in response to thinking about the former partner, and thus it is difficult to clearly identify which neural activations were specifically related to feelings of social pain. Nonetheless, together these studies support the notion that various types of socially painful experience activate pain-related neural regions such as the dACC.

The Face of Rejection

Based on our neuroimaging study of social exclusion as well as other studies of socially painful experiences, there is increasing evidence to suggest that the dACC is involved in the distressing experience associated with social pain experience in humans. Our next question was whether this neural region was also involved in responding to cues that signaled the possibility of socially painful experience. To examine this question, we investigated whether the dACC was involved in responding to "disapproving" facial expressions, a facial display that signified the possibility of social rejection (Burklund, Eisenberger, & Lieberman, 2007). Although many previous neuroimaging studies have investigated the neural responses associated with viewing specific emotional expressions (e.g., fear, anger, disgust), this is the first to explore the neural responses associated with viewing a disapproving face. We also examined whether there were differences in neural sensitivity to disapproving faces based on an individual's level of rejection sensitivity, an individual difference measure that should increase sensitivity to cues that signal social rejection (Downey & Feldman, 1996).

Participants were scanned while viewing a series of 3-second film clips depicting individuals making different emotional expressions. Participants viewed disapproving emotional expressions as well as angry and disgusted emotional expressions for comparison. Although all of these facial expressions can signal a threat to social connection, the "disapproving" facial expression is the only expression that has no other meaning but a threat to social connection. Thus, although anger and disgust expressions typically indicate physical and contamination threats, respectively, disapproval does not have a nonsocial interpretation.

Similar patterns of neural activity were found in response to each of the three facial expression conditions; participants showed significant activity in the amygdala and various regions of the PFC when viewing each of these emotional expressions compared to when viewing a neutral crosshair fixation. However, when examining individual differences in rejection sensitivity, we found that individuals who scored higher in rejection sensitivity showed greater dACC activity while viewing the disapproving faces but not while viewing the anger or disgust faces, highlighting a specific role for the dACC in responding to disapproving faces among rejection-sensitive individuals. Moreover, rejection sensitivity correlated specifically with dACC activity to disapproving faces but not with other limbic system activity (e.g., amygdala, insula), suggesting that dACC activity, rather than more general limbic system activity, may be specifically responsive to these cues of rejection.

This increased dACC activity to disapproving facial expressions among rejection-sensitive individuals could result from several factors. First, it is possible that rejection-sensitive individuals are more likely to feel socially distressed while viewing disapproving facial expressions and thus exhibit increases in distress-related dACC activity. Alternatively, it is possible that the dACC activity observed here is not directly related to the experience of social distress but, rather, that it is related to detecting cues that predict social distress, which may be more salient for rejection-sensitive individuals. Future research will be needed to disentangle between these two alternatives. In addition, future research will be needed to better understand why dACC activity in response to disapproving faces was limited to those high in rejection sensitivity and was not seen for the sample as a whole. It is possible that there was no main effect for dACC activity because the stimuli were not interpreted as personally relevant, except for those high in rejection sensitivity.

We also found that when viewing disapproving facial expressions, individuals who scored lower in rejection sensitivity exhibited greater activity in the subgenual ACC (subACC), a neural region that has been shown to play a role in the extinction of conditioned fear responses in humans (Phelps, Delgado, Nearing, & LeDoux, 2004) as well as in signaling a less threatening interpretation of a negative stimulus (Kim, Somerville, Johnstone, Alexander, & Whalen, 2003). Thus, it is possible that those low in rejection sensitivity may have shown greater subACC activity to disapproving faces because they were better able to regulate their negative responses to these disapproving facial expressions or better able to generate less threatening interpretations of these stimuli.

Finally, we found that neural activity in the subACC and dACC were negatively correlated with each other, such that individuals who showed greater activity in the subACC while viewing disapproving faces, compared to rest, also showed a corresponding reduction in dACC activity. These results are similar to previous findings showing an inverse relationship between subACC and amygdala activity when assessing the valence of certain stimuli (Kim et al., 2003). In that study, to the extent that surprised facial expressions were interpreted more positively, participants showed increased subACC activity and reduced amygdala activity; conversely, to the extent that surprised facial expressions were interpreted more negatively, participants showed reduced subACC activity and greater amygdala activity. In a similar manner, the present findings may suggest that individuals who interpret the disapproving facial expressions more positively (i.e., those low in rejection sensitivity) show greater subACC and reduced dACC activity, whereas individuals who interpret the disapproving facial expressions more negatively (i.e., those high in rejection sensitivity) show reduced subACC and greater dACC activity.

Shared Sensitivities to Physical and Social Pain

The studies reviewed thus far have used neuroimaging techniques to examine whether social pain processes in humans rely on some of the same neural structures that are involved in physical pain processes in humans and separation distress behaviors in nonhuman mammals. To examine the physical–social pain overlap in a different way, we conducted a behavioral study in which we used a measure of physical pain to investigate the extent to which people show similar patterns of sensitivity to physical and social pain. Specifically, we investigated: *(1)* whether individuals who are more sensitive to physical pain are also more sensitive to social pain and *(2)* whether inducing social pain potentiates sensitivity to physical pain stimuli, as triggering one type of pain should activate the underlying neural system that supports both types of pain processes (Eisenberger, Jarcho, Lieberman, & Naliboff, 2006).

Upon arriving in the lab, participants provided a baseline measure of sensitivity to heat pain by rating the temperature at which they perceived a painful heat stimulus delivered to their volar forearm to be very unpleasant (a 10 on a scale from 0 [“no sensation”] to 20 [“unbearable”]; Gracely, McGrath, & Dubner, 1978). After this, participants completed one round of the Cyberball game in which they were either included, not included (couldn’t play with the two other players because of technical difficulties), or overtly excluded (stopped receiving the ball from the two virtual players midway through the game) in a between-subjects manner. During the last 30 seconds of the Cyberball game, participants were exposed to three painful heat stimuli (at the temperature they reported to be “very unpleasant”) and were asked to rate the unpleasantness of each. They were also asked to rate how rejected they felt during the Cyberball game (level of social distress).

Results demonstrated that individuals who were more sensitive to physical pain at baseline (e.g., lower baseline pain thresholds) were also more distressed during social rejection (either non-inclusion or overt exclusion) but not during social inclusion, suggesting that individual sensitivity to one type of pain is related to sensitivity to the other. In addition, this relationship remained significant after controlling for neuroticism, suggesting that this relationship cannot simply be explained by a general tendency to report higher levels of negative experience. In addition, we found that individuals who felt the most distressed by the social rejection episodes also reported the highest pain ratings in response to the heat stimuli that were delivered at the end of the rejection episodes. Note that these heat stimuli were calibrated based on each subject’s baseline pain threshold, and thus this result is independent of the previous one. Although this finding was correlational, it suggests that augmented sensitivity to one type of pain is related to augmented sensitivity to the other. This relationship remained after controlling for neuroticism as well.

It should be noted that these findings are somewhat different from those of another study that examined the effect of social exclusion (using a different manipulation) on physical pain sensitivity (DeWall & Baumeister, 2006). In this study, social exclusion was manipulated by telling participants that they would be alone in the future. Participants in this "future alone" condition, compared to those who were given no feedback or who were told that they would have satisfying relationships in the future, showed a reduced (rather than an increased) sensitivity to physical pain. These different findings could result from the fact that the "future alone" manipulation may induce more depression-like affect, thus reducing pain sensitivity, whereas the Cyberball manipulation may induce more anxiety-like affect, making an increase in pain sensitivity more likely. Nonetheless, it is important to note that in both studies, sensitivity to physical pain still correlated directly with sensitivity to social pain. Thus, even among subjects in the "future alone" condition, those who showed the greatest sensitivity to physical pain also showed the greatest sensitivity to social pain as indicated by higher levels of empathy toward a rejected target individual. In other words, although the exclusion manipulations (future alone vs. Cyberball) had different effects on pain sensitivity, in both studies, sensitivity to physical pain still remained positively correlated with sensitivity to social pain.

Thus, overall, physical and social pain share not only similar neural substrates but similar experiential sensitivities as well, such that individual differences in sensitivity to physical pain experience covaried with individual differences in sensitivity to socially painful experience. Showing that social and physical pain experience track one another provides additional, behavioral evidence for the notion that physical and social pain share experiential, computational, and neural substrates.

Correlates of Neural Responses to Social Pain

Knowing that dACC responses to social rejection relate to feelings of social distress may help us to better understand the mechanisms underlying other phenomena that are likely to utilize this neural system. In the next section, I review three studies that utilized neural responses to social pain to help to better understand specific real-world social phenomena.

Cyberball and the Real World

Several studies now support the notion that experiences of social exclusion in the scanner lead to dACC activity and that the magnitude of dACC activity is associated with the degree to which individuals feel rejected or excluded. What is less clear, however, is whether these scanner-based responses to social rejection relate to how individuals experience real-world social interactions. In other words, do individuals who show greater dACC reactivity to social rejection in the scanner also report feeling more socially rejected or estranged in their real-world social interactions? In addition, are individuals who show greater dACC reactivity to scanner-based social rejection more likely to integrate their experiences of rejection into more negative global beliefs about themselves and their social worlds? Because it is not yet possible to directly assess whole-brain neural activity during naturalistic, real-world social encounters, we investigated whether neural responses during an experimental episode of social rejection within the fMRI scanner correlated with real-world experiences during ongoing social interactions (Eisenberger, Gable, & Lieberman, 2007).

To examine whether neural activity to social rejection in the scanner related to moment-to-moment feelings of social rejection in real-world interactions, participants completed the Cyberball social exclusion task in the scanner (as done in a previous sample; Eisenberger et al., 2003) and, at a separate point in time, completed a 10-day experience-sampling study in which they were randomly signaled at different times during the day and asked to report on their feelings of social distress in their most recent social interaction (*momentary social distress*: e.g., "I felt accepted/rejected by my interaction partner").

Results revealed that individuals who showed greater dACC activity to the Cyberball task in the scanner also reported feeling greater levels of momentary social distress during their real-world social interactions across the 10-day experience-sampling study. In addition, individuals who showed greater activity in response to social exclusion in the amygdala, a neural region involved in affective processing (Davis & Whalen, 2001), and in the periaqueductal gray (PAG), a neural region involved in pain processing and attachment-related behaviors (Bandler & Shipley, 1994), also reported feeling greater levels of momentary social distress across this 10-day period. This is a notable finding given that this neural activity was assessed during a brief episode of social rejection that is probably quite unlike what most individuals experience in their daily lives (presumably most real-world social interactions do not involve such overt social exclusion, at least in adults). However, the strong correlation between neural responses to scanner-based social rejection and self-reports of social distress during real-world interactions suggests a core sensitivity to experiences of social rejection, such that those who are the most sensitive to an experimental episode of social rejection are also the most sensitive to these types of experiences in their everyday lives.

As a second goal of the study, we were also interested in whether neural activity to social rejection in the scanner related to the extent to which momentary social distress was integrated into end-of-day global assessments of social disconnection. To examine this, participants provided a global assessment of social disconnection (*end-of-day social disconnection*: e.g., “Today, I generally felt accepted by others: strongly agree [1] to strongly disagree [7]”) at the end of each of the 10 days, and correlations were computed between momentary social distress and end-of-day social disconnection ratings across the 10-day period. This correlation provided an index of the extent to which momentary social distress scores corresponded with and perhaps contributed to end-of-day social disconnection ratings. Thus, individuals with a large, positive correlation were more likely to feel socially disconnected at the end of the day if they felt a lot of social distress during their moment-to-moment social interactions during the day, whereas individuals with a small correlation were those who showed no clear relationship between momentary and end-of-day reports. We then investigated how neural activity during social rejection in the scanner related to this correspondence measure. Thus, we were interested in the neural processes that occurred during an episode of social rejection that predict whether that experience will figure prominently into one’s later feelings about the whole day.

Here, activity in the dACC, amygdala, and PAG in response to experimental social rejection did *not* significantly relate to the correspondence between momentary social distress and end-of-day social disconnection; instead, activity in the left hippocampus and medial prefrontal cortex (mPFC; Brodmann’s Area [BA] 10) did. Individuals who showed greater hippocampal and mPFC activity during an experimental episode of social rejection demonstrated a greater correspondence between momentary social distress and end-of-day social disconnection, such that individuals who felt more social distress during their social interactions reported feeling more socially disconnected at the end of the day. Notably, the neural regions associated with this correspondence between momentary and retrospective reports are similar to those found in neuroimaging studies of memory encoding (Brewer, Zhao, Desmond, Glover, & Gabrieli, 1998; Wagner et al., 1998) as well as self-referential or autobiographical memory encoding (Cabeza et al., 2004; Macrae, Moran, Heatherton, Banfield, & Kelley, 2004). In these studies, individuals who demonstrated greater activity in the hippocampus when viewing presented stimuli or in the mPFC when viewing self-relevant stimuli were more likely to remember those stimuli in a subsequent memory test. In a similar fashion, the present data suggest that social experiences that are more deeply encoded when they occur may then be more easily retrieved when making global assessments of social disconnection in retrospective reports.

In sum, this study demonstrates that neural responses to an experimental episode of social rejection have meaningful real-world correlates, such that those who showed the greatest neural responses to social rejection in the scanner also reported feeling the most socially rejected in their real-world social interactions. In addition, these findings point to a double dissociation in the neural systems underlying momentary and retrospective reports of social disconnection (Lieberman, 2007). The neural regions associated with momentary social distress (dACC, amygdala, PAG) were not significantly associated with the correspondence between momentary and end-of-day assessments of social disconnection, and the neural regions associated with the correspondence between momentary and end-of-day social disconnection (mPFC, hippocampus) were not significantly associated with momentary social distress. These findings map nicely onto previous behavioral work demonstrating that moment-to-moment and retrospective reports of affect do not necessarily correspond (Fredrickson & Kahneman, 1993; Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993; Redelmeier & Kahneman, 1996; Updegraff, Gable, & Taylor, 2004) and suggest that part of the reason for this may result from the fact that these processes rely on the computational substrates of two separate neural systems. Future studies that continue to examine the relationships between neural responses within the fMRI scanner and real-world experiences may provide important information regarding how individuals experience their social worlds and the neurocognitive processes that underlie these experiences.

dACC Mediates the Effect of Social Support on Health-Related Outcomes

Although animal and human research has consistently demonstrated a relationship between a lack of social support and an increased risk of morbidity and mortality, the mechanisms underlying this relationship remain unknown and the neurocognitive mechanisms have been largely unexplored in humans. One hypothesis that has garnered some support is that social support reduces physiological stress reactivity (such as the release of cortisol, a neuroendocrine stress hormone) to threatening situations, which, over time, can have deleterious health consequences (Uchino et al., 1996).

Social support may modulate stress responses at two different points in the chain of events that lead from potential stressors to physiological stress responses (Cohen & Wills, 1981). First, social support may alter the appraisal or perception of potentially threatening conditions such that they are no longer perceived as stressful. Thus, feeling supported and cared for may lead an individual to be less likely to appraise certain conditions as threatening, preventing the onset of physiological stress reactivity. To the extent that social support downregulates threat-related reactivity, social support may be associated with less activity in limbic structures that are typically involved in responding to negative or threatening experiences, such as the amygdala, insula, or dACC. The second point at which social support may reduce physiological stress reactivity is after an event has been appraised as stressful but prior to the onset of prolonged physiological stress responses. Thus, individuals with greater social support may be better able to cope with or regulate negative stressful experiences, leading to reduced physiological stress responses through reappraisal or regulatory processes. To the extent that social support is important for regulating negative responses to stressors, social support may relate to increased activity in regions that are typically involved in regulating negative affect, such as VLPFC and mPFC (Ochsner & Gross, 2005).

To investigate the types of neural processes that underlie the stress-protective effects of social support, we investigated how daily levels of social support related to both neurocognitive and cortisol reactivity to a social rejection stressor. To assess daily levels of social support, participants completed a signal-contingent daily experience-sampling task, in which they were loaned a PalmPilot device and, for 10 days, were signaled at random times during the day to report on the degree to which their most recent interaction partner was someone they perceived to be generally supportive. To assess neural reactivity to social rejection, participants completed the Cyberball task within the fMRI scanner. To assess cortisol reactivity to a social stressor, all participants completed the Trier Social Stress Task (TSST; Kirschbaum et al., 1993), a task that requires participants to deliver an impromptu speech and perform mental arithmetic aloud in front of a nonresponsive, rejecting panel and, in a meta-analysis, has been shown to reliably elicit cortisol responses (Dickerson & Kemeny, 2004).[2]

Results showed that individuals who interacted regularly with supportive individuals across a 10-day period showed reduced activity in the dACC as well as reduced activity in BA 8 in the dorsal superior frontal gyrus, a region previously associated with the distress of social separation (Rilling, Winslow, O'Brien, Gutman, Hoffman, & Kilts, 2001). Moreover, reduced activity in these neural regions was associated with reduced cortisol reactivity to a social stressor. In addition, we found that individual differences in dACC and BA 8 reactivity mediated the relationship between high daily social support and low cortisol reactivity, such that supported individuals showed reduced neurocognitive reactivity to social stressors, which in turn was associated with reduced neuroendocrine stress responses. Thus, in the present study, social support related to reduced physiological stress reactivity by way of attenuated activity in neural regions that have previously been associated with distressing experience (dACC, BA 8), rather than by way of increased activity in regions previously associated with effortful, controlled processing or with regulating negative affect (LPFC, mPFC; Ochsner & Gross, 2005). Understanding how neural activity relates to social support and physiological stress reactivity thus helps to inform our understanding of the ways in which social support may relate to better health outcomes.

Using Neural Responses to Social Pain to Understand a Genetic Precursor to Aggression

In the past decade, there has been a surge of interest in understanding the genetic precursors of behavior and behavioral disorders. Neuroimaging techniques have played an integral role in this endeavor by allowing for the investigation of the neurocognitive mechanisms that may underlie gene–behavior relationships. For example, individuals with the short form of the serotonin transporter promoter polymorphism (5LC6A4), who are at a greater risk for anxiety disorders, have been shown to have stronger amygdala responses to negative stimuli (Hariri et al., 2002) and thus may be more dispositionally sensitive to fear-related stimuli. The implication of findings such as these is that neuroimaging techniques can be used to better understand the cognitive mechanisms that underlie gene–behavior or gene–disorder relationships.

Along these lines, we recently investigated whether neural responses to social rejection could inform our understanding of why individual differences in a gene that encodes monoamine oxidase-A (MAOA) relate to aggressive behavior (Eisenberger, Way, Taylor, Welch, & Lieberman, 2007). Previous work has demonstrated a link between MAOA, an enzyme that degrades monoamines such as serotonin (Caspi et al., 2002), and aggressive behavior. For example, MAOA-deficient men from a single Dutch kindred demonstrated elevated levels of impulsive aggression, arson, and attempted rape (Brunner, Nelen, Breakefield, Ropers, & van Oost, 1993). In addition, when exposed to early adversity, men with the low-expression allele (MAOA-L) of the 30-bp variable number tandem repeats polymorphism in the MAOA promoter (MAOA-uVNTR) were more likely to develop antisocial behavior than men with the high-expression allele (MAOA-H; Caspi et al., 2002). Despite mounting evidence suggesting a relationship between the MAOA-uVNTR and aggressive behavior, it is unclear how this genetic polymorphism predisposes individuals to aggressive behavior.

There are many possible mechanisms supporting this functional relationship between the MAOA polymorphism and aggressive behavior. We examined two possibilities, each related to social pain sensitivity. One possibility is that MAOA-L individuals show *blunted* socio-emotional sensitivity, making them less

[2] Although it would have been ideal to assess cortisol and neural responses simultaneously, the paradigm needed to produce cortisol responses was not amenable to the fMRI scanner. Previous research has demonstrated that the social-evaluative component of the TSST, the possibility that one could be evaluated and rejected, is critical for cortisol responses (Dickerson & Kemeny, 2004). Because of the difficulty in recreating an evaluative panel within the fMRI scanner, the Cyberball task, which has been shown to elicit feelings of rejection and is amenable to the fMRI scanner (Eisenberger et al., 2003), was used instead.

concerned with the feelings of others, less empathic, and thus more likely to commit violent crimes because they care less about harming others or the repercussions of doing so. Another possibility is that MAOA-L individuals show *heightened* socio-emotional sensitivity, making them more sensitive to negative social experiences like social rejection and more likely to respond to these experiences with defensively aggressive behavior. Numerous studies have shown that social rejection can trigger aggressive responses against the rejector (Crick & Dodge, 1996; Dodge et al., 2003; Dodge & Pettit 2003; Twenge, Baumeister, Tice, & Stucke 2001; Twenge, 2005).

To investigate these possibilities, we examined how different allelic variants in the MAOA polymorphism related to neural responses to the Cyberball game as well as to self-report measures of trait interpersonal hypersensitivity and trait aggression. To the extent that the MAOA–aggression link reflects *blunted* socio-emotional sensitivity, MAOA-L individuals should report less trait interpersonal hypersensitivity and show less dACC activity to social rejection than MAOA-H individuals. Alternatively, to the extent that the MAOA–aggression link reflects *heightened* socio-emotional sensitivity, MAOA-L individuals should report greater trait interpersonal hypersensitivity and show greater dACC activity to social rejection than MAOA-H individuals. In either case, MAOA-L individuals should report higher levels of trait aggression than MAOA-H individuals.

Consistent with previous work, we found that MAOA-L individuals did report higher levels of trait aggression than MAOA-H individuals. To examine the experiential or neurocognitive mediators of this gene–behavior link, we next investigated how the MAOA polymorphism related to trait interpersonal hypersensitivity and dACC activity to social rejection. Results indicated that MAOA-L individuals, compared to MAOA-H individuals, reported greater levels of trait interpersonal hypersensitivity as well as greater dACC responses to social rejection (*see* Fig. 16–2), suggesting that the relationship between MAOA and trait aggression may result from heightened, rather than blunted, socio-emotional sensitivity. We also found that the relationship between the MAOA polymorphism and trait aggression was partially mediated by self-reported trait interpersonal

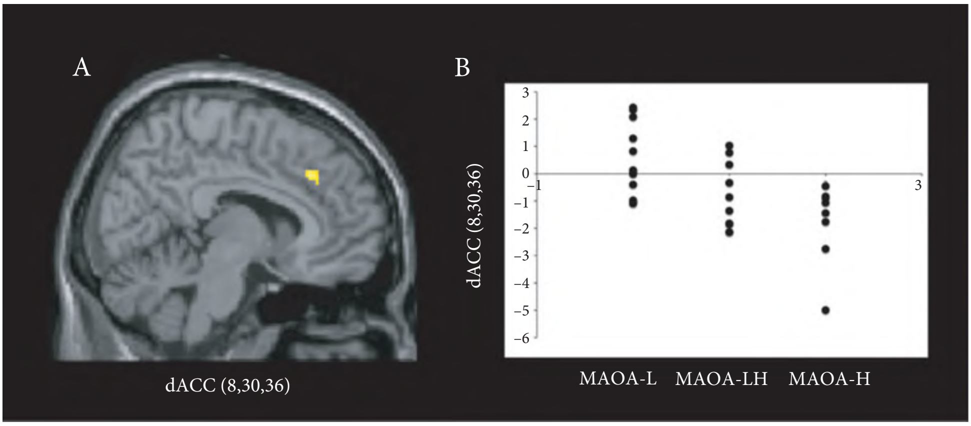


Fig. 16–2 **dACC activity (8,30,36) that varies as a function of the MAOA polymorphism. (A) Activity in the dACC, during social exclusion vs. inclusion, that correlates with individual differences in the MAOA polymorphism (maximum activation at 8,30,36) and shows greater activity for MAOA-L, compared to MAOA-H or MAOA-LH (females with one low expression and one high expression allele), individuals. (B) Scatterplot showing the relationship between the MAOA polymorphism and dACC (8,30,36) responses to social exclusion vs. inclusion.**

hypersensitivity as well as by dACC responses to social rejection.

These findings not only identify a possible genetic precursor to social pain sensitivity, but they also help to clarify some of the intervening mechanisms that link MAOA with aggressive behavior. Thus, instead of assuming that MAOA-related aggression results from psychopathy or a lack of social concern, it seems instead that MAOA-related aggression may be more closely tied to a heightened sensitivity to negative social cues, like social rejection, which may then trigger defensively aggressive behavior. Clarifying the underlying socio-emotional mechanisms that link MAOA to aggression is critical for both understanding the experience of individuals at risk for aggression and for identifying appropriate interventions for treating these aggressive behaviors. Moreover, identifying a genetic correlate of social pain sensitivity may aid not only in the identification and treatment of aggressive disorders but in the identification and treatment of other clinical disorders that relate closely to sensitivity to social pain as well (e.g., social anxiety, depression).

Summary

The studies reviewed here have several implications. First, they provide additional support for the notion that social pain and physical pain share some of the same experiential and neural substrates. We have shown that social pain in humans activates some of the neural structures that are involved in physical pain processing (Eisenberger et al., 2003), that cues of social rejection activate these regions among those who are the most rejection-sensitive (Burklund et al., 2007), and that sensitivity to physical pain is directly related to sensitivity to social pain (Eisenberger, Jarcho, Lieberman, & Naliboff, 2006). To the extent that these results point to common neural and behavioral mechanisms underlying physical and social pain, they support the notion that a lack of social connection can lead to pain experience and further the suggestion that social connection is indeed a fundamental need (Baumeister & Leary, 1995).

We have also shown that dACC activity during an experimental episode of social exclusion both relates to and helps us to understand real-world social experience and behavior. Thus, in one study, we demonstrated that neural responses to social rejection in the scanner corresponded strongly with the extent to which individuals felt socially rejected in their real-world social interactions (Eisenberger, Gable, & Lieberman, 2007). In a second study, we demonstrated that one way that social support relates to reduced physiological stress reactivity is through attenuated distress-related neural activity in regions like the dACC (Eisenberger, Taylor, Gable, Hilmert, & Lieberman, 2007). Finally, in a third study, we were able to use an assessment of neural activity to social rejection to elucidate a link between the MAOA gene and aggressive behavior. Here, we found that MAOA-related aggression was more closely related to heightened, rather than reduced, sensitivity to negative social experience, as evidenced by increased interpersonal sensitivity and increased dACC reactivity to social exclusion, among those with the low expression allele (Eisenberger, Way, Taylor, Welch, & Lieberman, 2007). Nonetheless, there are still unresolved issues regarding the role of the dACC in social pain processes and additional research that needs to be done. The final section of this review addresses one of the unresolved issues facing social pain research and highlights some key areas for future research.

Unresolved issues and future directions in social pain research

Relation of dACC Activity to Cognitive Processes

Finding such strong relationships between dACC activity and a negative socio-emotional experience like social rejection is somewhat at odds with previous cognitive neuroscience research. The most popular conceptions of dACC function have focused on its role in specific cognitive processes. For example, one prominent view of dACC function emphasizes its role in "conflict monitoring," in which the dACC monitors for conflicting response tendencies or goal representations to alert executive resources to implement cognitive control (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Botvinick, Cohen, & Carter, 2004; Carter, Braver, Barch, Botvinick, Cohen, & Noll, 1998; Carter et al., 2000; MacDonald, Cohen, Stenger, & Carter, 2000). This view is also closely related to a hypothesis suggesting that the dACC plays a role in error detection, detecting discrepancies between actual and intended events (Brown & Braver, 2005; Ito, Stuphorn, Brown, & Schall, 2003). Still, others have emphasized a role for the dACC in attentional processes more generally (Pardo, Pardo, Janer, & Raichle, 1990; Posner & Petersen, 1990). Moreover, a very influential review paper posited that the dorsal subdivision of the ACC is primarily involved in cognitive processes (i.e., conflict monitoring, attention-related processes), whereas the rostral-ventral subdivision of the ACC (rACC) is primarily involved in affective processes (Bush, Luu, & Posner, 2000). Indeed, this view has led some to suggest that the dACC activity seen in response to social exclusion during the Cyberball game may result from the fact that this exclusion is unexpected and that it is the ventral or subgenual portion of the ACC (subACC) that should be more directly activated by social rejection. However, in a study that attempted to dissociate expectancy violation from rejection, there was little evidence for the subACC playing a role in rejection-related distress; rather, subACC showed greater activity to the extent that subjects were accepted (Somerville, Heatherton, & Kelley, 2006).

Needless to say, although the proposed role for the dACC in cognitive processes specifically has been quite influential, it is at odds with the work reported here, showing a relationship between dACC activity and social pain, an experience that is undoubtedly affective in nature. Moreover, it is at odds with the work showing a relationship between dACC activity and physical pain distress (Price, 2000; Rainville, 2002; Rainville et al., 1997), anxiety (Bystritsky, Pontillo, Powers, Sabb, Craske, & Bookheimer 2001; Kimbrell et al., 1999; Nitschke, Sarinopoulos, Mackiewicz, Schaefer, & Davidson, 2006), and perceived stress (Wang et al., 2005). As can be seen in Figure 16–3, all of the activations reported in the present manuscript (related to social pain processes) fall within the dorsal, rather than the rostral-ventral, subdivision of the ACC, suggesting that the

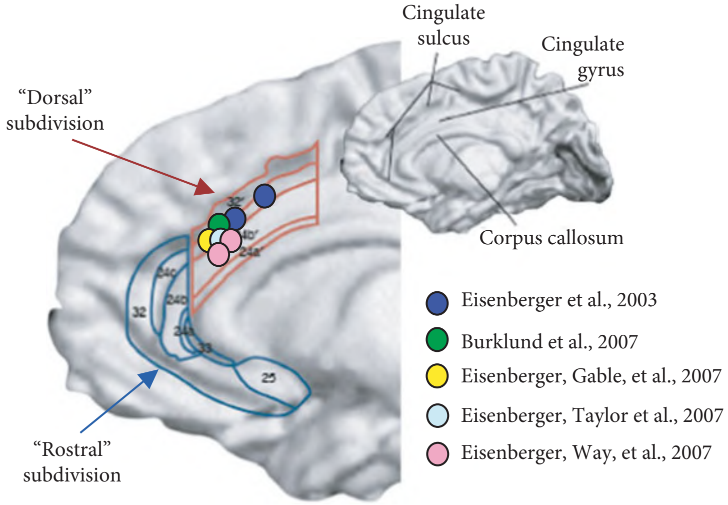


Fig. 16–3 A picture of the ACC adapted from Bush, Luu, & Posner (2000), showing dorsal and rostral-ventral subdivisions as well as how the neural responses reviewed in the present chapter map onto this figure.

dACC may be involved in affective processes as well.

As additional evidence supporting a role for the dACC in affective processes, numerous studies have shown that lesions to the dACC consistently result in reductions in distressing or anxious affective experience across many different patient populations (Baer et al., 1995; Ballantine, Bouckome, Thomas, & Giriunas, 1987; Ballantine, Cassidy, Flanagan, & Marino, 1967; Cohen, Paul, Zawacki, Moser, Sweet, & Wilkinson, 2001; Dougherty et al., 2002; Foltz & White, 1968); however, the data are less consistent with regard to how dACC lesions influence cognitive processes such as conflict monitoring. Across several studies that have examined Stroop performance (a task that assesses reaction times to trials containing conflicting information) following cingulotomy or naturally occurring dACC lesions, some studies have found reductions in conflict monitoring (as evidenced by reduced interference scores; Cohen, Kaplan, Moser, Jenkins, & Wilkinson, 1999; Cohen et al., 1999), one found increases in conflict monitoring (Ochsner et al., 2001), and some have found no differences in conflict monitoring (Fellows & Farah, 2005; Naccache et al., 2005; Turken & Swick, 1999) compared to controls. Thus, although cingulate lesions seem to relate uniformly to reductions in distressing affective experience, their impact on cognitive processes, like conflict monitoring, are still not well-understood.

Based on these additional data, it seems that this previous distinction between a "cognitive" and "affective" subdivision of the ACC needs to be revised and that the dominant focus on the role of the dACC in cognitive processes needs to be expanded. Rather than suggesting that the dACC is specifically involved in cognition or affect *per se*, we have posited that the dACC may be involved in both more basic cognitive processes, such as conflict monitoring or error detection, as well as in painful or distressing experience (Eisenberger & Lieberman, 2004). Thus, the dACC may function more generally as a "neural alarm system" that is involved in both detecting discrepancies from a desired standard (i.e., detecting threats to social connection) as well as in the phenomenological experience of distress (e.g., social pain experience) that is associated with bringing attention to the relevant problem and recruiting resources to fix or manage it. If the dACC functions more generally as a type of neural alarm system, it should be activated in response to the detection of simple discrepancies from desired standards (e.g., error detection), as suggested by research from cognitive neuroscience, and it should be activated in response to more complicated distressing experience (e.g., social rejection) that may represent a discrepancy from a desired standard (e.g., being socially connected), as suggested by the research reported here. Future research will be needed to determine whether these two processes, discrepancy detection and distressing experience, activate the same or different regions of the dACC.

Future Directions

Although some progress has been made in understanding the neural correlates of social pain processes in humans, there are still many issues that warrant further investigation. First and foremost, a careful examination needs to be carried out to investigate the specific overlap in the neural structures underlying physical and social pain experience by examining both of these processes within the same individuals. Identifying the overlap in the neural regions that underlie physical and social pain experience would be important for more clearly identifying the similarities and differences between these two processes. Some complicating issues with this approach include identifying a physical pain paradigm that most closely resembles social pain experience, as some types of physical pain (e.g., visceral pain) may approximate social pain more closely than others (e.g., somatosensory pain).

It will also be important to identify whether the neural responses to social exclusion are similar to or different from neural responses to other socially painful experiences. Neural responses to the Cyberball game provide us with information about the neural correlates associated with feeling rejected by individuals that one does not have a meaningful relationship with and presumably will not have future contact with. It is not yet clear if these are the same neural responses that would be seen with more self-relevant forms of socially painful experience. For example, do neural responses to experiences of discrimination look like neural responses to Cyberball or are discrimination-related neural patterns unique in some way? Are the neural correlates associated with the experience of bereavement the same as those involved in social rejection but more intense, or does bereavement activate different neural structures, based on specific processes that are unique to the loss an attachment figure? These questions are just beginning to be addressed (Gundel et al., 2003; Najib et al., 2006) and remain important and timely questions for future investigation.

In addition, there are presumably many more neural structures involved in the experience of social pain that have yet to be identified. For example, the insula is a neural structure that is involved in the processing of visceral sensation as well as negative affective experience (Aziz et al., 2000; Cechetto & Saper, 1987; Lane et al., 1997; Phan et al., 2004; Philips et al., 1997) and thus may play a role in social pain experience. Indeed, we found anterior insular activation in response to social exclusion in a previous study (Eisenberger et al., 2003). The PAG is another neural region that may play a role in social pain processes, as it has been shown to be involved in pain processing and attachment-related behaviors (Bandler & Shipley, 1994; Dunckley et al., 2005). Consistent with this, we found that PAG activity during social exclusion correlated with real-world reports of social distress in daily social interactions (Eisenberger, Gable, & Lieberman, 2007). Finally, the RVPFC, although not the primary focus of this chapter, is typically involved in regulating the distress of physical pain or negative affective experience (Hariri et al., 2000; Lieberman et al., 2007; Lieberman et al., 2004; Petrovic & Ingvar, 2002) and has been shown to play a role in regulating the distress of social pain as well (Eisenberger et al., 2003). Future studies will be needed to more completely identify the neural correlates of social pain experience.

Conclusions

Although we now know more about the neural correlates of social pain processes than we did 10 years ago, there is still much to learn. Regardless, it has been made clear across many different areas of research that social connection is critical for survival and well-being. From the earliest studies of mother–infant separation in rhesus monkeys (Harlow, 1958; Harlow & Zimmerman, 1959), demonstrating the importance of the mother–infant bond for normal socio-emotional development, to our present-day studies of the neural correlates of social pain, it is revealed over and over again that social relationships sustain, regulate, and promote physical, psychological, and emotional well-being. Although it can be debated as to whether a lack of social connection can truly engender pain experience, it is hard to argue with the notion that it "hurts" to be without the ones we love. Continuing to explore the neural substrates underlying our need for social connection may help us to better understand why.

References

Aziz, Q., Schnitzler, A., & Enck, P. (2000). Functional neuroimaging of visceral sensation. *Journal of Clinical Neurophysiology, 17*, 604–612.

Baer, L., Rauch, S.L., Ballantine, T., et al. (1995). Cingulotomy for intractable obsessive-compulsive disorder. *Archives of General Psychiatry, 52*, 384–392.

Ballantine, H.T., Bouckoms, A.J., Thomas, E.K., & Giriunas, I.E. (1987). Treatment of psychiatric illness by stereotactic cingulotomy. *Biological Psychiatry, 22*, 807–819.

Ballantine, H.T., Cassidy, W.L, Flanagan, N.B., & Marino, R. (1967). Stereotaxic anterior cingulotomy for neuropsychiatric illness and intractable pain. *Journal of Neurosurgery, 26*, 488–495.

Bandler, R. & Shipley, M.T. (1994). Columnar organization in the midbrain periaqueductal gray: Modules for emotional expression? *Trends in Neurosciences, 17*, 379–389.

Baumeister, R.F. & Leary, M.R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*, 497–529.

Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., & Cohen, J.D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*, 624–652.

Botvinick, M.M., Cohen, J.D., & Carter, C.S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences, 8*, 539–546.

Brewer, J.B., Zhao, Z., Desmond, J.E., Glover, G.H., & Gabrieli, J.D. (1998). Making memories: Brain activity that predicts how well visual experience will be remembered. *Science, 281*, 1185–1187.

Brown, J.W. & Braver, T.S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science, 307*, 1118–1121.

Brown, J.L., Sheffield, D., Leary, M.R., & Robinson, M.E. (2003). Social support and experimental pain. *Psychosomatic Medicine, 65*, 276–283.

Brunner, H.G., Nelen, M., Breakefield, X.O., Ropers, H.H., & van Oost, B.A. (1993). Abnormal behavior associated with a point mutation in the structural gene for monoamine oxidase A. *Science, 262*, 578–580.

Burklund, L.J., Eisenberger, N.I., & Lieberman, M.D. (2007). The face of rejection: Rejection sensitivity moderates dorsal anterior cingulate activity to disapproving facial expressions. *Social Neuroscience, 2*, 238–253.

Bush, G., Luu, P., & Posner, M.I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences, 4*, 215–222.

Bystritsky, A., Pontillo, D., Powers, M., Sabb, F.W., Craske, M.G., & Bookheimer, S.Y. (2001). Functional MRI changes during panic anticipation and imagery exposure. *Neuroreport, 12*, 3953–3957.

Cabeza, R., Prince, S.E., Daselaar, S.M., et al. (2004). Brain activity during episodic retrieval of autobiographical and laboratory events: An fMRI study using a novel photo paradigm. *Journal of Cognitive Neuroscience, 16*, 1583–1594.

Carter, C.S., Braver, T.S., Barch, D.M., Botvinick, M.M., Cohen, J.D., & Noll, D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science, 280*, 747–749.

Carter, C.S., MacDonald, A.W., Botvinick, M.M., et al. (2000). Parsing executive processes: Strategic vs. evaluative functions of the anterior cingulate cortex. *Proceedings of National Academy of Sciences, 97*, 1944–1948.

Caspi, A., McClay, J., Moffitt, T.E., et al. (2002). Role of genotype in the cycle of violence in maltreated children. *Science, 297*, 851–854.

Cechetto, D.F., & Saper, C.B. (1987). Evidence for a viscerotopic sensory representation in the cortex and thalamus in the rat. *Journal of Comparative Neurology, 262*, 27–45.

Chalmers, B., Wolman, W.L., Nikodem, V.C., Gulmezoglu, A.M., & Hofmeyer, G.J. (1995). Companionship in labour: Do the personality characteristics of labour supporters influence their effectiveness? *Curationis, 18*, 77–80.

Cohen, R.A., Kaplan, R.F., Moser, D.J., Jenkins, M.A., & Wilkinson, H. (1999). Impairments of attention after cingulotomy. *Neurology*, 53, 819–824.

Cohen, R.A., Kaplan, R.F., Zuffante, P., et al. (1999). Alteration of intention and self-initiated action associated with bilateral anterior cingulotomy. *Journal of Neuropsychiatry and Clinical Neuroscience, 11*, 444–453.

Cohen, R.A., Paul, R., Zawacki, T.M., Moser, D.J., Sweet, L., & Wilkinson, H. (2001). Emotional and personality changes following cingulotomy. *Emotion, 1*, 38–50.

Crick, N.R. & Dodge, K.A. (1996). Social information-processing mechanisms on reactive and proactive aggression. *Child Development, 67*, 993–1002.

Davis, M. & Whalen, P.J. (2001). The amygdala: Vigilance and emotion. *Molecular Psychiatry, 6*, 13–34.

DeWall, C.N. & Baumeister, R.F. (2006). Alone but feeling no pain: Effects of social exclusion on physical pain tolerance and pain threshold, affective forecasting, and interpersonal empathy. *Journal of Personality and Social Psychology, 91*, 1–15.

Dodge, K.A., Lansford, J.E., Salzer Burks, V., et al. (2003). Peer rejection and social information-processing factors in the development of aggressive behavior problems in children. *Child Development, 74*, 374–393.

Dodge, K.A. & Pettit, G.S. (2003). A biopsychosocial model of the development of chronic conduct problems in adolescence. *Developmental Psychology, 39*, 349–371.

Dougherty, D.D., Baer, L., Cosgrove, G.R., et al. (2002). Prospective long-term follow-up of 44 patients who received cingulotomy for treatment-refractory obsessive-compulsive disorder. *American Journal of Psychiatry, 159*, 269–275.

Downey, G. & Feldman, S.I. (1996). Implications of rejection sensitivity for intimate relationships. *Journal of Personality & Social Psychology, 70*, 1327–1343.

Eisenberger, N.I., Gable, S.L., & Lieberman, M.D. (2007). fMRI responses relate to differences in real-world social experience. *Emotion, 7*, 745–754.

Eisenberger, N.I., Jarcho, J.M., Lieberman, M.D., & Naliboff, B.D. (2006). An experimental study of shared sensitivity to physical pain and social rejection. *Pain, 126*, 132–138.

Eisenberger, N.I. & Lieberman, M.D. (2004). Why rejection hurts: The neurocognitive overlap between physical and social pain. *Trends in Cognitive Sciences, 8*, 294–300.

Eisenberger, N.I. & Lieberman, M.D. (2005). Why it hurts to be left out: The neurocognitive overlap between physical and social pain. In K.D. Williams, J.P. Forgas, & W. von Hippel (eds.), *The Social Outcast: Ostracism, Social Exclusion, Rejection, and Bullying* (pp. 109–127). New York: Cambridge University Press.

Eisenberger, N.I., Lieberman, M.D., & Williams, K.D. (2003). Does rejection hurt: An fMRI study of social exclusion. *Science, 302*, 290–292.

Eisenberger, N.I., Taylor, S.E., Gable, S.L., Hilmert, C.J., & Lieberman, M.D. (2007). Neural pathways link social support to attenuated neuroendocrine stress responses. *Neuroimage, 35*, 1601–1612.

Eisenberger, N.I., Way, B.M., Taylor, S.E., Welch, W.T., & Lieberman, M.D. (2007). Understanding genetic risk for aggression: Clues from the brain's response to social exclusion. *Biological Psychiatry, 61*, 1100–1108.

Fellows, L.K. & Farah, M.J. (2005). Is anterior cingulated cortex necessary for cognitive control? *Brain, 128*, 788–796.

Foltz, E.L. & White, L.E. (1968). The role of rostral cingulotomy in "pain" relief. *International Journal of Neurology, 6*, 353–373.

Frederickson, B.L. & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology, 65*, 45–55.

Gracely, R.H., McGrath, P., & Dubner, R. (1978). Validity and sensitivity of ratio scales of sensory and affective verbal pain descriptors: Manipulation of affect by diazepam. *Pain, 5*, 19–29.

Gundel, H., O'Connor, M.F., Littrell, L., Fort, C., & Lane, R.D. (2003). Functional neuroanatomy of grief: An fMRI study. *American Journal of Psychiatry, 160*, 1946–1953.

Hadland, K.A., Rushworth, M.F.S., Gaffan, D., & Passingham, R.E. (2003). The effect of cingulate lesions on social behaviour and emotion. *Neuropsychologia, 41*, 919–931.

Hariri, A.R., Bookheimer, S.Y., & Mazziotta, J.C. (2000). Modulating emotional response: Effects of a neocortical network on the limbic system. *NeuroReport, 11*, 43–48.

Hariri, A.R., Mattay, V.S., Tessitore, A., et al. (2002). Serotonin transporter genetic variation and the response of the human amygdala. *Science, 297*, 400–403.

Harlow, H.F. (1958). The nature of love. *American Psychologist, 13*, 673–685.

Harlow, H.F. & Zimmermann, R.R. (1959). Affectional responses in the infant monkey. *Science, 130*, 421–432.

Herman, B.H., & Panksepp, J. (1978). Effects of morphine and naloxone on separation distress and approach attachment: Evidence for opiate mediation of social affect. *Pharmacology and Biochemical Behavior, 9*, 213–220.

Ito, S., Stuphorn, V., Brown, J.W., & Schall, J.D. (2003). Performance monitoring by the anterior cingulate cortex during saccade countermanding. *Science, 302*, 120–122.

Kahneman, D., Fredrickson, B.L., Schreiber, C.A., & Redelmeier, D.A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science, 4*, 401–405.

Kalin, N.H., Shelton, S.E., & Barksdale, C.M. (1988). Opiate modulation of separation-induced distress in non-human primates. *Brain Research, 440*, 285–292.

Kennell, J., Klaus, M., McGrath, S., Robertson, S., & Hinkley, C. (1991). Continuous emotional support during labor in US hospital: A randomized control trial. *Journal of the American Medical Association, 265*, 2197–2201.

Kerns, J.G., Cohen, J.D., MacDonald, A.W., Cho, R.Y., Stenger, V.A., & Carter, C.S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science, 303*(5660), 1023–1026.

Kim, H., Somerville, L.H., Johnstone, T., Alexander, A.L., & Whalen, P.J. (2003). Inverse amygdala and medial prefrontal cortex responses to surprised faces. *Neuroreport, 14*, 2317–2322.

Kimbrell, T.A., George, M.S., Parekh, P.I., et al. (1999). Regional brain activity during transient self-induced anxiety and anger in healthy adults. *Biological Psychiatry, 46*, 454–465.

King, K.B., Reis, H.T., Porter, L.A., & Norsen, L.H. (1993). Social support and long-term recovery from coronary artery surgery: Effects on patients and spouses. *Health Psychology, 12*, 56–63.

Kulik, J.A. & Mahler, H.I. (1989). Social support and recovery from surgery. *Health Psychology, 8*, 221–238.

Lane, R.D., Reiman, E.M. Ahern, G.L., Schwartz, G.E., & Davidson, R.J. (1997). Neuroanatomical correlates of happiness, sadness, and disgust. *American Journal of Psychiatry, 154*, 926–933.

Leyton, A.S.F. & Sherrington, C.S. (1917). Observations of the excitable cortex of the chimpanzee, orangutan, and gorilla. *Quantitative Journal of Experimental Physiology, 11*, 135–222.

Lieberman, M.D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology, 58*, 259–289.

Lieberman, M.D., Eisenberger, N.I., Crockett, M.J., Tom, S., & Pfeifer, J.H. (2007). Putting feelings into words: Affect labeling disrupts amygdala activity in response to affective stimuli. *Psychological Science, 18*, 421–428.

Lieberman, M.D., Jarcho, J.M., Berman, S., et al. (2004). The neural correlates of placebo effects: A disruption account. *Neuroimage, 22*, 447–455.

MacDonald, A.W., Cohen, J.D., Stenger, V.A., & Carter, C.S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science, 288*, 1835–1838.

MacDonald, G. & Leary, M.R. (2005). Why does social exclusion hurt? The relationship between social and physical pain. *Psychological Review, 131*, 202–223.

MacLean, P.D. (1985). Brain evolution relating to family, play, and the separation call. *Archives of General Psychiatry, 42*, 405–417.

MacLean, P.D. & Newman, J.D. (1988). Role of midline frontolimbic cortex in production of the isolation call of squirrel monkeys. *Brain Research, 45*, 111–123.

Macrae, C.N., Moran, J.M., Heatherton, T.F., Banfield, J.F., & Kelley, W.M. (2004). Medial prefrontal activity predicts memory for self. *Cerebral Cortex, 14*, 647–654.

Naccache, L., Dehaene, S., Cohen, L., et al. (2005). Effortless control: Executive attention and conscious feeling of mental effort are dissociable. *Neuropsychologia, 43*, 1318–1328.

Najib, A., Lorberbaum, J.P., Kose, S., Bohning, D.E., & George, M.S. (2004). Regional brain activity in women grieving a romantic relationship breakup. *American Journal of Psychiatry, 161*, 2245–2256.

Nemoto, H., Toda, H., Nakajima, T., et al. (2003). Fluvoxamine modulates pain sensation and affective processing of pain in human brain. *Neuroreport, 14*, 791–797.

Nitschke, J.B., Sarinopoulos, I., Mackiewicz, K.L., Schaefer, H.S., & Davidson, R.J. (2006). Functional neuroanatomy of aversion and its anticipation. *Neuroimage, 29*, 106–116.

Ochsner, K.N., Kosslyn, S.M., Cosgrove, G.R., et al. (2001). Deficits in visual cognition and attention following bilateral anterior cingulotomy. *Neuropsychologia, 39*, 219–230.

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

Panksepp, J., Herman, B., Conner, R., Bishop, P., & Scott, J.P. (1978). The biology of social attachments: Opiates alleviate separation distress. *Biological Psychiatry, 13*, 607–618.

Pardo, J.V., Pardo, P.J., Janer, K.W., & Raichle, M.E. (1990). The anterior cingulate cortex mediates processing selection in the Stroop attentional conflict paradigm. *Proceedings of the National Academy of Sciences, 87*, 256–259.

Peyron, R., Laurent, B., & Garcia-Larrea, L. (2000). Functional imaging of brain responses to pain. A review and meta-analysis. *Neurophysiological Clinics, 30*, 263–288.

Petrovic, P. & Ingvar, M. (2002). Imaging cognitive modulation of pain processing. *Pain, 95*, 1–5.

Phan, K.L., Wager, T.D., Taylor, S.F., & Liberzon, I. (2004). Functional neuroimaging studies of human emotions. *CNS Spectrum, 9*, 258–266.

Phelps, E.A., Delgado, M.R., Nearing, K.I., & LeDoux, J.E. (2004). Extinction learning in humans: Role of the amygdala and vmPFC. *Neuron, 43*, 897–905.

Phillips, M.L., Young, A.W., Senior, C., et al. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature, 389*, 495–498.

Ploghaus, A., Tracey, I., Gati, J.S., et al. (1999). Dissociating pain from its anticipation in the human brain. *Science, 284*, 1979–1981.

Ploog, D. (1981). Neurobiology of primate audio-vocal behavior. *Brain Research, 3*, 35–61.

Posner, M.I. & Petersen, S.E. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13*, 25–42.

Price, D.D. (2000). Psychological and neural mechanisms of the affective dimension of pain. *Science, 288*, 1769–1772.

Rainville, P. (2002). Brain mechanisms of pain affect and pain modulation. *Current Opinions in Neurobiology, 12*, 195–204.

Rainville, P., Duncan, G.H., Price, D.D., Carrier, B., & Bushnell, M.D. (1997). Pain affect encoded in human anterior cingulate but not somatosensory cortex. *Science, 277*, 968–971.

Redelmeier, D.A. & Kahneman, D. (1996). Patients memories of painful medial treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66*, 3–8.

Rilling, J.K., Winslow, J.T., O'Brien, D., Gutman, D.A., Hoffman, J.M., & Kilts, C.D. (2001). Neural correlates of maternal separation in rhesus monkeys. *Biological Psychiatry, 49*, 146–157.

Robinson, B.W. (1967). Neurological aspects of evoked vocalizations. In S.A. Altmann (ed.), *Social Communication Among Primates* (pp. 135–147). Chicago, IL: The University Press.

Sawamoto, N., Honda, M., Okada, T., et al. (2000). Expectation of pain enhances responses to non-painful somatosensory stimulation in the anterior cingulate cortex and parietal operculum/posterior insula: An event-related functional magnetic resonance imaging study. *Journal of Neuroscience, 20*, 7438–7445.

Shimodozono, M., Kawahira, K., Kamishita, T., Ogata, A., Tohgo, S., & Tanaka, N. (2002). Reduction of central poststroke pain with the selective reuptake inhibitor fluvoxamine. *International Journal of Neuroscience, 112*, 1173–1181.

Singh, V.P., Jain, N.K., & Kulkarni, S.K. (2001). On the anitnociceptive effect of fluoxetine, a selective serotonin reuptake inhibitor. *Brain Research, 915*, 218–226.

Smith, W. (1945). The functional significance of the rostral cingular cortex as revealed by its responses to electrical excitation. *Journal of Neurophysiology, 8*, 241–255.

Somerville, L.H., Heatherton, T.F., & Kelley, W.M. (2006). Anterior cingulated cortex responds differentially to expectancy violation and social rejection. *Nature Neuroscience, 9*, 1007–1008.

Turken, A.U. & Swick, D. (1999). Response selection in the human anterior cingulate cortex. *Nature Neuroscience, 2*, 920–924.

Twenge, J.M. (2005). When does social rejection lead to aggression? The influences of situations, narcissism, emotion, and replenishing connections. In K.D. Williams, J.P. Forgas, & W. von Hippel (eds.), *The Social Outcast: Ostracism, Social Exclusion, Rejection, and Bullying* (pp. 201–212). New York: Cambridge University Press.

Twenge, J.M., Baumeister, R.F., Tice, D.M., & Stucke, T.S. (2001). If you can't join them, beat them: effects of social exclusion on aggressive behavior. *Journal of Personality and Social Psychologoy, 81*, 1058–1069.

Updegraff, J.A., Gable, S.L., & Taylor, S.E. (2004). What makes experiences satisfying? The interaction of approach-avoidance motivations and emotions in well-being. *Journal of Personality and Social Psychology, 86*, 496–504.

Wagner, A.D., Schacter, D.L., Rotte, M., et al. (1998). Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science, 281*, 1188–1191.

Wang, J., Rao, H., Wetmore, G.S., et al. (2005). Perfusion functional MRI reveals cerebral blood flow pattern under psychological stress. *Proceedings of the National Academy of Sciences, 102*, 17,804–17,809.

Ward, A.A. (1948). The cingular gyrus: Area 24. *Journal of Neurophysiology, 11*, 13–23.

Williams, K.D., Cheung, C.K.T., & Choi, W. (2000). Cyberostracism: Effects of being ignored over the Internet. *Journal of Personality and Social Psychology, 79*, 748–762.

Zaza, C. & Baine, N. (2002). Cancer pain and psychosocial factors: A critical review of the literature. *Journal of Pain and Symptom Management, 24*, 526–542.

CHAPTER 17
Could an Aging Brain Contribute to Subjective Well-Being? The Value Added by a Social Neuroscience Perspective

John T. Cacioppo, Gary G. Berntson, Antoine Bechara, Daniel Tranel, & Louise C. Hawkley

We thank Hanna Damasio for her assistance in the anatomical analyses of the neural lesions. Funding was provided by the National Institute of Aging Grant No. PO1 AG18911 and the John Templeton Foundation. Correspondence may be addressed to John T. Cacioppo, Center for Cognitive and Social Neuroscience, University of Chicago, 5848 S. University Avenue, Chicago, IL 60637, or Cacioppo@uchicago.edu.

People's moods modulate social cognition, interpersonal interactions, and social relationships. For example, negative moods can alter the perceived likelihood of occurrence for consequences presented when forming impressions or attitudes (Wegener, Petty, & Klein, 1994), and negative moods can have adverse effects on interpersonal interactions (Hawkley, Preacher, & Cacioppo, 2007). When feelings of dysphoria extend beyond minutes and hours to weeks and months, the individual transitions from negative moods to depressive symptomatology, the consequences of which can be devastating.

Individuals with elevated depressive symptoms are at risk for a host of problems, including cardiovascular disease (e.g., Barefoot & Schroll, 1996; Barth, Schumacher, & Herrmann-Lingen, 2004; Carney & Sheps, 2004), functional impairments (Mehta, Yaffe, & Covinsky, 2002), diminished immunosurveillance (Hawkley, Bosch, Engeland, Marucha, & Cacioppo, 2007), higher health-care resource utilization (Wells et al., 1989; Wygaard & Albreksten, 1992), social disruptions (Cacioppo et al., 2006), and feelings of social isolation (Cacioppo et al., 2006). Even treatments for depressive symptoms carry significant risks. Selective serotonin reuptake inhibitors (SSRIs) are a class of anti-depressant medications that help alleviate depressive symptoms in many individuals, but these drugs adversely affect osteoblasts (cells from bones), resulting in more brittle bones and increased risk of bone fractures and disability in the elderly (Richards et al., 2007). When totaled, the estimated annual economic cost of depressive symptomatology exceeds $43 billion (Greenberg, Stiglin, Finkelstein, & Berndt, 1993).

Depressive symptomatology is surprisingly prevalent in industrialized countries. Analyses from the first wave of the Health and Retirement Survey (HRS), a nationally representative longitudinal survey, indicated that approximately one-third of adults (33.6% of those ages 51–55 years and 31.2% of those ages 56–61 years) reported moderate to high levels of depressive symptoms (Steffick, 2000, Table 22). Data from the second and third wave of the HRS and the first two waves of the Assets and Health Dynamics Study of the Oldest Old, in which the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977) was used, indicated that from 14% to 19% responded that they "felt depressed," 21% to 25% responded that "everything was an effort," and 17% to 23% responded that they "could not get going" (Steffick, 2000, Table 9). These data suggest that depressive symptomatology has reached epidemic proportions and plays a significant role in social cognition, interpersonal relationships, and social behavior in industrialized nations. Indeed, the Federal Interagency Forum on Aging Related Statistics (2004) uses depressive symptomatology as an important indicator of general well-being and health among adults.

With the aging of industrialized nations, concerns arose about a corresponding increase in depressive symptomatology and a compounding of the health-care costs associated with an older population. It came as a surprise, then, when Carstensen and colleagues reported that at least until very late in life, healthy older adults reported *lower* levels of depressive symptomatology and *higher* levels of subjective well-being (e.g., Carstensen, Isaacowitz, & Charles, 1999). These findings were surprising not only because they went against social stereotypes of the misery of old age, but because cognitive declines were also evident in older adults (e.g., Petersen, Doody, Kurz et al., 2001). Age-related changes in cognition include a reduction in processing speed, episodic memory, and executive functioning, including problem solving and inhibitory control (e.g., Hasher & Zacks, 1988; Salthouse, 2001; Stern & Carstensen, 2000; von Hippel & Dunlop, 2005). Moreover, 5% of the United States population ages 65 to 69 years shows moderate or severe memory impairment, and 32% of those 85 years and older show moderate or severe memory impairment (Federal Interagency Forum on Aging-Related Statistics, 2004).

Carstensen et al.'s (1999) important work has led to efforts to determine the underlying cause of the age-related decline in depressive feelings in the hopes of improving treatments for depressive symptomatology across the age range. Early work focused on the temporal perspective and self-regulatory strategies that characterize healthy older adults, but attention to age-related changes in brain function provide an alternative explanation for these findings. Our goal in this chapter is to contrast these two perspectives to examine how a neuroscientific approach to a social problem can produce insights that would not be discernible from a social or behavioral perspective alone. We also illustrate the complementary nature of research using fMRI and lesion patients.

Age-related psychological changes

Cognitive declines are viewed generally as the consequence of an aging brain (e.g., McArdle et al., 2004), whereas the improved affect associated with aging has been attributed to changes in motivation derived from differences in time perspective (e.g., Carstensen, Fung, & Charles, 2003). According to Carstensen's socio-emotional selective theory, people have a sense of their time left in life, and perceived boundaries on time leads to attention be directed to emotionally meaningful aspects of life. When time is perceived as abundant, an individual's motivation and goals center on acquiring new information, expanding horizons, and pursuing achievements. When time is perceived to be limited, positive emotional experience becomes the preeminent motivation, and the individual tunes attentional, cognitive, and social investments to enhance emotional closeness and positive affect.

There is considerable evidence consistent with the predictions of socio-emotional selectivity. In an illustrative study, Nolen-Hoeksema and Ahrens (2002) investigated the levels of depressive symptoms in 25- to 35-year-old, 45- to 55-year-old, and 65- to 75-year-old adults. These groups were selected to represent different life circumstances and social histories. Results indicated that as a group, the older adults reported the lowest levels of depressive symptomatology.

We recently examined the determinants of subjective well-being in our population-based study of 50- to 67-year-olds in the Chicago Health, Aging, and Social Relations Study (CHASRS; Cacioppo et al., 2006). Consistent with prior research, our cross-sectional analyses indicated that dispositional variables, such as emotional stability, relationship satisfaction, and self-esteem, were associated with subjective well-being. Cross-sectional analyses may provide useful information on dispositional characteristics of happy people as well as risk factors;

longitudinal analyses are more useful to investigate likely causal influences. Longitudinal analyses of data from the first 3 years of CHASRS revealed an effect of age on changes in subjective well-being, as predicted by socio-emotional selectivity theory.

Given the replicability of age-related decreases in depressive symptomatology and increases in subjective well-being, investigations have turned from determining the association to explicating the underlying mechanism. Socio-emotional selectivity theory predicts that older adults will self-regulate their own affective states by choosing to attend to and think more about positive, in contrast to negative, stimuli and events in their daily lives. Consistent with this hypothesis, Carstensen and colleagues (Charles, Mather, & Carstensen, 2003) demonstrated that age-related decrements in memory performance are greater for negative than positive stimuli. Recall and recognition memory for positive, neutral, and negative pictures were measured in young (ages 18–29 years), middle-aged (ages 41–53 years), and older adults (ages 76–80 years). Results confirmed that young adults recalled comparable numbers of positive and negative stimuli, whereas middle-aged and older adults recalled more positive than negative.

However, an alternative to socio-emotional selectivity theory is suggested by the work on age-related changes in adrenergic and amygdala functioning. According to an aging-brain model (ABM): *(1)* the amygdala activation in response to negative stimuli decreases with age whereas amygdala activation to positive stimuli is maintained across age; *(2)* the decreased amygdala activation is associated with the diminution in emotional arousal to negative stimuli; and *(3)* the diminution of emotional arousal to negative stimuli that is associated with aging correspondingly reduces the memorial advantage conferred to emotionally arousing events and elevates subjective well-being. According to the ABM, these changes carry an additional cost: the maintenance of felt arousal to positive emotional outcomes and the diminution of felt arousal to negative emotional outcomes can also impair decision making in situations in which weighting negative feedback is essential (e.g., gambling).

The neuroscientifically inspired ABM differs from socio-emotional selectivity theory in several respects. Socio-emotional theory posits that the greater priority placed on emotional goals by older adults leads them to choose to attend to and allocate cognitive resources toward positive, rather than negative, stimuli (e.g., people, events) as a means of mood regulation and maintaining emotional closeness. According to the ABM, age-related changes in brain function include impairments in amygdala function, which results in reductions in emotional impact of negative, but not positive, stimuli. Both models predict that amygdala activation will be comparable to positive stimuli in young adults and will be smaller to negative than to positive stimuli in older adults. However, socio-emotional selectivity theory predicts that these amygdala changes are the consequence of the reduced attention to and cognitive emphasis on negative information in older adults that comes from their increased focus on emotional goals and emotional regulatory strategies (Carstensen et al., 2003). ABM predicts that these amygdala changes are the cause of the reduced impact of negative stimuli and, consequently, diminished depressive symptomatology and improved subjective well-being.

Imaging research has confirmed the pattern of amygdala activation predicted by socio-emotional selectivity theory and by the ABM. Mather et al. (2004) used event-related fMRI in a study of 17 healthy young adults (ages 18–29 years) and 17 older healthy adults (ages 70–90 years). The participants viewed 192 randomly ordered negative, neutral, and positive pictures from the International Affective Picture System (Lang, Bradley, & Cuthbert, 1999) in addition to 64 fixation trials (a large cross on the center for the screen). Each of these 256 stimuli were presented for 3 seconds, and after each the participants rated "how excited or calm you feel when you view each picture" using a scale from 1 to 4, with 1 labeled "completely relaxed, calm, sluggish, dull, sleepy, unaroused" and 4 labeled "stimulated, excited, frenzied, jittery, wide-awake, aroused." For older adults, the average signal change in the amygdala was larger for positive than negative pictures, whereas for young adults the average signal change in the amygdala was comparably large to positive and negative pictures. That is, young adults showed amygdala activation to positive and negative pictures, whereas older adults showed amygdala activation only in response to the positive pictures. Furthermore, Mather et al. (2004) found that young and older adults rated the positive pictures as comparably emotionally arousing, but older adults rated the negative pictures as less arousing than did the young adults.

The results of Mather et al. (2004) and Charles et al. (2003) are consistent with Carstensen's socio-emotional selectivity theory wherein older adults prefer emotionally meaningful experiences, and this increased focus on emotional goals and emotional regulatory strategies leads to a reduced cognitive focus on negative information. Their results are also consistent with the ABM, wherein the reduced amygdala activation in older, compared to young, adults found in response to negative (but not positive) stimuli is an age-related change in brain function and is causal: the lower amygdala activation shown selectively to negative stimuli drains them of emotional arousal, which, as Cahill and colleagues (1994, 1995) have shown, eliminates the memorial advantage usually found for emotional stimuli.

Socio-emotional selectivity theory and ABM differ in their predictions about attention to negative versus positive stimuli. Based on the former, Charles et al. (2003, Study 2) hypothesized that "older adults, compared with younger adults, would spend less time viewing the negative images than positive images" (p. 319) in their work of aging and emotional memory. Contrary to socio-emotional selectivity theory and consistent with the ABM, however, Charles et al. (2003) found that both young and old adults spent more time viewing negative than positive stimuli, and no differences between young and old adults were found in the time spent viewing negative (or positive) images.

The extant data are all correlational in nature, and they do not address whether the age-related changes in amygdala function observed in the literature: *(1)* are the consequence of changes in the perceived time left in life, which motivates changes in the attention to and cognitive operations on positive and negative information (the socio-emotional selectivity process-hypothesis), or *(2)* lessen the emotional arousal elicited specifically by negative stimuli, which, in turn, produces an effective landscape in which positive and negative stimuli are recognized as such but in which positive stimuli are associated with greater emotional arousal (and greater memorial and general affective impact) than negative stimuli (the ABM process-hypothesis). One of the fundamental tenets of the ABM is that patients with amygdala/anterior temporal lesions, even young patients with such lesions, will be selectively impaired in the arousal response to negative stimuli. This, of course, is because amygdala function is seen as cause rather than consequence. The investigation of patients with selective lesions of the amygdala/anterior temporal regions rather than fMRI studies of young and old adults permits the better test of this hypothesis.

Amygdala lesions and emotional arousal

To test the tenet that patients with amygdala/anterior temporal lesions will be selectively impaired in the felt arousal elicited by negative stimuli, Berntson, Bechara, Damasio, Tranel, and Cacioppo (2007) examined the separate valence and arousal aspects of evaluative judgments in the context of a comprehensive evaluative space model (Cacioppo & Berntson, 1994). Specifically, we compared the affective ratings (positivity, negativity, and arousal) of graded emotional picture stimuli (very positive, moderately positive, neutral, moderately negative, and very negative; International Affective Picture Series [IAPS], Lang et al., 1999) by six patients (four males and two females; ages 22–65 years, mean = 37.8) with amygdala/anterior temporal lesions to the ratings of a lesion control group (three males and three females; ages 33–61 years, mean = 51.2) with lesions sparing the amygdala and other areas that have been implicated in emotional processes (e.g., ventromedial prefrontal cortex, insula/SSII; Berntson et al., 2006). Results were also compared to a large set of normative data on these pictures (Lang, Bradley, & Cuthbert, 1999). Participants rated each of 48 pictures on a five-point bivariate scale of positivity and negativity, and a univalent scale of arousal. Pictures were matched on evaluative extremity from the mid (neutral)-point of the normative scale (12 very positive, 6 moderately positive, 12 neutral, 6 moderately negative, and 12 very negative) and normative arousal ratings. Pictures were presented in random order on a computer monitor for 6 seconds. Participants were instructed to focus on the emotional content of the pictures and to rate them on positivity and negativity by moving a mouse pointer and clicking a location in a 5 (positivity, 0 = not at all, 4 = extremely) x 5 (negativity, 0 = not at all, 4 = extremely) grid presented on the screen immediately after termination of the slide (Cacioppo et al., 2004). Immediately after responding, a second screen displayed a single-response continuum, and the participant was instructed to rate the arousability of the stimulus (0–4; 0 = not at all, 4 = extremely), again by the use of a mouse. Three seconds after completing the ratings, the next slide was presented. In addition to the separate ratings of positivity, negativity, and arousal, a net valence rating was calculated as the positivity rating minus the negativity rating for each picture.

As illustrated Figure 17–1, the amygdala/anterior temporal group showed markedly reduced arousal ratings to negative emotional stimuli, despite ratings of neutral and positive stimuli that were highly similar to those of the clinical control group and to a normative adult sample (Lang et al., 1999). An analysis of variance, with polynomial trends analysis, revealed a significant effect of picture category (very positive, moderately positive, neutral, moderately negative, very negative) on arousal ratings, characterized by a significant overall quadratic trend across picture categories. The latter reflects the minimal arousal to neutral stimuli and the progressively increasing arousal to either positive or negative stimuli, as has been reported previously. There also emerged a significant group x

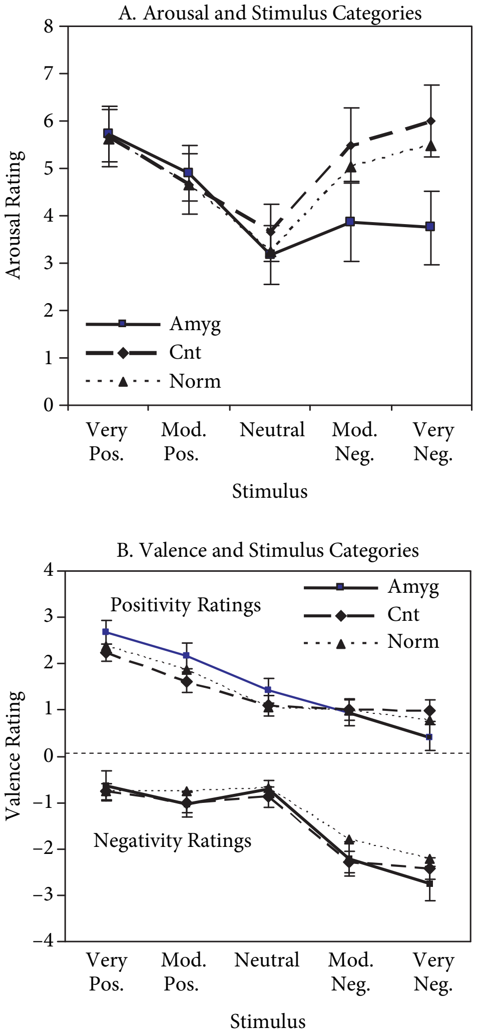


Fig. 17–1 Arousal and valence ratings. Mean (s.e.m.) arousal (A) and valence (B) ratings across stimulus categories, for patients with amygdala lesions (Amyg) compared to the clinical contrast group (Cnt) and normative control data (Norm). All groups effectively discriminated the stimulus categories and applied valance ratings accordingly. All groups also displayed comparable arousal functions to positive stimuli, but the amygdala group showed diminished arousal selectively to the negative stimuli. (From Berntson, G. G., Bechara, A., Damasio, H., Tranel, D., & Cacioppo, J. T. [2007]. Amygdala contribution to selective dimensions of emotion. *Social, Cognitive, and Affective Neuroscience, 2*, 123–129.)

picture-category interaction, characterized by a significant difference between the groups in the linear trend component across picture categories. This reflected the reduced arousal ratings, selectively for the negative pictures, by the amygdala/anterior temporal group. In contrast to the clinical control group and the normative group, the amygdala/anterior temporal group displayed minimal arousal ratings to negative picture content.

An important question arises as to whether the diminished arousal to negative stimuli in the amygdala/anterior temporal group may be attributable to impaired recognition or discriminative processing of the negative features of the pictures. As illustrated in Figure 17–1B, however, the amygdala/anterior temporal group was able to categorize and label the negative picture content accurately, suggesting a fundamental dissociation between the cognitive and affective processing of the stimuli in this group. Analysis of variance, with polynomial trends analysis, revealed the expected significant effects of picture category on positivity ratings and negativity ratings, each being characterized by a significant linear component (for positivity ratings and for negativity). There were no significant main effects or interactions for group on either positivity or negativity ratings, although for positivity ratings, an interaction on the linear component, reflecting the somewhat higher slope of the positivity-rating function of the amygdala/anterior temporal group, approached significance.

Two patients in the amygdala/anterior temporal patients had a bilateral lesion, and four had unilateral lesions. Bilateral amygdala lesions have generally been found to yield larger effects than unilateral lesions, although unilateral lesions have been reported to have similar although attenuated effects (Glascher & Adolphs, 2003; LaBar et al., 1995; Phelps et al., 1997) or, in some cases, even comparable effects to bilateral lesions (Buchanan et al., 2004). We examined this issue in further analyses of the change in response (from neutral) to positive and to neutral stimuli in bilateral and unilateral patients.

As illustrated in Figure 17–2, the lesion control group evidenced a somewhat greater arousal response to negative stimuli than to positive stimuli, whereas the reverse pattern was apparent for the amygdala/anterior temporal group. An analysis of variance revealed a significant group x picture-category (positive/negative) interaction on arousal ratings. This interaction was attributable to the similar arousal responses of the groups to positive stimuli, and the considerably smaller responses of the amygdala/anterior temporal group to negative stimuli.

Figure 17–2 also reveals differences in the arousal response of patients with bilateral and unilateral amygdala/anterior temporal lesions. The arousal response of the bilateral patients to negative stimuli were smaller than those of unilateral participants, but even those with unilateral lesions showed a substantial attenuation of arousal to negative stimuli.

The present findings indicate that patients with amygdala/anterior temporal lesions are not impaired at recognizing and labeling negative

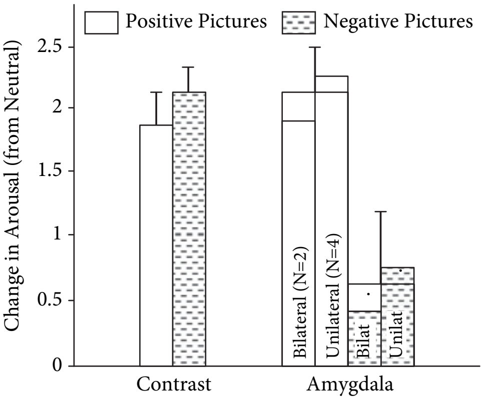


Fig. 17–2 Effects of bilateral and unilateral lesions. Heavy bars show overall mean (s.e.m.) of the change in arousal ratings to positive and negative stimuli, compared to neutral stimuli, for patients with amygdala damage and clinical contrast (Control) subjects. Lighter bars within the heavy bars of the Amygdala group illustrate effects of unilateral vs. bilateral lesions. (From Berntson, G. G., Bechara, A., Damasio, H., Tranel, D., & Cacioppo, J. T. [2007]. Amygdala contribution to selective dimensions of emotion. *Social, Cognitive, and Affective Neuroscience, 2*, 123–129.)

as well as positive emotional content in pictures. They are also capable of emotional arousal, as evidenced by typical arousal functions to positive emotional stimuli. In contrast, however, they display an attenuated arousal to negative emotional stimuli. These results are consistent with previous reports of diminished arousal to negative stimuli in patients with amygdala lesions (Adolphs, Russell, & Tranel, 1999; Winston et al., 2005) but extend these findings by showing that this arousal deficit does not arise from a cognitive/perceptual deficit in recognizing and labeling graded negative picture content (*see also* Adolphs et al., 1999).

Neither memory nor depressive symptomatology was measured in this preliminary study, but these data indicate that amygdala damage causes reduced emotional arousal specifically to negative stimuli. Given this effect, it is feasible that the memorial benefit typically characteristic of negative stimuli is diminished in individuals with amygdala damage, and relatedly the reduced emotional impact to and memory for negative events lessens depressive symptomatology and may impair decision making when consequences are negative. This hypothesis warrants further investigation.

We have further suggested that aging may be associated with a reduction in amygdala function, leaving older adults to evince cognitive and affective effects similar to those shown by middle-aged patients with amygdala damage. The patients with amygdala damage and the lesion controls whose data are shown in Figures 17–1 and 17–2 were selected to be matched on gender and age. In addition, Berntson et al. (2007) identified two patients older than age 80 years with damage that spared the amygdala/anterior temporal region. Figure 17–3 presents the results for the arousal ratings. The five categories of images along the *x*-axis are very positive, moderately positive, neutral, moderately negative, and very negative, and their arousal ratings are depicted along the ordinate. The dotted line represents the normative ratings obtained from undergraduates, the dashed line represents the mean ratings of emotional arousal by middle-aged lesion controls, and the solid line represents the mean ratings of emotional arousal by the two elderly lesion patients. The arousal ratings from the elderly patients look similar to those obtained from the middle-aged patients with amygdala damage. Although preliminary, these data, together with the investigations using fMRI showing that the amygdala activation observed in young adults to negative or threatening stimuli are reduced or absent in older adults (e.g., Gunning-Dixon et al., 2003), fit the notion that age-related brain changes include reductions in amygdala function, which in turn leads to reduced emotional arousal specifically to negative stimuli and, consequently, to reduced emotional impact to and memory for negative events. If this is the case, reduced depressive symptomatology and improved subjective well-being could easily be conceived to follow (Wood & Conway, 2006).

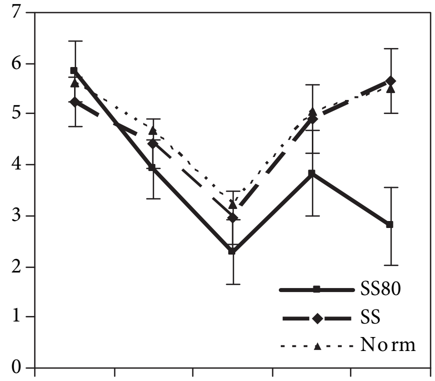


Fig. 17–3 Age and arousal ratings. Mean (s.e.m.) arousal ratings across stimulus categories, for two elderly patients with lesions sparing the amygdala and regions implicated in emotion (SS80) compared to the clinical contrast group (Cnt) and normative control data (Norm).

In sum, work on the aging brain suggests there is a shift from the amygdala-hippocampal system to the prefrontal cortex over time (Leigland, Schulz, & Janowsky, 2004). This shift may be more nuanced than previously thought, however, with diminished amygdala function more evident in response to negative than positive stimuli, and with consequences for affect, cognition, and social behavior that have more in common than previously appreciated. Preliminary evidence for this implication was recently reported by Williams et al. (2006) in the neurosciences.

Conclusion

There have been tremendous advances in our understanding of the links between the mind, brain, and behavior over the past quarter century, but people have generally been considered as isolated units in these analyses. People are inherently social creatures, however, and the tools are now available to determine the biological mechanisms underlying social cognition, emotion, and interpersonal interactions. Discovering the biological mechanisms underlying social and affective processes and behavior is undoubtedly one of the major problems for the neurosciences to address in the twenty-first century.

An assumption underlying social neuroscience is that all human social behavior is implemented biologically. The generative power of social neuroscience comes in part from a focus on fundamental mechanisms and processes, such as age-related changes in brain function, and from the derivation of novel and testable predictions about common underlying mechanisms for what appear to be disparate effects. We have attempted to illustrate these points in the current chapter. Age-related improvements in subjective well-being and age-related decrements in cognitive functioning and decision making have been treated initially as separate effects. These effects were discovered in disparate fields and continue to be regarded by many as unrelated changes associated with aging. Although we do not mean to suggest that all age-related memory impairments, decision-making problems, and mood improvements derive from a single cause, age-related changes in amygdala functioning may contribute to each. Based on a developing model of the neural organization of evaluative processes and animal and human research on the contribution of specific neural substrates to valence judgments, we reasoned that the amygdala complex is important in affective processing of negatively valenced stimuli and specifically that amygdala damage is associated with the reduction of emotional arousal elicited by negative but not neutral or positive stimuli. Reductions in emotional arousal are commonplace in aging, and we reasoned that both amygdala activation and the emotional arousal elicited by negative stimuli are lower in older adults than in young adults. In accord with prior work showing that emotional arousal promotes memory for emotional events, we further reasoned that the reduction in the age-related changes in amygdala activation and emotional arousal to negative events is associated with poorer memory for these events, greater reductions in overall negative affect (i.e., improved subjective well-being), and impaired decisions in circumstances that require the consideration of negative information. The extant evidence is consistent with this analysis, but additional data are required to test related predictions of the ABM.

Whether or not this model proves to be correct is less important here than its origins in the neuroscience literature, its suggestion that what appear to be disparate effects in socio-emotional processes may have a common underlying cause, its very different explanation from socio-emotional selectivity theory for age-related changes in subjective well being, its integration of data from multiple levels of organization (e.g., neurochemical, brain, affect, cognition, and social behavior), its generativeness and falsifiability, and its demonstration of the complementary roles played by lesion and imaging research. Indeed, the value added by a social neuroscience perspective is that novel, testable hypotheses for complex social processes and behaviors are not only possible but natural. As a consequence, this perspective pulls together literatures, techniques, and investigators from scientific disciplines that were once thought to have nothing in common.

These developments are no more a threat to traditional social psychological and social science approaches than the proposed model is a threat to socio-emotional selectivity theory. The two models are incompatible in some respects (e.g., the role of the amygdala), but evidence for the ABM does not mean that socio-emotional selectivity theory cannot also operate in some circumstances, and vice versa. Indeed, Carstensen and Fredrikson (1998) demonstrated that the regulation of emotional states and the consequent improvements in affect are not limited to older adults; younger adults who are approaching the end of life show similar motivational adjustments. This result suggests both processes may contribute to improvements in subjective well-being later in life.

Finally, an assumption underlying social neuroscience is that all human social behavior is implemented biologically. It does not follow that the concepts of biology can by themselves directly describe or explain social behavior or that "molecular" forms of representation provide the only or best level of analysis for understanding human behavior or mental disorders. Scientific constructs developed by behavioral and social scientists provide a means of understanding highly complex activity without needing to specify each individual action by its simplest components, thereby providing an efficient approach to describing complex systems. More importantly, there are concepts essential to the description and understanding of social behavior that are not contained in biology. By analogy, chemists who work with the periodic table on a daily basis use recipes rather than the periodic table to cook, not because a particular food preparation cannot be coded by complex chemical expressions (Cacioppo et al., 2000). However, efficiency of expression is not the only issue: the concepts defining fine cuisine are not part of the discipline of chemistry. The theoretical terms of the behavioral and social sciences are similarly valuable in relation to those of biology but can be informed, complemented, and refined through integration with theories and methods from the neurosciences.

References

Adolphs, R., Russell, J. A., & Tranel, D. (1999) A role for the human amygdala in recognizing emotional arousal from unpleasant stimuli. *Psychological Science, 10*, 167–171.

Adolphs, R., Tranel, D., & Damasio, A. R. (2001). Emotion recognition from faces and prosody following temporal lobectomy. *Neuropsychology, 15*, 396–404.

Adolphs, R. (1999). Social cognition and the human brain. *Trends in Cognitive Sciences, 3*, 469–479.

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature, 433*, 68–72.

Adolphs, R., Tranel, D., & Damasio, A. R. (1998). The human amygdala in social judgment. *Nature, 393*, 470–474.

Aston-Jones, G., Rajkowski, J., Kubiak, P., Valentino, R. J., & Shipley, M. T. (1996). Role of the locus coeruleus in emotional activation. *Progress in Brain Research, 107*, 379–402.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Basso, A., & Piantanelli, L. (2002). Influence of age on circadian rhythms of adrenoceptors in brain cortex, heart and submandibular glands of BALB/c mice: when circadian studies are not only useful but necessary. *Experimental Gerontology, 37*, 1441–1450.

Bechara, A., Damasio, H., Damasio, A. R., & Lee, G. P. (1999). Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *Journal of Neuroscience, 19*, 5473–5481.

Beilin, L. J. (1999). Lifestyle and hypertension—an overview. *Clinical & Experimental Hypertension, 21*, 749–762.

Beninger, R. J., Dringenberg, H. C., Boegman, R. J., & Jhamandas, K. (2001). Cognitive effects of neurotoxic lesions of the nucleus basalis magnocellularis in rats: differential roles for corticopetal versus amygdalopetal projections. *Neurotoxicity Research, 3*, 7–21.

Berntson, G. G., Bechara, A., Damasio, H., Tranel, D., & Cacioppo, J. T. (2006). Amygdala contribution to selective dimensions of emotion. *Social, Cognitive, and Affective Neuroscience, 2*, 123–129.

Berntson, G. G., Sarter, M., & Cacioppo, J. T. (1998). Anxiety and cardiovascular reactivity: the basal forebrain cholinergic link. *Behavioural Brain Research, 94*, 225–248.

Berntson, G. G., Sarter, M., & Cacioppo, J. T. (2003). Ascending visceral regulation of cortical affective information processing. *European Journal of Neuroscience, 18*, 2103–2109.

Berntson, G. G., Shafi, R., & Sarter, M. (2002). Specific contributions of the basal forebrain corticopetal cholinergic system to electroencephalographic activity and sleep/waking behavior. *European Journal of Neuroscience, 16*, 2453–2461.

Berntson, G. G., Shafi, R., Knox, D., & Sarter, M. (2003). Blockade of epinephrine priming of the cerebral auditory evoked response by cortical cholinergic deafferentation. *Neuroscience, 116*, 179–186.

Berntson, G. G., Quigley, K. S., & Lozano, D. (2007). Cardiovascular psychophysiology. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of Psychophysiology, 3rd edition* (pp. 182–210). New York: Cambridge University Press.

Bradley, M. M., & Lang, P. J. (1999a). *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings (Tech. Rep. No. C–1).* Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

Bradley, M. M., & Lang, P. J. (1999b). *International Affective Digitized Sounds (IADS): Stimuli, Instruction Manual and Affective Ratings (Tech. Rep. No. B–2).* Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage, 5*, 49–62.

Brodde, O. E., & Leineweber, K. (2004). Autonomic receptor systems in the failing and aging human heart: similarities and differences. *European Journal of Pharmacology, 500*, 167–176.

Buchanan, T. W., Tranel, D., & Adolphs, R. (2004). Antermedial temporal lobe damage blocks startle modulation by fear and disgust. *Behavioral Neuroscience, 118*, 429–437.

Buchanan, T. W., Tranel, D., & Adolphs, R. (2005). Memories for emotional autobiographical events following unilateral damage to medial temporal lobe. *Brain*, 2005 November 16; [Epub ahead of print].

Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: a critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin, 115*, 401–423.

Cacioppo, J. T., Berntson, G. G., Klein, D. J., & Poehlmann, K. M. (1998). The psychophysiology of emotion across the lifespan. *Annual Review of Gerontology and Geriatrics, 17*, 27–74.

Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., & Ito, T. A. (2000). The psychophysiology of emotion. In R. Lewis & J. M. Haviland-Jones (Eds.), *The Handbook of Emotion, 2nd edition* (pp. 173–191). New York: Guilford Press.

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.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1999). The affect system has parallel and integrative processing components: form follows function. *Journal of Personality and Social Psychology, 76*, 839–855.

Cacioppo, J. T., Hughes, M. E., Waite, L. J., Hawkley, L. C., & Thisted, R. A. (2006). Loneliness as a specific risk factor for depressive symptoms: cross sectional and longitudinal analyses. *Psychology and Aging, 21*, 140–151.

Cacioppo, J. T., Larsen, J. T., Smith, N. K., & Berntson, G. G. (2004). The affect system: what lurks below the surface of feelings? In A. S. R. Manstead, N. H. Frijda, & A. H. Fischer (Eds.), *Feelings and Emotions: The Amsterdam Conference* (pp. 223–242). New York: Cambridge University Press.

Cahill, L., & Alkire, M. T. (2003). Epinephrine enhancement of human memory consolidation: interaction with arousal at encoding. *Neurobiology of Learning and Memory, 79*, 194–198.

Cahill, L., Babinsky, R., Markowitsch, H. J., & McGaugh, J. L. (1995). The amygdala and emotional memory. *Nature, 377*, 295–296.

Cahill, L., Haier, R. J., White, N. S., et al. (2001). Sex-related difference in amygdala activity during emotionally influenced memory storage. *Neurobiology of Learning and Memory, 75*, 1–9.

Cahill, L., Prins, B., Weber, M., & McGaugh, J. L. (1994). β-adrenergic activation and memory for emotional events. *Nature, 371*, 702–704.

Cape, E. G., Manns, I. D., Alonso, A., Beaudet, A., & Jones, B. E. (2000). Neurotensin-induced bursting of cholinergic basal forebrain neurons promotes γ and θ cortical activity together with waking and paradoxical sleep. *Journal of Neuroscience, 20*, 8452–8461.

Carstensen, L. L., & Fredrickson, B. F. (1998). Influence of HIV status and age on cognitive representations of others. *Health Psychology, 17*, 494–503.

Carstensen, L. L., Isaacowitz, D. M., & Charles, S. T. (1999). Taking time seriously: a theory of socioemotional selectivity. *American Psychologist, 54*, 165–181.

Carstensen, L. L., Fung, H., & Charles, S. (2003). Socioemotional selectivity theory and the regulation of emotion in the second half of life. *Motivation and Emotion, 27*, 103–123.

Charles, S. T., Mather, M., & Carstensen, L. L. (2003). Aging and emotional memory: the forgettable nature of negative images for older adults. *Journal of Experimental Psychology: General, 132*, 310–324.

Casey, B. J., Thomas, K. M., Welsh, T., Livnat, R., & Eccard, C. H. (2000). Cognitive and behavioral probes of development using functional magnetic resonance imaging. In M. Ernst & J. M. Rumsey (Eds.), *Functional Neuroimaging in Child Psychiatry* (pp. 155–168). New York: Cambridge University Press.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29*, 162–173.

Crawford, L. E., & Cacioppo, J. T. (2002). Learning where to look for danger: integrating affective and spatial information. *Psychological Science, 13*, 449–453.

Damasio, A. R. (1995). Toward a neurobiology of emotion and feeling: operational concepts and hypotheses. *The Neuroscientist, 1*, 19–25.

Damasio, A. R. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt Brace & Company.

Damasio, H., Bechara, A., Tranel, D., & Damasio, A. R. (1997). Double dissociation of emotional conditioning and emotional imagery relative to the amygdala and right somatosensory cortex. *Society for Neuroscience Abstracts, 23*, 1318.

D'Argembeau, A., & Van der Linden, M. (2005). Influence of emotion on memory for temporal information. *Emotion, 5*(4), 503–507.

Dawson, M., & Schell, D. (2007). The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of Psychophysiology, 3rd edition* (pp. 159–181). New York: Cambridge University Press.

Detari, L. (2000). Tonic and phasic influence of basal forebrain unit activity on the cortical EEG. *Behavioral Brain Research, 115*, 159–170.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*, 71–75.

Dolcos, F., LaBar, K. S., & Cabeza, R. (2004). Dissociable effects of arousal and valence on prefrontal activity indexing emotional evaluation and subsequent memory: an event-related fMRI study. *Neuroimage, 23*, 64–74.

Federal Interagency Forum on Aging-Related Statistics (2004). *Older Americans 2004: Key Indicators of Well-being*. Washington, DC: U.S. Government Printing Office.

Flynn, F. G., Benson, D. F., & Ardila, A. (1999). Anatomy of the insula—functional and clinical correlates. *Aphasiology, 13*, 55–78.

Glascher, J., & Adolphs, R. (2003). Processing of the arousal of subliminal and supraliminal emotional stimuli by the human amygdala. *Journal of Neuroscience, 23*(32), 10,274–10,282.

Glover, G. H., & Law, C. S. (2001). Spiral–in/out BOLD fMRI for increased SNR and reduced susceptibility artifacts. *Magnetic Resonance in Medicine, 46*, 515–522.

Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage, 9*, 416–429.

Grober, E., & Sliwinski, M. (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical and Experimental Neuropsychology, 13*, 933–949.

Gu, Z., Wortwein, G., Yu, J., & Perez–Polo, J. R. (2000). Model for aging in the basal forebrain cholinergic system. *Antioxidants & Redox Signaling, 2*, 437–447.

Gunning-Dixon, F. M., Gur, R. C., Perkins, A. C., et al. (2003). Age-related differences in brain activation during emotional face processing. *Neurobiology of Aging, 24*, 285–295.

Hamann, S. B., Ely, T. D., Grafton, D. T., & Kilts, C. D. (1999). Amygdala activity related to enhanced memory for pleasant and aversive stimuli. *Nature Neuroscience, 2*, 289–293.

Hart, S., Sarter, M., & Berntson, G. G. (1999). Cholinergic inputs to the medial prefrontal cortex mediate potentiation of the cardiovascular defensive response by the anxiogenic benzodiazepine receptor partial inverse agonist FG 7142. *Neuroscience, 94*, 1029–1038.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: a review and a new view. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory, Vol. 22* (pp. 193–225). San Diego, CA: Academic Press.

Hornak, J., O'Doherty, J., Bramham, J., et al. (2004). Reward-related reversal learning after surgical excisions in orbito-frontal or dorsolateral prefrontal cortex in humans. *Journal of Cognitive Neuroscience, 16*, 463–478.

Ito, T. A., & Cacioppo, J. T. (2000). Electrophysiological evidence of implicit and explicit categorization processes. *Journal of Experimental Social Psychology, 36*, 660–676.

Ito, T. A., Cacioppo, J. T., & Lang, P. J. (1998). Eliciting affect using the International Affective Picture System: trajectories through evaluative space. *Personality and Social Psychology Bulletin, 24*, 855–879.

Jasmin, L., Burkey, A. R., Granato, A., & Ohara, P.T. (2004). Rostral agranular insular cortex and pain areas of the central nervous system: a tract-tracing study in the rat. *Journal of Comparative Neurology, 468*, 425–440.

Jolkkonen, E., Miettinen, R., Pikkarainen, M., & Pitkanen, A. (2002). Projections from the amygdaloid complex to the magnocellular cholinergic basal forebrain in rat. *Neuroscience, 111*, 133–149.

Kaufman, J., Birmaher, B., Brent, D. A., Ryan, N. D., & Rao, U. (2000). K-SADS-PL. *Journal of American Academy of Child & Adolescent Psychiatry, 39*, 1208.

Kilgard, M. P., Pandya, P. K., Vazquez, J., Gehi, A., Schreiner, C. E., & Merzenich, M. M. (2001). Sensory input directs spatial and temporal plasticity in primary auditory cortex. *Journal of Neurophysiology, 86*, 326–338.

Kluver, H., & Bucy, P. C. (1938). An analysis of certain effects of bilateral temporal lobectomy in the rhesus monkey with special reference to "psychic blindness." *Journal of Psychology, 5*, 33–54.

Kluver, H., & Bucy, P. C. (1939). Preliminary analysis of functions of the temporal lobes in monkeys. *Archives of Neurology and Psychiatry, 42*, 979–1000.

Knox, D., Sarter, M., & Berntson, G. G. (2004). Visceral afferent bias on cortical processing: role of adrenergic afferents to the basal forebrain cholinergic system. *Behavioral Neuroscience, 118*, 1455–1459.

Koenigs, M., Tranel, D., & Damasio, A. R. (2007). The lesion method in cognitive neuroscience. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of Psychophysiology* (pp. 139–158). New York: Cambridge University Press.

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*(5), 937–945.

LaBar, K. S., LeDoux, J. E., Spencer, D. D., & Phelps, E. A. (1995). Impaired fear conditioning following unilateral temporal lobectomy in humans. *Journal of Neuroscience, 15*(10), 6846–6855.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1999). *International Affective Picture System (IAPS): Instruction Manual and Affective Ratings (Tech. Rep. No. A–4).* Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

Larsen, J. T., McGraw, A. P., Mellers, B. A., & Cacioppo, J. T. (2004). The agony of victory and the thrill of defeat: mixed emotional reactions to disappointing wins and relieving losses. *Psychological Science, 15*, 325–330.

Larsen, J. T., Norris, C. J., McGraw, A. P., Hawkley, L. C., & Cacioppo, J. T. (2009). The evaluative space grid: a single-item measure of positivity and negativity. *Cognition and Emotion, 23*, 453–480.

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

Leigland, L. A., Schulz, L. E., & Janowsky, J. S. (2004). Age related changes in emotional memory. *Neurobiology of Aging, 25*, 1117–1124.

Levenson, R. W., Carstensen, L. L., Friesen, W. V., & Ekman, P. (1991). Emotion, physiology, and expression in old age. *Psychology and Aging, 6*, 28–35.

Lezak, M. D. (1995) *Neuropsychological Assessment*, 3rd edition. New York: Oxford University Press.

Loring, D., Marin, R., & Meador, K. (1990). Psychometric construction of the Rey-Osterreith complex figure. *Archives of Clinical Neuropsychology, 5*, 1–14.

Martire, L. M., Schulz, R., Mittelmark, M. B., & Newsom, J. T. (1999). Stability and change in older adults' social contact and social support: the cardiovascular health study. *Journals of Gerontology, 54B*; S302–S311.

Masi, C. M., Hawkley, L. C., Rickett, E. M., & Cacioppo, J. T. (2007). Respiratory sinus arrhythmia and diseases of aging: obesity, diabetes mellitus, and hypertension. *Biological Psychology, 74*, 212–223.

Mather, M., Canli, T., English, T., et al. (2004). Amygdala responses to emotionally valenced stimuli in older and younger adults. *Psychological Science, 15*, 259–263.

McCardle, J. J., Hamgami, F., Jones, K., et al. (2004). Structural modeling of dynamic changes in memory and brain structure using longitudinal data from the normative aging study. *Journal of Gerontology: Psychological Sciences, 59B*, P294–P304.

McGaugh, J. L. (2004). The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience, 27*, 1–28.

Mesulam, M. M., Mufson, E. J., Wainer, B. H., & Levey, A. I. (1983). Central cholinergic pathways in the rat: an overview based on an alternative nomenclature (Ch1–Ch6). *Neuroscience, 10*, 1185–1201.

Miasnikov, A. A., McLin, D. 3rd, & Weinberger, N. M. (2001). Muscarinic dependence of nucleus basalis induced conditioned receptive field plasticity. *Neuroreport, 12*, 1537–1542.

Noesselt, T., Driver, J., Heinze, H. J., & Dolan, R. (2005). Asymmetrical activation in the human brain during processing of fearful faces. *Current Biology, 15*(5), 424–429.

Nolen-Hoeksema, S., & Ahrens, C. (2002). Age differences and similarities in the correlates of depressive symptoms. *Psychology and Aging, 17*, 116–124.

Norris, C. J. (2004). *Exploring the Negativity Bias: A Social Neuroscience Perspective*. Unplublished dissertation, University of Chicago, December, 2004.

Norris, C. J., Chen, E. E., Zhu, D. C., Small, S. L., & Cacioppo, J. T. (2004). The interaction of social and emotional processes in the brain. *Journal of Cognitive Neuroscience, 16*, 1818–1829.

Petersen, R. C., Doody, R., Kurz, A., et al. (2001). Current concepts in mild cognitive impairment. *Archives of Neurology, 58*, 1985–1992.

Pfeiffer, E. (1975). Short portable mental status questionnaire for assessment of organic brain deficit in elderly patients. *Journal of the American Geriatrics Society, 23*, 433–441.

Phelps, E. (2006). Emotion and cognition: insights from studies of human amygdala. *Annual Review of Psychology, 57*, 27–53.

Phelps, E. A., LaBar, K. S., & Spencer, D. D. (1997). Memory for emotional words following unilateral temporal lobectomy. *Brain and Cognition, 35*(1), 85–109.

Phelps, E. A., O'Connor, K. J., Gatenby, J. C., Grillon, C., Gore, J. C., & Davis, M. (2001). Activation of the human amygdala to a cognitive representation of fear. *Nature Neuroscience, 4*, 437–441.

Power, A. E. (2004). Muscarinic cholinergic contribution to memory consolidation: with attention to involvement of the basolateral amygdala. *Current Medical Chemistry, 11*, 987–996.

Preston, A. R., Thomason, M. E., Ochsner, K. N., Cooper, J. C., & Glover, G. H. (2004). Comparison of spiral-in/out and spiral-out BOLD fMRI at 1.5 and 3 T. *NeuroImage, 21*, 291–301.

Pribram, K. H., & Melges, F. T. (1969). Psychophysiological basis of emotion. In P. J. Vinken, & G. W. Bruyn (Eds.), *Handbook of Clinical Neurology* (pp. 316–342). Amsterdam: North-Holland.

Radloff, L. S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385–401.

Richards, J. B., Papaioannou, A., Adachi, J. D., et al. (2007). Effect of selective serotonin reuptake inhibitors on the risk of fracture. *Archives of Internal Medicine, 67*, 188–194.

Rowe, J. W., & Kahn, R. L. (1998). *Successful Aging*. New York: Random House.

Salthouse, T. A. (2001). Structural models of the relations between age and measures of cognitive functioning. *Intelligence, 29*, 93–115.

Sarter, M., & Bruno, J. P. (2004). Developmental origins of the age-related decline in cortical cholinergic function and associated cognitive abilities. *Neurobiology of Aging, 25*, 1127–1139.

Sarter, M., Bruno, J. P., & Givens, B. (2003). Attentional functions of cortical cholinergic inputs: what does it mean for learning and memory? *Neurobiology of Learning and Memory, 80*, 245–256.

Schafer, M. K., Eiden, L. E., & Weihe, E. (1998). Cholinergic neurons and terminal fields revealed by immunohistochemistry for the vesicular acetylcholine transporter. I. Central nervous system. *Neuroscience, 84*, 331–359.

Schocken, D. D., & Roth, G. S. (1977). Reduced badrenergic receptor concentrations in ageing man. *Nature, 267*, 856–858.

Simpson, D. M., & Wicks, R. (1988). Spectral analysis of heart rate indicates reduced baroreceptor-related heart rate variability in elderly persons. *Journal of Gerontology, 43*, M21–M24.

Sivan, A. B. (1992). *Revised Visual Retention Test*, 5th edition. New York: The Psychological Corporation.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological Methodology 1982* (pp. 290–312). Washington DC: American Sociological Association.

Spitzer, R. L., Williams, J. B., Gibbon, M., & First, M. B. (1992). The structured clinical interview for DSM–III–R (SCID): I. History, rationale, and description. *Archives of General Psychiatry, 49*, 624–629.

Stern, P., & Carstensen, L. (2000). *The Aging Mind*. Washington, D.C.: National Academy Press.

Talairach, J., & Tournoux, P. (1988). *Co-Planar Stereotaxic Atlas of the Human Brain: 3D Proportional System: An Approach to Cerebral Imaging.* New York, New York: Georg Thieme Verlag.

Tessitore, A., Hariri, A. R., Fera, F., et al. (2005). Functional changes in the activity of brain regions underlying emotion processing in the elderly. *Psychiatry Research, 139*, 9–19.

von Hippel, W., & Dunlop, S. M. (2005). Aging, inhibition, and social inappropriateness. *Psychology and Aging, 20*, 519–523.

Wager, T. D., Hernandez, L., Jonides, J., & Lindquist, M. (2007). Elements of functional neuroimaging. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of Psychophysiology, 3rd edition* (pp. 19–55). New York: Cambridge University Press.

Ward, B. D. (2001). *Deconvolution Analysis of FMRI Time Series Data* (Technical Report). Milwaukee, Wisconsin: Biophysics Research Institute, Medical College of Wisconsin.

Wegener, D. T., Petty, R. E., & Klein, D. J. (1994). Effects of mood on high elaboration attitude change: the mediating rle of likelihood judgments. *European Journal of Social Psychology, 24*, 25–43.

Williams, C. L., & Clayton, E. C. (2001). Contribution of brainstem structures in modulating memory storage processes. In P. E. Gold, J. L. McGaugh, & W. T. Greenough (Eds.), *Memory Consolidation: Essays in Honor of James L. McGaugh* (pp 141–163). Washington, DC: American Psychological Association.

Williams, L. M., Brown, K. J., Palmer, D., et al. (2006). The mellow years? Neural basis of improving emotional stability over age. *The Journal of Neuroscience, 26*, 6422–6430.

Winston, J. S., Gottfried, J. A., Kilner, J. M., & Dolan, R. J. (2005). Integrated neural representations of odor intensity and affective valence in human amygdala. *Journal of Neuroscience, 25*(39), 8903–8907.

Wood, W., & Conway, M. (2006). Subjective impact, meaning making, and current and recalled emotions for self-defining memories. *Journal of Personality, 74*, 811–845.

CHAPTER 18
Social Neuroscience and the Soul's Last Stand

Joshua D. Greene

> "Eventually, as brain imaging is refined, the picture may become as clear and complete as those see-through exhibitions, at auto shows, of the inner workings of the internal combustion engine. At that point it may become obvious to everyone that all we are looking at is a piece of machinery, an analog chemical computer, that processes information from the environment. 'All,' since you can look and look and you will not find any ghostly self inside, or any mind, or any soul."
>
> —Tom Wolfe, "Sorry, But Your Soul Just Died" *Forbes,* 1996

Most people are dualists (Bloom, 2004). Intuitively, we think of ourselves not as physical devices but as immaterial minds or souls housed in physical bodies. Most experimental psychologists and neuroscientists disagree, at least officially. The modern science of mind proceeds on the assumption that the mind is simply what the brain does. We don't talk much about this, however. We scientists take the mind's physical basis for granted. And among the general public, it's a touchy subject. So why bring it up?

We scientists, of course, have our own touchy subjects. One of them concerns the value of neuroscientific research in psychology. Many argue that neuroscience, and brain imaging in particular, is highly overrated (Uttal, 2003). Bloom (2006) attributes the seductive power of neuroscience to our intuitive dualism: "We intuitively think of ourselves as non-physical, and so it is a shock, and endlessly interesting, to see our brains at work in the act of thinking." Bloom, like many experimental psychologists, worries that we are spending millions of dollars on flashy experiments that do little to expand our knowledge of the mind but that instead prompt us to contemplate our ontological navels.

I believe that our ontological navels are in desperate need of contemplation. On some level, we appreciate the urgency of this enterprise, which is why so many psychologists—psychologists who are ardent non-dualists—are fascinated by neuroscience. But the dominant conception of psychological research and its aims obliges us to regard our fascination as an irrelevant distraction. We are forced into a kind of doublethink by which our deeper motivations are at odds with our official reasons for doing what we're doing. Like all scientists, neuroscientists who study mental phenomena are uncovering details. And as scientists we are supposed to be able to say why it matters if the details turn out one way rather than another. But when we try to explain why neuroscientific details matter to psychology, our work often sounds like either an overpriced substitute for more traditional behavioral research or a plodding exercise in "brain mapping." What, then, are we really trying to do? And is it worth doing?

What we really want, I think, is to see the mind's clockwork "as clear and complete as those see-through exhibitions at auto shows." That's not all we're after, of course. We'd like to cure diseases and do other patently useful things. But the promise of useful applications is not what fascinates us. Our fascination is existential. We are hooked on the idea of understanding ourselves in transparently mechanical terms. But a strange feature of this impulse to see the mind's clockwork is that, so far as this impulse is concerned, the clockwork's details are almost irrelevant. We don't care how it works, exactly. We just want to see it in action. Is that foolish? I don't think so. On the contrary, when we think about how our minds work more generally, this bare yearning to perceive the mechanical details of our minds—whatever they happen to be—makes perfect sense.

There are different ways of knowing (Kahneman, 2003; Lieberman, Gaunt, Gilbert, & Trope, 2002). It's one thing to know something intellectually, to believe it in a thin, abstract sort of way—to say "yes" when asked if it's true and not be lying. It's quite another thing to know something in a deep way, to have one's knowledge woven into the fabric of one's worldview, guiding one's thoughts and actions implicitly. Take, for example, the events of September 11th. Most Americans were shocked that such a thing could happen. But why were people so surprised? Eight years earlier, Islamic terrorists attempted to destroy the very same buildings and nearly succeeded. Between 1993 and 2001, every rational person "knew" that America was vulnerable to a large-scale terrorist attack. But it took September 11th to make people really *know.* To take another gruesome example, consider the need people sometimes feel to see the body of someone who has died. After reading the telegram, you may "know" that your long-lost brother is dead, but there is a kind of closure that comes only with seeing the body. The opposite happens at the movies. When the star-crossed lovers, locked in each others' arms, tumble tragically into the lava pit, you "know" that it's only a movie, but tell that to your thumping heart, your tearing eyes, and the parts of your brain that control them. All of this makes perfect evolutionary sense. The basic structure of our brains was in place long before we acquired the ability to use language and with it the ability to acquire beliefs independent of sensory experience. There is a reason why we humans, who specialize in believing in things unseen, still insist that "seeing is believing."

Our self-knowledge may be similarly fractured. Officially, we scientists already know (or think we know) that dualism is false and that we are simply complex biological machines. But insofar as we know this, we know this in a thin, intellectual way. We haven't *seen* the absence of the soul. Rather, we have inferred its absence, based on the available evidence and our background assumptions about what makes one scientific theory better than another. But to truly, deeply believe that we are machines, we must see the clockwork in action. We've all heard that the soul is dead. Now we want to see the body. This is what modern neuroscience promises to deliver, and it is no small thing.

One may argue that achieving a deeper understanding of ourselves is important in itself, on a par with understanding how the universe began and how life first arose on Earth. But I wish to make a more practical argument for deeper self-knowledge. Like a handful of others (Bloom, 2005; Dawkins, 2006; Dennett, 2006; Harris, 2004), I believe that our intuitive dualism causes a lot of problems. And if anything can talk us out of our dualist tendencies it is neuroscience—more specifically *social neuroscience.* According to Wegner and Gilbert (2000), social psychology has evolved from being a fairly circumscribed science of human social interaction into a sprawling science encompassing all of human subjective experience. If that is right, then the mission of social neuroscience, as the offspring of social psychology and neuroscience, is to understand all of human subjective experience in physical terms. The rise of social neuroscience is the demise of the soul.

My aim in this article is to consider the broader implications of social neuroscience, so conceived. Although I am an unabashed enthusiast, I agree with critics who say that there is a real danger of our wasting precious time and money on misguided research. The challenge, as I see it, is to achieve the kind of short-term incremental progress that journals and funding agencies demand, while honoring our broader, and all but unspoken, philosophical mission. The key, I think, is to make our task the *functional decomposition of the brain*—that is, breaking down complex psychological processes into simpler processes that are associated with different parts of the brain. This is not a new idea. It is not even close to being a new idea. It is what the best cognitive neuroscientists and, more recently, social neuroscientists have been doing all along. But it is an idea that many people don't seem to get, especially people who come to social neuroscience without training in experimental psychology. (Witness the proliferation of scientifically undermotivated brain imaging experiments [Cacioppo et al., 2003].) And because social neuroscience is so inherently fascinating to our dualist minds, it is possible to get away with not getting it. In what follows I will argue that functionally decomposing the social brain is a worthy thing to do in the short-term and, perhaps, one of the most worthy things that we social scientists could ever do in the long-term. As I make my case, I will use my own work on moral judgment as an illustrative example. There may be bad reasons for doing this (laziness, egocentricism), but there is at least one good reason. When it comes to undoing dualism, the neuroscientific study of moral judgment occupies a unique position.

These days, even the most ardent dualists recognize that we have brains and that our brains must do *something*. In recent decades we've learned that brains do many things that are historically within the province of the soul: perception, memory, the production and comprehension of language, and so forth. The soul has, as it were, "outsourced" these operations to the brain. This outsourcing process, still ongoing, raises a question: How many of the soul's functions can be taken up by the brain before the soul is completely out of a job? In other words, what is the soul's "core competence?" The answer, I believe, is *moral judgment*. After all, in many religious traditions it is the quality of a soul's moral judgment and character that determines where it ends up, either permanently or on the next go-around. Thus, if the soul is not in the moral judgment business, it is not in any business at all. And, thus, what it would take to send the soul packing for good is a purely physical account of how the human mind does its moral business. If our goal is to determine once and for all whether there is a soul in there, there is no better place to start than with the neuroscience of moral judgment.

The dual-process model of moral judgment

Consider the following moral dilemma, which we will call the *crying baby* case:

> It is wartime, and you and some of your fellow villagers are hiding from enemy soldiers in a basement. Your baby starts to cry, and you cover your baby's mouth to block the sound. If you remove your hand, your baby will cry loudly, the soldiers will hear, and they will find you and the others and kill everyone they find, including you and your baby. If you do not remove your hand, your baby will smother to death. Is it okay to smother your baby to death to save yourself and the other villagers?

This is a difficult question. People take a relatively long time to answer, and there is no consensus about the right answer (Greene, Nystrom, Engell, Darley, & Cohen, 2004). Why is this question so difficult? And what's going on in people's heads when they are deciding?

According to my dual-process theory (Greene et al., 2001, 2004, 2008), it goes something like this: On the one hand, we have an intuitive emotional response to the thought of smothering one's own baby (or anyone else) that makes us say, "No! It's wrong!" On the other hand, there's a different voice in our heads, a more dispassionate and controlled voice that says, "But there's nothing to be gained and much to be lost by not acting. The baby will die no matter what. You ought to save yourself and the others." These two voices, the intuitive emotional voice and the controlled "cognitive"[1]

[1] In some cases, the word "cognitive" refers to information processing in general (as in "cognitive science"), whereas in other cases, it refers to a kind of information processing that is contrasted with emotional or affective processes. Here and elsewhere, I place "cognitive" in quotation marks to indicate the latter usage. As I explain elsewhere (Greene, 2007), I believe that the latter usage refers to a natural category of processes involving representations that are inherently neutral but that may be contingently connected to valenced representations. This allows for the production of behavior that is both flexible and goal-directed.

voice, fight it out in your head, until one of them wins and you render your judgment. This theory, despite its introspective plausibility, stands in tension with two leading schools of thought in moral psychology, one of which denies that emotions play an important role in the moral judgments of mature adults (Kohlberg, 1969), while the other denies that moral reasoning and controlled "cognitive" processes play a direct causal role in shaping ordinary people's moral decisions (Haidt, 2001).

We conducted a simple behavioral experiment to test our dual-process theory. People responded to dilemmas like the *crying baby* dilemma under normal conditions and under cognitive load (i.e., while simultaneously performing a distracting task that requires cognitive resources). Once again, our claim is that there is an intuitive emotional voice that says "No! Don't!" as well as a controlled, "cognitive" voice that says, "Please, go ahead." And if that is correct, then a drain on limited cognitive resources should interfere with the processes that are pushing for "yes" but not with the more intuitive processes that are pushing for "no." In the best case, we would expect the imposition of a cognitive load to make people say "no" more often by selectively interfering with the "yes"-friendly processes. And if the load manipulation is not strong enough to actually change people's judgments, then it might at least make "yes" answers slower, without slowing down the "no" answers.

This is exactly what we found (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008). When people responded to dilemmas like the *crying baby* case under cognitive load, the load made people slower when they endorsed harming someone in the name of the greater good but had no effect when they said that it would be wrong to cause the harm. Thus, these results support our dual-process theory of moral judgment: If it were all about intuitive emotional responses, then there would be no reason for the cognitive load to slow down "yes" answers any more than "no" answers. And if it were all about controlled processes, then, again, we would have no reason to expect a difference in reaction times between "yes" and "no" answers, as all answers would be slowed by the load. Only a dual-process theory makes sense of this interaction, whereby cognitive load has a different effect on reaction time depending on the content of the judgment.

This experiment tells us that the psychological processes pushing for "no" answers in these cases are rather automatic, charging ahead, impervious to the cognitive load. But this study doesn't necessarily tell us whether these automatic responses are emotional.[2] A different experiment, conducted by Valdesolo and DeSteno (2006), addresses this question. They presented people with two moral dilemmas, which we'll call the *switch* and *footbridge* dilemmas. In the *switch* dilemma (elsewhere referred to as the *trolley* dilemma; (Greene et al., 2001) and the *bystander* dilemma (Thomson, 1986)), a runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save these people is to hit a switch that will turn the trolley onto a side-track, where it will run over and kill one person instead of five. Is it okay to turn the trolley to save five people at the expense of one? Most people say "yes" (Greene et al., 2001; Petrinovich, O'Neill, & Jorgensen, 1993). This case contrasts with the *footbridge* (Thomson, 1986) dilemma: As before, a runaway trolley threatens to kill five people, but this time you are standing next to a large stranger on a footbridge spanning the tracks, in between the oncoming trolley and the five people. The only way to save the five people is to push this stranger off the bridge and onto the tracks below. (You're not big enough to block the trolley by jumping yourself.) He will die as a result, but his body will stop the trolley from reaching the others. Is it okay to save the five people by pushing this stranger to his death? Most people say "no" (Greene et al., 2001; Petrinovich et al., 1993). One might suppose that the action proposed in the *footbridge* case triggers more of an intuitive emotional response than the action proposed in the *switch* case,

[2] It depends on how one defines "emotion." If a process that is automatic and valenced is necessarily emotional, then this study may be sufficient to implicate emotion.

which would explain why most people tend to say "no" to the *footbridge* case and "yes" to the *switch* case. (Here, too, we are supposing that the controlled process favors the greater good, i.e. the "yes" response.) Valdesolo and DeSteno tested this hypothesis using an emotion induction paradigm. They presented two groups of people with the *trolley* and *footbridge* dilemmas. The control group, before responding to these dilemmas, watched an emotionally neutral film clip (5 minutes from a documentary about a Spanish village). The experimental group watched a 5-minute comedy clip from *Saturday Night Live*. Valdesolo and DeSteno reasoned as follows: If the typical "no" responses to the *footbridge* dilemma are driven by negative emotional responses, then hitting people with a dose of positive emotion (using the comedy clip) should counteract the negative emotional response and make those participants more likely to say "yes." But if the typical "yes" response to the *switch* case is not driven by emotional processes, then watching the clip should have no effect on people's responses to that case. And, of course, the control group should experience no change as a result of the neutral film. This is what Valdesolo and Desteno found. The neutral film had no effect at all, and the *Saturday Night Live* clip had no effect on people's responses to the *trolley* dilemma. But the comedy clip did have a significant effect on people's responses to the *footbridge* dilemma, tripling the number of people who approved of pushing the man off of the footbridge. Thus, it seems that emotional response plays a key role in producing people's negative judgments to the *footbridge* case.

These two studies support the dual-process model of moral decision making, providing evidence that moral judgment involves both intuitive emotional responses and more controlled "cognitive" processes. Perhaps surprisingly, these two studies were designed to bolster the conclusions drawn from previous neuroimaging studies. In one of these studies (Greene et al., 2001), my collaborators and I presented people with a series of moral dilemmas, including versions of the *switch* and *footbridge* dilemmas. In response to cases like the *footbridge* dilemma,[3] people exhibited relatively higher levels of activity in brain regions associated with emotion and social cognition (medial prefrontal cortex, posterior cingulate cortex, and superior temporal sulcus). In contrast, responses to cases like the *switch* case were associated with increased activity in brain regions associated with classically "cognitive" functions such as working memory (dorsolateral prefrontal cortex [(DLPFC)] and corresponding regions in the parietal lobes). This double dissociation between activity in brain regions associated with emotion, on the one hand, and more classically "cognitive" brain regions, on the other, provided the first piece of evidence in support of our dual-process theory of moral judgment. A second study (Greene et al., 2004) focused on difficult cases like the *crying baby* case. We found that cases like this, relative to easier cases, were associated with increased activity in the anterior cingulate cortex (ACC) and in the DLPFC. Previous work suggests that the ACC plays a role in detecting response conflict (the simultaneous activation of two incompatible behavioral responses; Botvinick, Nystrom, Fissell, Carter, & Cohen, 1999; Botvinick, Braver, Barch, Carter, & Cohen, 2001; MacDonald, Cohen, Stenger, & Carter, 2000), whereas other work has implicated the DLPFC in cognitive control functions aimed at resolving response conflict (Kerns et al., 2004; MacDonald et al., 2000). Thus, increased activity in the ACC and DLPFC is what one would expect if cases like the *crying baby* dilemma create a conflict between two incompatible responses that must be resolved. We also found that when people responded to such cases with a utilitarian judgment (favoring the greater good, even at the cost of harming

[3] In our initial investigation (Greene et al., 2001), we used a set of three criteria to distinguish dilemmas like the footbridge case (which we called "personal") from dilemmas like the switch case (which we called "impersonal"). The personal/impersonal distinction was devised as a "first cut" for identifying the psychologically salient features that distinguish these two dilemmas. Based on more recent research (Greene et al., 2009)., I now believe that the personal/impersonal distinction should be replaced by a more cognitively sophisticated set of criteria concerning the nature of the agent's intention and the kind of force applied by the agent.

someone), they exhibited increased activity in the DLPFC. This is what we would expect if the utilitarian "yes" responses were driven by controlled "cognitive" processes.

Similarly to the cognitive load and emotion induction studies described above, these two neuroimaging studies support the dual-process theory of moral judgment, implicating both intuitive emotional responses and controlled "cognitive" responses in moral decision-making. But these neuroimaging studies, compared to the behavioral studies that followed, were a lot more expensive. One has to wonder, then, if they were worth it. Are we getting any additional bang for our neuroscience buck?

Who needs social neuroscience? the scientist's answer

Neuroscientific data, and neuro-imaging data in particular, offer new insights into the relationships among the myriad psychological processes identified using more traditional means. Consider the following remarks from Daniel Gilbert (1999), introducing an edited volume devoted to dual-process theories in social psychology:

> The neuroscientist who says that a particular phenomenon is the result of two processes usually means to say something unambiguous—for example, that the inferior cortex does one thing, that the limbic system does another… In such instances the phrase "dual processes" refers to the activities of two different brain regions that may be physically discriminable, and the neuroscientist says there are "two processes" because the neuroscientist is talking about things that can be counted. But few of the psychologists whose chapters appear in this volume would claim that the dual processes in *their* models necessarily correspond to the activity of two distinct brain structures (pg. 3).

Oh, what a decade can bring. Now the social psychologists are also neuroscientists, and they have started counting. This transformation, of course, has not yielded precise process counts, and perhaps it never will, given that processes are inherently fuzzy-boundaried things. But the advent of social neuroscience has given students of social cognition, best known for their analytical splitting, the opportunity to engage in an exciting new kind of lumping. By looking directly at the brain we can see whether the processes that Psychologist A has been dutifully teasing apart in her quest to understand Behavior X are, in fact, some of the same processes that Psychologist B has been teasing apart in his quest to understand Behavior Y. Humpty Dumpty may never fit back together again, but thanks to neuroscience, he may escape being ground to fine dust.

Thus, neuroscience holds the promise of a new kind of synthetic psychology. It also holds the promise of a different kind of analytic psychology. At present, there is a gulf between the high-level language of the mind ("belief," "impulse," "thought," "attitude," "emotion") and the low-level language of the brain ("lobe," "gyrus," "neural activity"). We can correlate things like "attitudes" with things like "neural activity," but we have only the foggiest picture of how the former arise from the latter. How will we acquire a clearer one? We don't know how, exactly, or we'd have already done it, but it is hard to imagine that our learning process won't involve a fair amount of "top-down" investigation—that is, using what we know about psychology to map the brain in a psychologically meaningful way. An investment in brain-mapping (if done well) will pave the way for a deeper science of mind that is seamlessly integrated with the physical sciences (but *see* Uttal, 2005).

In short, social neuroscience, at least in the long-run, is likely to yield scientific theories that are richer and more coherent than the ones that to which social psychologists are accustomed. This is the standard justification for doing what we're doing, and it is a good one. It is the one that we relay, in various forms, to the editors of journals and administrators of funding agencies, and that is as it should be. But in the long-long run, the greatest value of social neuroscience may lie elsewhere.

Who needs social neuroscience? the philosopher's answer

When I tell people that I study the neuroscience of moral decision making, I am often asked, "Where is the brain's moral center?" Apparently, people find the idea of a "moral center" in the brain very compelling. There may be many reasons for this, but I think it has something to do with the fact that a center, unlike a distributed system, need not have *parts*. The moral center of popular conception, I'm guessing, is not a computational system housing an array of dissociable subsystems that perform relatively simple, complementary functions. It is instead more like a *portal,* out of which fully formed moral thoughts emerge. A moral portal is what a dualist, when confronted with the fact of the brain, naturally imagines the moral brain to be. The portal theory acknowledges that moral judgments must get into our brains somehow, while leaving open the possibility that their true origin lies beyond.

This dualist conception of moral judgment breaks down when the moral brain is functionally decomposed. This involves two things. First, the process of moral decision making is itself broken down into distinct psychological processes. Second, it is shown that distinct parts of the brain are respectively responsible for carrying out these distinct processes. Both aspects of functional decomposition are necessary if they are to count against dualism. If the component processes are distinguished psychologically, but not attributed to different parts of the brain, then we are free to think of these psychological processes as operations of a multifaceted soul that renders its judgments before transmitting them to the material realm via the brain's moral portal. Consider, for example, the cognitive load and emotion induction studies described above. The cognitive load study tells us that, when confronted with cases like the *crying baby* dilemma, we have an intuitive response that tells us one thing ("Don't smother the baby!") and a more controlled response that tells us the opposite ("But there is nothing to lose and much to gain by smothering the baby."). Valdesolo and DeSteno's (2006) emotion induction study teaches a complementary lesson, demonstrating that our judgments, in some cases more than others, are driven by emotional responses. These results provide evidence for the dual-process model of moral judgment, but they do little, if anything, to dispel dualism. Dualists and materialists alike are familiar with having emotional impulses, and with resisting them. A dualist can happily regard these mental operations as operations of the soul, and these behavioral experiments provide no evidence to the contrary.

The situation changes little if we point to multiple brain regions that are "involved" in moral judgment, but say nothing about how these various brain regions contribute to the decision-making process. We can stick someone in a brain scanner, have that person make moral judgments, and then report on the brain regions that exhibit increased activation, but this will do nothing to embarrass a dualist. He can happily regard this smattering of brain regions as a distributed portal, an archipelago of mysterious mind-body interaction.

It is only when we ascribe distinct psychological subfunctions of moral decision-making to distinct physical parts of the brain that moral decision-making starts to look like a mechanical process. And this is exactly what the neuroimaging experiments described above do. Our claim is not that the ACC is "involved" in moral judgment in some nebulous way. If our theory (Greene et al., 2004) is correct, then the ACC performs the specific function of detecting conflict between competing responses, both in moral contexts and in other contexts that have nothing to do with morality *per se* (Botvinick et al., 2001; Cohen, 2005). Our theory likewise attributes a control function to the DLPFC and an emotional function to the medial prefrontal cortex.[4] Of course, the dualist can dig in his heels, even in the face of functional decomposition of this kind. He might imagine, for example,

[4] More recent neuropsychological data provide additional support for our conclusion that parts of the medial prefrontal cortex generates or mediates emotional responses that drive non-utilitarian judgments in response to moral dilemmas such as the *footbridge* case. Mendez et al. (2005) found that patients with frontotemporal dementia (known for their "emotional blunting" and ventromedial prefrontal damage) were disproportionately likely to approve of pushing the man off of the *footbridge*. Similar results have been obtained in patients with more well-defined ventromedial prefrontal lesions (Ciaramelli et al., 2007; Koenigs et al., 2007).

that each of these brain regions simply receives transmissions from different functional parts of the soul: The ACC receives transmissions from the part of the soul that detects conflict, and so on. But this can't go on forever. As our body of knowledge grows, the moral brain will be decomposed into smaller and smaller physical parts, associated with narrower and narrower psychological functions, until the corresponding bits of soul are reduced to an array of manifestly superfluous micro-ghosts.

As Wolfe's remarks suggest, the soul will officially expire when the mechanics of the moral mind become transparent. I believe that the death of the soul may prove to be one of psychology and neuroscience's most lasting contributions. That is, if we are around long enough to get the job done.

Dualism: what is at stake?

Why does it matter if people are dualists? As long as we scientists know what we need to know to do our work, is it our business, or in anyone's best interest, to provide compelling demonstrations of the fact that we have no souls? I think that it is. Dualist beliefs may be harmless enough most of the time, but they divide us in destructive ways, enable us to do some of the worst things that we do, and may ultimately lead to our demise.

Consider, once again, the events of September 11th. Nineteen men killed nearly 3,000 people, setting in motion a series of events that, in addition to killing many thousands more people, has destabilized the Middle East at a time when nuclear weapons are becoming increasingly accessible. Those 19 men destroyed their bodies, along with thousands of others, believing that they—their souls—would go on to enjoy a pleasant post-corporeal existence. Of course, it is possible that their beliefs about the next world played no role in their decision to leave this one, but that seems unlikely. Rather, as others have noted (Dawkins, 2006; Harris, 2004), it seems that their beliefs about the afterlife enabled them to do what they did.

Here in the West, our leaders are dualists, too. George Bush does not speak openly about the details of his metaphysical worldview, but as an evangelical Christian he presumably believes that all people have souls and that the souls of those who share his faith will be saved, whereas the remainder will spend eternity in hell. This remainder presumably includes the vast majority of Muslims. Thus, transforming a Muslim society into a more Westernized society with greater exposure to Christianity may translate into huge gains in terms of the number of souls saved. The prospect of saving millions of souls may warrant taking large risks. More specifically, it might warrant risking the lives of millions of Muslims, assuming that in the absence of Christian intervention, their souls are doomed. According to one report, over half a million Iraqis have died as a result of the American invasion of Iraq (Tavernise & McNeil, 2006).

Dualism is at the heart of many bioethical debates (Bloom, 2005). It is often said that the abortion controversy is really a debate about when life begins. But "life" is not the real issue, as everyone agrees that a fertilized human egg, like an unfertilized human egg, is alive. Nor is it a matter of when a developing human acquires significant psychological characteristics such as the ability to feel pain. Most opponents of abortion are perfectly happy to eat animals that are capable of feeling pain and that, more generally, have richer mental lives than human fetuses. Nor is it a matter of destroying potential human life. Both birth control and abstinence rob potential humans of their existence. Rather, the debate over abortion is ultimately a metaphysical one. The question is not whether a fertilized egg is alive, but whether it is host to a "human life"—that is, a human soul. Without a soul in the balance, there is no abortion debate. This is also true for the debates over human stem cell research and euthanasia.

As I (Greene & Cohen, 2004) and others (Bloom, 2005) have argued, certain aspects of our criminal justice system are implicitly dualist. If you ask people why we ought to punish criminals, people most often cite the law's deterrent effect. But when people respond to concrete cases, their judgments are surprisingly insensitive to factors that are relevant to the prevention of future crime (Carlsmith, Darley, & Robinson, 2002). Rather, it seems that people's intuitions about punishment are *retributivist*. We want to punish criminals, not because of the future benefits, but simply as an end in itself. These retributivist tendencies are, I believe, implicitly dualist. If someone has a brain tumor that causes aggressive behavior, people are far more willing to forgive that person. "After all," we say, "It's not *him*, it's *his brain*." When we attribute bad behavior to a purely physical cause (such as a brain tumor), the retributivist impulse fades. Our aim is to punish guilty minds (*mens rea*), not broken brains. (A broken brain may be worth containing, deterring, and rehabilitating, but there is no good reason to punish someone simply for having a broken brain.)

From a neuroscientific perspective, of course, all behavior (good and bad) has purely physical causes, and anyone who does unusually bad things must have something, however subtle, wrong with his brain. Combine this ordinary scientific assumption (all bad behavior is caused by brains that are, in some sense, broken) with people's ordinary assumption about punishment (there is no inherent value in punishing someone for having a broken brain), and we get a very different sort of legal system. We get one focused exclusively on the practical business of preventing future crime, rather than on the metaphysical business of making guilty minds suffer for their sins. In the United States, at least, our prison system is very good at making people suffer, but its merits as a system for preventing future crime are highly questionable (Tonry, 2004). If we were more interested in reducing crime, and less interested in making guilty minds suffer, we might all be better off.

Dualism plays a parallel role in people's thinking about mental illness. Intuitively, we all agree that people with cancer deserve our sympathy and financial support because cancer is a serious medical problem. But if someone is depressed, that person's condition is, to many people at least, just "psychological," and the prescription is to "snap out of it." Dualism draws an illusory distinction between having a weakened body and having a weakened mind.

Finally, dualism may play an important role in people's attitudes concerning the environment. According to a variety of polls, about 40% of Americans believe that we are living in the "end times"—that is, that the world will end relatively soon, at which point all followers of the Christian faith will be swept up into heaven, while the rest of us descend into hell (Sahagun, 2006). If you think that God is going to end the world relatively soon, you are unlikely to be terribly concerned about the level of carbon dioxide in the Earth's atmosphere. Of course, most of the people who are worried about global warming are dualists, too. But to be indifferent to the long-term health of the environment, it certainly helps to believe that our need for it is temporary.

Conclusion

Social neuroscience is exciting, but it is hard for some of us to say why. Most would agree that looking directly into the human brain will, sooner or later, provide us with better theories about how our minds work. But the prospect of better psychological theories, arriving sooner or later, hardly explains the excitement we feel. I believe that social neuroscience is exciting primarily because of its broader philosophical implications, and only secondarily because of the empirical details we expect it to yield. But to speak of social neuroscience's philosophical "implications" is a bit awkward. Officially, we scientists already know that the operations of the mind are the operations of the brain and not those of an immaterial soul. This is, at the very least, our working assumption. In making this assumption, however, we part ways with the rest of humanity, the vast majority of whom explicitly believe that we are souls housed in bodies. Such dualist tendencies are, in my opinion, a major social problem and may become increasingly destructive. If that is correct, then dispelling dualism is serious business, at least as serious as curing cancer, and probably more so. If anything can cure us of our dualist tendencies, it is social neuroscience—the physical science of human experience. By decomposing the social brain into its mechanical components, we can do good science in the conventional sense, but that is, I think, the least of what we're doing. Social neuroscience is, above all else, the construction of a metaphysical mirror that will allow us to see ourselves for what we are and, perhaps, change our ways for the better.

Acknowledgments

Many thanks to Andrea Heberlein and Joe Paxton for helpful suggestions.

References

Bloom, P. (2004). *Descartes' Baby: How the Science of Child Development Explains What Makes Us Human*. New York: Basic Books.

Bloom, P. (2005). Worse than creationism. *American Psychological Society Observer, 18*(10).

Bloom, P. (2006, June, 27, 2006). Seduced by the flickering lights of the brain. *Seed*.

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature, 402*(6758), 179–181.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*(3), 624–652.

Cacioppo, J., Bernston, G., Lorig, T., Norris, C., Rickett, E., & Nusbaum, H. (2003). Just because you're imaging the brain doesn't mean you can stop using your head: A primer and set of first principles. *Journal of Personality and Social Psychology, 85*(4), 650–661.

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology, 83*(2), 284–299.

Ciaramelli, E., Muccioli, M., Ladavas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience, 2*, 84–92.

Cohen, J. (2005). The vulcanization of the human brain: A neural perspective on interactions between cognition and emotion. *Journal of Economic Perspectives, 19*, 3–24.

Dawkins, R. (2006). *The God Delusion*. Boston, MA: Houghton Mifflin.

Dennett, D. C. (2006). *Breaking the Spell: Religion as a Natural Phenomenon*. New York: Viking.

Gilbert, D. T. (1999). What the mind's not. In S. Chaiken & Y. Trope (Eds.), *Dual Process Theories in Social Psychology*. New York: Guilford.

Greene, J., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transacations of the Royal Society of London. Series B, Biological Sciences, 359*(1451), 1775–1785.

Greene, J., Morelli, S., Lowenberg, K., Nysrom, L., & Cohen, J. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*, 1144–1154.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364–371.

Greene, J. D. (2007). The secret joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology: Morality in the Brain* (Vol. 3). Cambridge, MA: MIT Press.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389–400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834.

Harris, S. (2004). *The End of Faith: Religion, Terror, and the Future of Reason*. New York: W. W. Norton & Co.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *The American Psychologist, 58*(9), 697–720.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., 3rd, Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science, 303*(5660), 1023–1026.

Koenigs, M., Young, L., Adolphs, R., et al. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature, 446*, 908–911.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of*

Socialization Theory and Research (pp. 347–480). Chicago: Rand McNally.

Lieberman, M. D., Gaunt, R., Gilbert, D. T., & Trope, Y. (2002). Reflection and reflexion: A social cognitive neuroscience approach to attributional inference. *Advances in Experimental Social Psychology, 34*, 199–249.

MacDonald, A. W., 3rd, Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science, 288*(5472), 1835–1838.

Mendez, M. F., Anderson, E., & Shapira, J. S. (2005). An investigation of moral judgement in frontotemporal dementia. *Cognitive and Behavioral Neurology, 18*(4), 193–197.

Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An empirical study of moral intuitions: Toward an evolutionary ethics. *Journal of Personality and Social Psychology, 64*, 467–478.

Sahagun, L. (2006, June 22, 2006). 'End Times' religious groups want apocalypse soon. *Los Angeles Times*.

Tavernise, S., & McNel, D. (2006, October 11, 2006). Iraqi dead may total 600,000, study says. *New York Times*.

Thomson, J. J. (1986). *Rights, Restitution, and Risk : Essays, in Moral Theory*. Cambridge, MA: Harvard University Press.

Tonry, M. (2004). *Thinking about Crime: Sense and Sensibility in American Penal Culture*. New York: Oxford University Press.

Uttal, W. R. (2003). *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain*. Cambridge, MA: MIT Press.

Uttal, W. R. (2005). *Neural Theories of Mind: Why the Mind–Brain Problem May Never Be Solved*. Mahwah, NJ: Lawrence Erlbaum Associates.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science, 17*(6), 476–477.

Wegner, D., & Gilbert, D. T. (2000). Social psychology—the science of human experience. In H. Bless & J. P. Forgas (Eds.), *Subjective Experience in Social Cognition and Behavior*. Philadelphia, PA: Psychology Press.

CHAPTER 19
Building a Social Brain

Todd F. Heatherton

What do you need to make a social brain? Or what does the brain need to do to allow it to be social? The chapters in this book provide clear evidence for the quick maturation of the field of social neuroscience, which incorporates scholars from widely diverse areas (e.g., social psychology, neuroscience, philosophy, anthropology, economics, sociology, psychiatry, medicine), working together and across levels of analysis to understand fundamental questions about human social nature (Cacioppo et al., 2007). Social neuroscience attempts to identify and characterize the neural mechanisms that support social behavior, broadly defined. From this perspective, the brain has evolved specialized mechanisms for processing information about the social world, including the ability to know ourselves, to know how others respond to us, and to regulate our actions to co-exist with other members of society.

The chapters in this section focus on social emotions, and that is not surprising. Ralph Adolphs (2003) argued that social neuroscience resulted from a sometimes uneasy alliance between evolutionary psychology and social cognition and that the success of the area results from the adoption of neuroscience methods and largely restricting the domain of empirical study to emotional aspects of cognition. After all, thinking about other people entails emotional responses that thinking about, say, vegetables, does not. The social and emotional aspects of the brain are inexorably linked, with the adaptive significance of emotions being closely linked to their social value, and nearly all social interaction produces affective responses. Research in social psychology over the past several decades has established the central role of emotional processes in facilitating social relationships and guiding group behavior. Social emotions, which are complex subjective experiences (e.g., pride, guilt, admiration, jealousy, envy, irritation, and flirtatiousness), serve many important social functions that promote long-term relationships and group stability. From a functional perspective, social emotions enable people to be good group members, thereby avoiding rejection and enhancing their survival and reproduction. Thus, social emotions such as guilt are essential to human social life (Baumeister, Stillwell, & Heatherton, 1994).

Social emotions not only encourage successful relationships by providing incentives to engage in social interactions (e.g., affection, love, feelings of pride or admiration for those with whom we interact), they also increase the likelihood that people will adhere to societal norms and moral values that are necessary for group living. When such norms are violated, people experience negative social emotions (e.g., feelings of guilt, embarrassment, or shame) that subsequently encourage people to act within the bounds of socially acceptable conduct, thereby reducing the risk of social exclusion and promoting positive social interactions. Moreover, long-lasting social emotions (such as remembering an embarrassing moment from adolescence) reduce the likelihood of repeat violations. Importantly, humans have evolved a fundamental need to belong, which encourages behavior that helps people be good group members (Bowlby, 1969; Baumeister & Leary, 1995). Humans are social beings who live in groups. According to the need-to-belong theory, the need for interpersonal attachments is a fundamental motive that has evolved for adaptive purposes. Effective groups shared food, provided mates, and helped care for offspring. As such, human survival has long depended on living within groups; banishment from the group was effectively a death sentence. Baumeister and Leary (1995) argued that the need to belong is a basic motive that activates behavior and influences cognition and emotion, and that it leads to ill effects when not satisfied. Indeed, even today, not belonging to a group increases a person's risk for a number of adverse consequences, such as illnesses and premature death (*see* Cacioppo et al., 2006). Here I argue that an evolutionary need to belong has guided the evolution of a social brain. I propose that building a social brain requires four essential components: self-awareness, theory of mind, threat detection, and self-regulation.

Building the social brain: components

Given the fundamental need to belong, there needs to be a social brain system that monitors for signs of social inclusion/exclusion and alters behavior to forestall rejection or resolve other social problems (Heatherton & Krendl, 2009). Such a system requires four components, each of which is likely to have a discrete neural signature. First, people need self-awareness—to be aware of their behavior so as to gauge it against societal or group norms. Second, people need to understand how others are reacting to their behavior so as to predict how others will respond to them. In other words they need "theory of mind" or the capacity to attribute mental states to others. This implies the need for a third mechanism that detects threat, especially in complex situations. Finally, there needs to be a self-regulatory mechanism for resolving discrepancies between self-knowledge and social expectations or norms, thereby motivating behavior to resolve any conflict that exists. Space limitations preclude a thorough discussion of all brain mechanisms likely to be involved in all of these processes. Here I highlight the most central areas for the theory as well as discuss the implications of the four chapters in this section for the model.

Self-Awareness

Survival in human social groups requires people to monitor their behavior and thoughts in order to assess whether those thoughts and behaviors are in keeping with prevailing group (social) norms. Social neuroscience has made excellent strides in identifying brain regions that are involved in processing information about the self (Heatherton, Kelley, & Macrae, 2004). Both neuro-imaging and patient (lesion) research has identified various regions of the prefrontal cortex (PFC) as being crucial for the normal functioning of self. For example, a series of imaging studies conducted over the past 10 years has documented a substantial role of the medial region of the PFC (mPFC) in processing self-relevant information (Craik et al., 1999; Heatherton et al., 2006; Johnson et al., 2002; Kelley et al., 2002; Macrae et al., 2004; Moran et al., 2006; Schmitz et al., 2004; Ochsner et al., 2004). This region is more active, for example, when people report on their personality traits, make self-relevant judgments about pictures, or retrieve autobiographical memories of past events. The issue of whether the self is somehow "special" is somewhat contentious (*see* Gillihan & Farah, 2005), but the imaging literature is quite clear regarding tasks that involve self-awareness; they activate mPFC in imaging studies (Gusnard, 2005).

It is interesting to note that converging evidence from patient research indicates that frontal lobe lesions, particularly to the mPFC and adjacent structures, have a deleterious effect on personality, mood, motivation, and self-awareness. Patients with frontal lobe lesions show dramatic deficits in recognizing their own limbs, engaging in self-reflection and introspection, and even reflecting on personal knowledge. Indeed, frontal lobe patients are particularly impaired in social emotions (Beer et al., 2003). As Josh Greene notes in his chapter in this volume, people with frontal lobe injuries are also impaired in moral judgments. The convergence of patient and imaging data support the conclusion that mPFC plays a prominent role in self-awareness, a necessary and critical contributor to the experience of social emotions.

I hasten to add that that there is no specific "self" spot of the brain, no single brain region that is responsible for all psychological processes related to self. Rather, psychological processes are distributed throughout the brain, with contributions from multiple subcomponents determining discrete mental activities that come together to give rise to the human sense of self (Turk, Heatherton, Macrae, Kelley, & Gazzaniga, 2003). Various cognitive, sensory, motor, somatosensory, and affective processes are essential to self, and these processes likely reflect the contribution of several cortical and subcortical regions. Josh Greene makes the same point regarding moral emotions. His dual-process model dictates that different brain regions and their associated psychological functions contribute to moral judgments. In his chapter, Greene provides a forceful argument against dualistic conceptions of human nature like those that commonly occur for the soul, or the self for that matter. There is neither homuncular self nor soul. There is a brain that carries out adaptive functions through the activity of various regions that are responsible for the very psychological processes responsible for emotion, cognition, and behavior.

Theory of Mind

In addition to recognizing our own mental states, living harmoniously in social groups requires that we are able to interpret the emotional and mental states of others (Heatherton & Krendl, 2009). For example, social emotions require that we are able to draw inferences about the emotional states of others (even if those inferences are inaccurate). For example, to feel guilty about hurting a loved one, people need to understand that other people have feelings. Similarly, interpersonal distress results from knowing that people are evaluating you (thereby giving rise to emotions such as embarrassment), which at its core means recognizing that other people make evaluative judgments.

The ability to infer the mental states of others is commonly referred to as mentalizing, or having the capacity for theory of mind (ToM). ToM enables the ability to empathize and cooperate with others, accurately interpret other people's behavior, and even deceive others when necessary. The rapidly emerging neuro-imaging literature on ToM has consistently implicated MPFC as a central component of the neural systems that support mentalizing (Amodio & Frith, 2006; Gallagher & Frith, 2003; Macrae et al., 2004). Interestingly, neuro-imaging research has demonstrated that the ability to mentalize relies heavily on similar neural networks engaged in processing self-relevant information—notably mPFC. However, this region of mPFC tends to be more dorsal in ToM studies than in self-reference studies, where the activity tends to be more ventral. Sometimes overlap between ventral and dorsal mPFC is observed when perceivers are asked to infer the mental states of targets that are most similar to them (Mitchell, Banaji, & Macrae, 2005). This latter finding suggests that mental simulation is possibly engaged in ToM tasks (i.e., "what would I do if I were that person?") illustrating a possible common role for the mPFC in both self-awareness and ToM. Although activity in other brain regions has been observed during ToM tasks—notably, the superior temporal sulcus (STS), the temporo-parietal junction (TPJ), and less often the amygdala—dorsal mPFC appears to play a central role in the ability to make mental state attributions about other people. Indeed, as Rilling notes in his chapter, this area reliably differentiates between when people are interacting with people versus computers, even engaging in the same tasks. That is, "people" are given privileged status by dmPFC as it processes information in the environment (*see* Mitchell, Heatherton, & Macrae, 2002)

Theory of mind processes play an important role in the type of social interaction tasks described in Rilling's chapter, such as the Prisoner's Dilemma game used by Rilling and colleagues (2004) or the Ultimatum game used by Sanfey and colleagues (2003). In the latter task, research participants often act irrationally, such as refusing unfair offers that result in receiving even less—that is, a person would rather have nothing than something, if it means that the other person in dyadic interaction comes out on top. Rilling notes an interesting power struggle between prefrontal regions and the insular cortex, an area that has been identified in the visceral aspects of emotion. When the frontal lobes win the struggle, the person acts "rationally" and takes the best offer, whereas when the insula is more active, the person chooses to reject the unfair offer. The reciprocal relations between the frontal "cognitive" regions and the limbic "emotional" regions are observed in Greene's chapter of moral judgment and in Eisenberger's chapter on social pain. Indeed, this struggle between cognition and emotion has been around since the ancient Greeks, but these studies show the ability to study the dynamic processes that underlie the resolution of these conflicts.

Detection of Threat

Over the course of human evolution, a major adaptive challenge to survival was other people. Put simply, other people can be dangerous. There are two basic social threats: those from the in-group and those from the out-group. The nature of these threats is distinctly different with the major threat from the in-group being social exclusion. As mentioned earlier, humans have a fundamental need to belong because during the course of evolutionary history, being kicked out of the group was a potentially fatal sentence. By contrast, out-group members are threatening because they want to take your group's resources or they may even want to kill you. Thus, the social brain requires threat mechanisms that differentiate in-group from out-group or that are differentially sensitive to the nature of the social threat. A variety of brain regions have been identified as relevant to the detection of threat, but the two most prominent regions are the amygdala and the anterior cingulate cortex. Both regions have been implicated in social cognition.

Let's start with the out-group. In the social neuroscience literature, the most common area identified as relevant to threat from outgroup members is the amygdala. For example, studies have associated amygdala activity with negative response to Blacks (Cunningham et al., 2004; Phelps et al., 2000, Richeson et al., 2003). People who possess stigmatizing conditions that make them seem less than human, such as the homeless, also activate regions of the amygdala (Harris & Fiske, 2006). We also have found amygdala responses to the physically unattractive or people with multiple facial piercings (Krendl et al., 2006). Considered together, it is clear that evaluating out-group members involves activity of the amygdala. So, what does the amygdala do in the social context? It has long been thought to play a special role in responding to stimuli that elicit fear (Blanchard & Blanchard, 1972; Feldman Barrett & Wager, 2006; LeDoux, 1996). From this perspective, affective processing in the amygdala is a hardwired circuit that has developed over the course of evolution to protect animals from danger. For example, much data support the notion that the amygdala is robustly activated in response to primary biologically relevant stimuli (e.g., faces, odors, tastes, etc.), even when these stimuli remain below the subjects' level of reported awareness (e.g., Morris et al., 1998; Whalen et al., 1998).

However, many recent imaging studies have observed amygdala activity to stimuli of both negative and positive valence, indicating that the amygdala is not solely concerned with fear. Indeed, some have argued that the amygdala is important for drawing attention to novel stimuli that have biological relevance. For example, Hamann and colleagues (2004) found that activity within the amygdala increased when both men and women viewed sexually arousing stimuli, such as short film clips of sexual activity or nude pictures of the opposite sex. Under this argument, it is plausible that the amygdala plays a role in processing social emotions because they have direct relevance in maintaining long-term social relations, which has been argued to reflect a fundamental need that is biologically relevant. My colleague, Paul Whalen, has argued that the amygdala is especially concerned with ambiguous stimuli that provide insufficient information to discern the nature of the threat (Whalen, 1998; 2007). This may be why fearful faces activate the amygdala to a greater extent than do angry faces (Whalen et al., 2001).

The chapter by Cacioppo et al. in this volume provides a very interesting examination of how the amygdala may be related to psychological well-being and how this may change across the life-course. Interestingly, aging can be considered a relatively recent adaptive problem in that only recently have people achieved life spans into the 70s, 80s, and beyond. An open question, therefore, is whether social mechanisms that have evolved over time are preserved in aging. Put another way, one might argue that the adaptive challenges associated with group living may differ substantially for older adults and that the mechanisms that guide social behavior reflect solutions that evolved for solving challenges that are less relevant to young humans (e.g., mate competition). In any case, Cacioppo and colleagues note that amygdala lesions are associated with a selective deficit in processing negative information, which is similar to the pattern observed along older adults. Interestingly, this pattern may be interpreted to suggest that threat detection among older adults is diminished. This may explain why older adults are susceptible to scam artists, perhaps in part because they are too trusting. This may be caused, in part, by faulty threat detection.

How about threat from in-group members? If humans have a fundamental need to belong, then there should be mechanisms for detecting inclusionary status (Leary, Tambor, Terdal, & Down, 1995; MacDonald & Leary, 2005). Put another way, given the importance of group inclusion, humans need to be sensitive to signs that the group might exclude them. According to the sociometer model, self-esteem functions as a monitor of the likelihood of social exclusion. When people behave in ways that increase the likelihood they will be rejected, they experience a reduction in state self-esteem. There has recently been a series of studies that have examined social rejection. Most prominent is the study by Eisenberger and colleagues (2003), which she describes in her chapter in this volume. Specifically, Eisenberger et al. (2003) found that the dorsal region of the anterior cingulate cortex (dACC) was responsive during a video game designed to elicit feelings of social rejection when virtual interaction partners suddenly and surprisingly stopped cooperating with the research participant. Since this initial study, other studies have also implicated the ACC, although some of them find a more ventral rather than dorsal region. For example, in our hands, we found that social feedback about acceptance or rejection was associated with differential activity in the ventral ACC (vACC; Somerville, Heatherton, & Kelley, 2006). A more recent study using paintings portraying rejection imagery observed a somewhat different pattern than found in either of the previous studies (Kross, Egner, Ochsner, Hirsch, & Downey, 2007). Although these authors also found dACC to be responsive to rejection imagery, the response was in a different area of dACC from that found by Eisenberger et al., and the relation between feelings of rejection and activity in this area was opposite that reported by Eisenberger et al. Another recent study (Burklund, Eisenberger, & Lieberman, 2007) found a relationship between both dACC and vACC activity and rejection sensitivity during emotional processing, albeit the vACC activity was in a more subgenual region than that reported by Somerville, Heatherton, and Kelley (2006). The somewhat disparate findings of these studies indicate the need for further research to more clearly identify the neural correlates of states of social distress, especially in terms of the functional roles of dACC and vACC in processing and responding to threat cues. Eisenberger does an excellent job in her chapter of describing a specific role for dACC in social pain. What is especially impressive is the link between specific genetic polymorphisms and brain responses to rejection and self-reports of emotional hypersensitivity. This reflects an excellent multimethod approach to understanding the neurophysiological basis of social threat detection.

As mentioned, there is some ambiguity regarding the role of different regions of ACC in detecting and responding to interpersonal communication, such as that which occurs during rejection. I have become particularly interested in trying to understand the vACC response to social feedback, in part because various clinical populations have abnormal vACC structure and function. For example, a voxel-based morphometry study found reduced grey matter volume among medication-naïve major depressives (Tang et al., 2007). There are also imaging reports implicating vACC in emotional disorders. For example, a recent report of participants with post-traumatic stess disorder (PTSD)showed decreased vACC activity to trauma-script imagery (Frewen et al., 2008). Activity in vACC during the processing of emotional cues also predicts overall number of symptoms in PTSD (Shin et al., 2005) as well as which patients will respond to cognitive behavioral therapy (Bryant et al. 2007). At the same time, imaging studies have produced some conflicting results, with activity in vACC found for both positive and negative emotional tasks. As noted by many researchers, such differential patterns of activity may reflect anatomical distinctions within vACC (Gotlib et al., 2005; van den Bos et al., 2007).

One fairly consistent finding is that clinical disorders, such as PTSD and depression, are associated with reduced volume and activity in vACC (e.g., Drevets et al., 1997; Milad et al., 2007; Tang et al., 2007; although *see* Mayberg et al., 1999). One possibility is that a sustained (tonic) reduction in vACC activity may heighten responsitivity to transient positive social cues. According to this perspective, transient vACC activity in response to positive social feedback, like we observed in Somerville et al. (2006), may serve a regulatory or corrective role in dealing with negative affect. Indeed, a recent theory by Gotlib and colleagues (Cooney et al., 2007) proposes, "the subgenual cingulate is important in regulating negative affect through the processing of mood-enhancing information." These authors argue that there is a ventral regulatory network that promotes mood-incongruent processing, such as would be the case for positive feedback during a state of social distress. Additional evidence for a ventral regulatory mechanism can be found in Kevin Ochsner's work. In one study, when participants reappraised negative images in a self-relevant manner, increases in vACC activity were associated with decreasing negative mood (Ochsner et al., 2004). In all of these studies, task conditions conspired to promote negative emotional states, such as having subjects look at a series of negative expresses or images. In the presence of these negative states, vACC may modulate transiently to mitigate negative affect. Put another way, perhaps paradoxically, a tonic vACC reduction (such as occurs for clinical states or induced dysphoria) may accentuate transient increases in vACC during positive social feedback. These transient increases in vACC may be therapeutic in the sense that they help restore baseline levels observed prior to negative mood induction. A recent PET study found that successful cognitive behavioral therapy for depression was associated with increased metabolic activity in the vACC (Kennedy et al., 2007). This suggests that the vACC might not be involved in threat detection *per se* but, rather, may reflect the neural basis of self-regulatory efforts to handle such threats. This view of the ACC, as being involved in resolving conflict, is in keeping with the more traditional view of ACC in the neuroscience literature. I now turn to the regulatory component of the social brain.

Self-Regulation

People who defy group norms—such as by cheating, lying, or being incompetent—often experience social emotions that indicate that something is wrong. We feel embarrassed when we goof, guilty when we harm, and ashamed when we get caught. Similarly, encounters with out-group members can leave us wary or even afraid, even if they can ultimately override our prejudices and treat them fairly. The important point is that emotions that arise from social interactions serve as guides for subsequent behavior. This is what makes something like feeling guilty adaptive (Baumeister, Stillwell, & Heatherton, 1994). Feeling socially excluded, which threatens the need to belong, motivates behavior to repair social relationships. Feeling ashamed about considering cheating on our partner helps reign in temptations. In other words, social emotions promote self-regulation, which allows us to alter our behavior, make plans, choose from alternatives, focus attention on pursuit of goals, inhibit competing thoughts, and regulate social behavior (Baumeister, Heatherton, & Tice, 1994).

Neuroscience research indicates that various regions of PFC are responsible for the human capacity for self-regulation (*see* the review by Banfield, Wyland, Macrae, Munte, & Heatherton, 2004). For example, functional neuro-imaging studies have implicated the ACC in decision monitoring, initiating the selection of an appropriate novel response from several alternatives, performance monitoring, action monitoring, detection or processing of response conflict, and internal cognitive control (Wyland, Kelley, Macrae, Gordon, & Heatherton, 2003). More recently, we found an important role for the ACC in efforts to suppress unwanted thoughts (Mitchell et al., 2007). What we observed was that ACC was transiently engaged following the occurrence of unwanted thoughts, whereas dorsolateral PFC was most active during tonic efforts to suppress those thoughts. This finding is in keeping with the important role of prefrontal regions in executive functions more generally, all of which are necessary for successful self-regulation. Since the days of Phineas Gage, we have known the damage to certain prefrontal regions is associated with a lack of impulse control and self-regulatory difficulties more generally. The role of lateral PFC regions in regulating social emotions was also noted in Greene's, Eisenberger's, and Rilling's chapters and appears to be among the most robust findings in social neuroscience.

Conclusions

The four chapters in this section are proof that the new field of social neuroscience has not only sprung to life but is now walking on its own—perhaps out of infancy and into toddlerhood. That is, much remains to be known about the social brain and there is every reason to believe we will make great strides as the field matures. In this chapter, I have proposed that building a social brain requires four components, each of which involves distinct functional brain regions. First, people need self-awareness—to be aware of their behaviors so as to gauge them against societal or group norms and the available evidence indicates that ventral mPFC is especially important for the experience of self. Second, people need to understand how others are reacting to their behavior so as to predict how others will respond to them. This capacity for ToM has been most closely associated with a region of mPFC that is more dorsal than that observed for self-referential processing. Threat detection involves at least the amygdala and the ACC, although the precise nature of their roles in threat detection remains somewhat unclear. For example, the amygdala may be especially important in ambiguous situations, such as when people are anticipating negative social judgments, whereas the ACC may be more important once negative feedback has been received. Finally, self-regulation involves a number of prefrontal brain regions, including ACC, lateral PFC, and ventral-medial PFC. It is possible that these areas play different roles in self-regulation failure depending on whether the failure is related to an impaired sense of self (vmPFC), impaired ToM (dorsal PFC), or failure to detect threat or conflict (ACC). There is much yet to discover.

I hope this commentary has shown the value of social neuroscience in identifying the important components of the social brain. At the same time, by analyzing its component parts, we may know the ingredients necessary for making a social brain, but we have not yet perfected the recipe for actually cooking one. Only further research will help us understand how these various components interact to produce the fully cooked social brain.

References

Adolphs, R. (2003). Cognitive neuroscience of human social behaviour. *Nature Reviews Neuroscience*, 4(3), 165–178.

Amodio, D.M., & Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, 7(4), 268–277.

Banfield, J.F., Wyland, C.L., Macrae, C.N., Munte, T.F., & Heatherton, T.F. (2004). The cognitive neuroscience of self-regulation. In Baumeister, R.F., & Vohs, K.D. (Eds.), *Handbook of Self-Regulation: Research, Theory, and Applications* (pp. 63–83). New York: Guilford Press.

Barrett, L.F., & Wager, T.D. (2006). The structure of emotion: evidence from neuroimaging studies. *Current Directions in Psychological Science*, 15, 79–83.

Baumeister, R.F., Heatherton, T.F., & Tice, D.M. (1994). *Losing Control: How and Why People Fail at Self-regulation*. San Diego, CA: Academic Press.

Baumeister, R.F., Stillwell, A.M., & Heatherton, T.F. (1994). Guilt: an interpersonal approach. *Psychological Bulletin*, 115, 243–267.

Baumeister, R.F., & Leary, M.R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3), 497–529.

Beer, J.S., Heerey, E.A., Keltner, D., Scabini, D. & Knight, R.T. (2003). The regulatory function of self-conscious emotion: insights from patients with orbitofrontal damage. *Journal of Personality and Social Psychology*, 85, 594–604.

Blanchard, D.C., & Blanchard, R.J. (1972). Innate and conditioned reactions to threat in rats with amygdaloid lesions. *Journal of Compartive and Physiological Psychology*, 81, 281–290.

Bowlby, J. (1969). *Attachment and Loss (Vol. 1)*. New York: Basic Books.

Bryant, R.A., Felmingham, K., Kemp, A., et al. (2007). Amygdala and ventral anterior cingulated activation predicts treatment response to cognitive behaviour therapy for post-traumatic stress disorder. *Psychological Medicine*, 16, 1–7.

Burklund, L.J., Eisenberger, N.I., & Lieberman, M.D. (2007). The face of rejection: rejection sensitivity moderates dorsal anterior cingulated activity to disapproving facial expressions. *Social Neuroscience*, 2, 238–253.

Cacioppo, J.T., Amaral, D., Blanchard, J.J., et al. (2007). Social neuroscience: progress and promise. *Perspectives on Psychological Science*, 2, 99–123.

Cooney, R.E., Joorman, J., Atlas, L.Y., Eugene, F., & Gotlib, I.H. (2007). Remembering the good times: neural correlates of affect regulation. *Neuroreport*, 18, 1771–1774.

Craik, F.I.M., Moroz, T.M., Moscovitch, M., et al. (1999). In search of the self: a positron emission tomography study. *Psychological Science*, 10, 26–34.

Cunningham, W.A., Johnson, M.K., Raye, C.L., Gatenby, C.J., Gore, J.C., & Banaji, M.R. (2004). Separable neural components in the processing of black and white faces. *Psychological Science*, 15(12), 806–813.

Drevets, W.C., Price, J.L., Simpson, J.R., Jr., et al. (1997). Subgenual prefrontal cortex abnormalities in mood disorders. *Nature*, 386(6627), 824–827.

Eisenberger, N.I., Lieberman, M.D., & Williams, K.D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, 302(5643), 290–292.

Frewen, P.A., Lanius, R.A., Dozois, D.J., et al. (2008). Clinical and neural correlates of alexithymia in posttraumatic stress disorder. *Journal of Abnormal Psychology*, 117, 171–181.

Gallagher, H.L, & Frith, C.D. (2003). Functional imaging of 'theory of mind.' *Trends in Cognitive Sciences*, 7(2), 77–83.

Gillihan, S.J., & Farah, M.J. (2005). Is self special? A critical review of evidence from experimental psychology and cognitive neuroscience. *Psychological Bulletin*, 13(1), 76–97.

Gotlib, I.H., Sivers, H., Gabrieli, J.D., et al. (2005). Subgenual anterior cingulated activation to valenced emotional stimuli in major depression. *Neuroreport*, 16(6), 1731–1734.

Gusnard, D.A. (2005). Being a self: considerations from functional imaging. *Consciousness and Cognition*, 14, 679–697.

Hamann, S., Herman, R.A., Nolan, C.L., & Wallen, K. (2004). Men and women differ in amygdala response to visual sexual stimuli. *Nature Neuroscience*, 7, 411–416.

Harris, L.T., & Fiske, S.T. (2006). Dehumanizing the lowest of the low: neuroimaging responses to extreme out-groups. *Psychological Science*, 17(10), 847–853.

Heatherton, T.F., & Krendl, A.C. (2009). Social emotions: neuroimaging. In Squire, L. (Ed.), *Encyclopedia of Neuroscience*, Vol. 9 (pp. 35–39). Oxford: Academic Press.

Heatherton, T. F., Macrae, C. N., & Kelley, W.M. (2004). What the social brain sciences can tell us about the self. *Current Directions in Psychological Science*, 13(5), 190–193.

Heatherton, T.F., Wyland, C.L., Macrae, C.N., Demos, K.E., Denny, B.T., & Kelley, W.M. (2006). Medial prefrontal activity differentiates self from close others. *Social Cognitive and Affective Neuroscience*, 1, 18–25.

Johnson, S.C., Baxter, L.C., Wilder, L.S., Pipe, J.G., Heiserman, J.E., & Prigatano, G.P. (2002). Neural correlates of self-reflection. *Brain*, 125(Pt 8), 1808–1814.

Kelley, W.M., Macrae, C.N., Wyland, C.L., Caglar, S., Inati, S., & Heatherton, T.F. (2002). Finding the self?: an event-related fMRI study. *Journal of Cognitive Neuroscience*, 14(5), 785–794.

Kennedy, S.H., Konarski, J.Z., Segal, Z.V., et al. (2007). Differences in brain glucose metabolism between responders to cbt and venlafaxine in a 16-week randomized controlled trial. *American Journal of Psychiatry*, 164, 778–788.

Krendl, A.C., Macrae, C.N., Kelley, W.M., Fugelsang, J.F., & Heatherton, T.F. (2006). The good, the bad, and the ugly: an fMRI investigation of the functional anatomic correlates of stigma. *Social Neuroscience*, 1(1), 5–15.

Kross, E., Egner, T., Ochsner, K., Hirsch, J., & Downey, G. (2007). Neural dynamis of rejection sensitivity. *Journal of Cognitive Neuroscience*, 19, 945–956.

Leary, M.R., Tambor, E.S., Terdal, S.K., & Downs, D.L. (1995). Self-esteem as an interpersonal monitor: the sociometer hypothesis. *Journal of Personality and Social Psychology*, 68(3), 518–530.

LeDoux, J.E. (Ed.). (1996). *The Emotional Brain*. New York: Simon and Schuster.

Macdonald, G., & Leary, M.R. (2005). Why does social exclusion hurt? The relationship between social and physical pain. *Psychological Bulletin*, 131, 202–223.

Macrae, C.N., Moran, J.M., Heatherton, T.F., Banfield, J.F., & Kelley, W.M. (2004). Medial prefrontal activity predicts memory for self. *Cerebral Cortex*, 14(6), 647–654.

Mayberg, H.S., Liotti, M., Brannan, S.K., et al. (1999). Reciprocal limbic-cortical function and negative mood: converging PET findings in depression and normal sadness. *American Journal of Psychiatry*, 156, 675–682.

Mitchell, J.P., Heatherton, T.F., & Macrae, C.N. (2002). Distinct neural systems subserve person and object knowledge. *Proceedings of the National Academy of Sciences of the United States of America*, 99(23), 15238–15243.

Mitchell, J.P., Banaji, M.R., & Macrae, C.N., (2005). General and specific contributions of the medial prefrontal cortex to knowledge about mental states. *Neuroimage*, 28(4), 757–762.

Mitchell, J.P., Heatherton, T.F., Kelley, W.M., Wyland, C.L., Wegner, D.M., & Macrae, C.N. (2007). Separating sustained from transient aspects of cognitive control during thought suppression. *Psychological Science*, 18, 292–297.

Moran, J.M., Macrae, C.N., Heatherton, T.F., Wyland, C.L., & Kelley, W.M. (2006). Neuroanatomical evidence for distinct cognitive and affective components of self. *Journal of Cognitive Neuroscience*, 18, 1586–1594.

Morris, J.S., Ohman, A., & Dolan, R.J. (1998). Conscious and unconscious emotional learning in the human amygdala. *Nature*, 393, 467–470.

Ochsner, K., Ray, R.D., Cooper, J.C., et al. (2004). For better or for worse: neural systems supporting the cognitive down- and up-regulation of negative emotion. *Neuroimage*, 2, 483–499.

Phelps, E.A., O'Connor, K.J., Cunningham, W.A., et al. (2000). Performance on indirect measures of race evaluation predicts amygdala activation. *Journal of Cognitive Neuroscience*, 12(5), 729–738.

Richeson, J.A., Baird, A.A., Gordon, H.L., et al. (2003). An fMRI investigation of the impact of interracial contact on executive function. *Nature Neuroscience*, 6(12), 1323–1328.

Schmitz, T.W., Kawahara-Baccus, T.N., & Johnson, S.C. (2004). Metacognitive evaluation, self-relevance, and the right prefrontal cortex. *Neuroimage*, 22(2), 941–947.

Shin, L.M., Wright, C.I., Cannistraro, P.A., et al. (2005). A functional magnetic resonance imaging study of amygdala and medial prefrontal cortex responses to overtly presented fearful faces in post-traumatic stress disorder. *Archives of General Psychiatry*, 62, 273–281

Somerville, L.H., Heatherton, T.F., & Kelley, W.M. (2006). Dissociating expectiancy violation from social rejection. *Nature Neuroscience*, 9(8), 1007–1008.

Tang, Y., Wang, F., Xie, G., et al. (2007). Reduced ventral anterior cingulated and amygdala volumes in medication-naïve females with major depressive disorder: a voxel-based morphometric magnetic resonance imaging study. *Psychiatry Research*, 156, 83–86.

Turk, D.J., Heatherton, T.F., Macrae, C.N., Kelley, W.M., & Gazzaniga, M.S. (2003). Out of contact, out of mind: the distributed nature of the self. *Annals of the New York Academy of Sciences*, 1001, 65–78.

van den Bos, W., McClure, S.M., Harris, L.T., Fiske, S.T., & Cohen, J.D. (2007). The role of ventral medial prefrontal cortex in processing affective and social stimuli. *Cognitive, Affective, and Behavioral Neuroscience*, 7, 337–346.

Whalen, P.J. (1998). Fear, vigilance, and ambiguity: initial neuroimaging studies of the human amygdala. *Current Directions in Psychological Science*, 7(6), 177–188.

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(1), 411–418.

Whalen, P.J., Shin, L.M., McInerney, S.C., Fischer, H., Wright, C.I., & Rauch, S.L. (2001). A functional MRI study of human amygdala responses to facial expressions of fear versus anger. *Emotion*, 1, 70–83.

Whalen, P.J. (2007). The uncertainty of it all. *Trends in Cognitive Sciences*, 11, 499–500.

Wyland, C.L., Kelley, W.M., Macrae, C.N., Gordon, H.L, & Heatherton, T.F. (2003). Neural correlates of thought suppression. *Neuropsychologia*, 41(14), 1863–1867.

GENERAL COMMENTARY
Hanging with Social Neuroscientists

Marcia K. Johnson

The first meeting I attended that had an explicit social neuroscience agenda was hosted by Todd Heatherton and his Dartmouth colleagues (Jay Hull, Robert Kleck, Neil Macrae, & Jennifer Richeson) at the Minary Conference Center in New Hampshire in August, 2001. It was a small group and, I believe, one of the first gatherings under the nascent umbrella of social neuroscience. There was much discussion about whether a new name such as *social neuroscience* or *social/cognitive neuroscience* or *social/affective neuroscience* was necessary or desirable to point the field of social psychology in new directions, or whether the already established term *cognitive neuroscience* was comprehensive enough to include the neural substrates of processes and phenomena that were the focus of social psychology.

In 2005, I attended two meetings of somewhat larger groups. One at Princeton in May was hosted by Alex Todorov, Susan Fiske, and Debbie Prentice and is the origin of this collection of chapters from pioneers extending the boundaries of social psychology. The psychological issues addressed, the study designs described, and the findings reported at this meeting had all the features of an exciting and promising new field. As I recall, at that time, there was still some discussion about the best name for this new field. I believe I suggested *human neuroscience* because it could encompass many interrelated strands (including clinical neuroscience, developmental neuroscience, and neuro-economics). However, *human neuroscience* has a drawback in that it seems to exclude neuroscience research using nonhuman animals—clearly an important domain. *Neuropsychology* would be seemingly straightforward and comprehensive but, historically, has the connotation of being limited to studies of patient populations (another important source of evidence). Thus, I think we will continue to see "neuroscience" affixed to our traditional group labels (clinical, developmental, social, cognitive, etc.) and combinations of them (social/affective; social/cognitive; social/developmental) as individuals and programs signal their interest in neural substrates of various psychological phenomena.

The other meeting I attended in 2005 (*Social Neuroscience and Behavior: From Basic to Clinical Science*) was in June, sponsored by NIMH and hosted by Kevin Quinn and Mike Kozak. Under the leadership of John Cacioppo, this meeting focused on the potential of social neuroscience to make contributions to the field of mental health (Cacioppo et al., 2007). Thus, by 2005, social neuroscience not only existed, but the promise for a productive merger of basic social neuroscience and clinical research was already recognized—that is, social neuroscience was becoming social/affective/clinical neuroscience, just as cognitive neuroscience had expanded into social, clinical, and developmental domains.

Recently, I attended the second annual meeting of the *Social & Affective Neuroscience Society*

(Boston, May, 2008). The program was organized by Jason Mitchell, David Amodio, Jennifer Beer, Wil Cunningham, Matt Lieberman, and Kevin Ochsner. Approximately 250 people attended, and the program included two keynote addresses (Chris Frith, Tom Insel), 2 days of symposia (person perception, stereotyping and prejudice, morality, deception and trust, the self in social cognition, and self-regulation), and two lively poster sessions. In my talk, I joked that social and affective neuroscience had staked a flag in the medial prefrontal cortex (PFC) because cognitive neuroscientists had already taken up residence in the lateral PFC.

I am a little wistful that "cognitive" was not incorporated into the Society's name. I understand that there are important functions of group identity, but we should be mindful of the various effects of in-group/out-group identity, including even arbitrary divisions (Brewer, 1979; Van Bavel, Packer & Cunningham,2008). On the other hand, social and affective processes are so infused with cognition (and vice versa, Johnson & Sherman, 1990) that perhaps *cognition* doesn't need to be explicitly stated. In any event, a major challenge for the future is to better understand how lateral and medial PFC, along with other brain areas, work together to create socially relevant phenomenal experiences (e.g., thoughts, feelings, memories), including implicit effects (e.g., on judgments, actions).

During the same period that these and other related meetings were taking place, new journals encouraging reports of research in social neuroscience were launched: *Cognitive, Affective, & Behavioral Neuroscience* (*CABN*, 2001, John Jonides, editor), *Emotion* (2001, Richard Davidson and Klaus Scherer, joint-editors), *Social Cognitive and Affective Neuroscience* (*SCAN*, 2006, Matt Lieberman, editor), and *Social Neuroscience* (2006, Jean Decety, editor). Of course, there was social neuroscience research before 2001 (e.g., the work of John Cacioppo and others) and outlets for papers (e.g., *Cognition & Emotion*), just as there had been cognitive neuroscience research before Mike Gazzaniga and friends launched the *Journal of Cognitive Neuroscience* (1989) and the *Cognitive Neuroscience Society* (1993). However, societies and journals not only reflect but prompt research activity. Thus, here we are in 2008 with a dramatically increasing number of papers being published each year at the interface of cognition, emotion, social psychology, and neuroscience. The current level of activity, criss-crossing these particular disciplinary borders, would have been hard to imagine even 10 years ago.

Psychology is a wonderful discipline in its diversity. Its domains are varied, rich, and challenging; it can be approached at many levels of analysis and with many techniques; and its edges go everywhere (e.g., biology, chemistry, neurology, psychiatry, sociology, political science, economics, philosophy, linguistics, law, art, literature). Interestingly, neuroimaging as a research technique has not made psychology more reductionistic, isolated, and less relevant as some feared. Rather, I think it has given researchers a new common currency (in this sense, Neuro is like the Euro), helping people from diverse disciplines or orientations within disciplines to relate to each other's work and expand their own vision as they explore complex concepts such as memory, control, intentionality, self, empathy, attitudes, choice, and so forth. If anything, psychology has looked outward more (not less), become more (not less), interdisciplinary—more omnivorous and (appropriately, in my view) imperialistic. New methods and findings can reinvigorate researchers to tackle tough issues from a fresh perspective. Neuro-imaging data does not replace behavioral data or a psychological level of analysis; rather, interpreting neuro-imaging data depends on and furthers our understanding of psychological processes. Psychology has much to contribute to helping understand brain function, and the synergy between behavioral and neuroscience approaches is a key to continued progress. Appropriate clinical applications or applications in domains of social and economic policy (education, legal decisions) will depend on such synergistic progress.

Compared to when I started my first faculty position in 1970, now it is much harder for any one researcher to learn all that it would be helpful to know. This is true in all areas of psychology, but especially so for those trying to work at the intersection of neuroscience and social/cognitive/affective psychology. To capitalize on different types of expertise, working more in research teams that cut across labs would be a great benefit. But investigators—especially younger researchers—receive mixed messages. On the one hand, they should do cutting edge research and be productive. This goal may, in fact, be best served by a group effort, moving fastest, drawing on the highest levels of expertise available. On the other hand, they should be independent. Independence is easier to demonstrate if you work alone or only with more junior colleagues such as graduate students. But, I think we hurt the field when we discourage young investigators from working with senior colleagues or in teams with peers. We also slow progress when we suggest that it is better to demonstrate a new phenomenon (or relabel an old one) than to solve an old problem or advance an established theory. For psychology to take full advantage of the strengths provided by teamwork that draws on multiple types and sources of expertise, we need to take the time and make the effort to assess individual contributions within a team context. For psychology to be a cumulative science, we need to make sure rewards (publications, promotions, grants, etc.) recognize teamwork and cumulative contributions as well as individual efforts and the buzz of the new.

It has been fascinating to watch this new field take off. I've been delighted to be invited to meetings. I've learned a great deal and also have very much enjoyed the enthusiasm and high spirits of the attendees. I had one of my best poker hands ever at a raucous game at the Minary conference, and dinners hosted by Susan Fiske in Princeton and by Jason Mitchell in Cambridge were great fun. I admit it gave me pause when Liz Phelps noted a similarity between my career and the TV show *Survivor* and when Kevin Ochsner and Jason Mitchell started probing for my "historical perspective." Historical perspective is definitely a plus for a cumulative science, but I would also reiterate the observation that Jim Sherman and I (1990) quoted from Carly Simon: "These are the good old days." History is every day. And every day is an opportunity for perspective.

Marcia K. Johnson
New Haven
June, 2008

References

Van Bavel, J.J., Packer, D.J., & Cunningham, W.A. (in press). The neural substrates of in-group bias: An functional magnetic resonance imaging investigation. *Psychological Science, 19*, 1131–1139.

Brewer, M.B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin, 86*, 307–324.

Cacioppo, J.T., Amaral, D.G., Blanchard, J.J., et al. (2007). Social neuroscience: Progress and implications for mental health. *Perspectives on Psychological Science, 2*, 99–123.

Johnson, M. K., & Sherman, S. J. (1990). Constructing and reconstructing the past and the future in the present. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and social cognition: Foundations of social behavior* (pp. 482–526). New York: Guilford Press.

AUTHOR INDEX

Note: Page references followed by "*f*" and "*t*" denote figures and tables, respectively.

SUBJECT INDEX

Note: Page numbers followed by "*f*" and "*t*" denote figures and tables, respectively.